EECS 6414 - Final 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

ABSTRACT

Recently there has been a growing trend in publishing research works by utilizing the powerful language generation and reasoning capabilities of the Large Language Models. Subsequently, the goal of this work is to examine the interactions between researchers who frequently publish in Arxiv, that are related to Large Language Models. To support this analysis, results are drawn from a set of 'sub-arxivs' of the 'Computer Science' and 'Statistics' discipline. Due to the lack of readily available data, considerable effort was devoted to collecting, preprocessing, and analyzing metadata from the whole Arxiv dataset. The initial focus of this study is to determine how Co-Authorship Networks are formed between authors who are affiliated with both industry and academia. Notably, our analysis demonstrates an increase in papers published in recent times by academic institution-affiliated researchers, alongside the greater movement of industry-affiliated researchers. Finally, we perform a plethora of data analysis in different dimensions to further reinforce our findings.

1 INTRODUCTION

1.1 Motivation

In this work, we 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. We analyze the arXiv dataset's publication patterns over the years, which shows us researchers from which institutions 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 reveals 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 [13], and their ongoing influence on interdisciplinary research over time.

Recently, Large Language Models (LLMs) like GPT-4 [14], Gemini [8], LLaMA-2 [17] have taken the world by storm. Research into Large Language Models (LLMs) is essential given their far-reaching consequences that extend well beyond the traditional boundaries of computer science. Their transformative potential is undeniable, with applications ranging from creative text generation to sophisticated code understanding, article summarization, and advanced reasoning. Driven by massive datasets and ever-increasing computational resources, LLMs offer a paradigm shift in language processing, opening new possibilities in fields like linguistics (where complex patterns can be analyzed) and healthcare (where patient communication becomes more nuanced thanks to their language abilities). This is been further exemplified by their integration in the research works of not only Computer Science-related fields but also in other fields, such as Audio and Speech Processing, Quantitative Biology, etc. We discuss our potential findings regarding this in §3.

We aim to answer the following questions with our project:

- What trend can we infer from the Co-Authorship Networks? (§2.4)
- How are the industry-academia collaboration network formed? (§2.5)
- How are institutions collaborating? (§2.5 & §3)
- Which topics and authors are driving the growth of LLM research? (§2.4 & §2.2)
- What does the publication trend look like? (§3)

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, 9, 10]. 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.

Notably, the arXiv has frequently been utilized as a research dataset in various studies. Examples include Liben-Nowell et al.'s [11] 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. [12] in machine learning and physics. The arXiv data also underpinned the 2003 KDD Cup [7], 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

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

[2] https://arxiv.org

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 into 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.

Following the preprocessing, we shift focus to a specialized form of network analysis. Utilizing tools like "Python", "NetworkX", and "Gephi", we 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 explore the impact of recent advancements in LLMs and analyze their influence on interdisciplinary research.

2 METHODOLOGY

2.1 Dataset

The dataset we 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-arxivs. 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. To overcome the abovementioned limitation, we utilize another dataset from a GitHub repository[4]. The arxiv metadata was collected from January 2018 to September 2023. The metadata points consist of arXiv ID, author list, title, abstract, submission date, and subject categories assigned within arXiv. To focus on papers related to either computer science (CS) or statistics (Stat), data was refined to include only those listed in at least one CS or Stat sub-arXiv (presumably large language model-related papers are more likely to appear in these domains), resulting in 418K papers. 2018 was chosen as a starting point to align with the increasing use of pre-trained language model representations such as BERT [3] and ELMo [16]. In addition, to further extend our work, we performed an Author-affiliation network analysis which required not only the metadata of the arxiv papers but also the PDF version of the papers as well. Subsequently, PDF full-texts for these papers were obtained from a GCP bucket co-hosted by Kaggle and arXiv [5]. These PDFs were converted into plaintext files by utilizing the pdftotext [6] Python tool. Additionally, to gather insights pertaining to influential papers, citation data for the 16,979 LLM papers described below were retrieved via the Semantic Scholar API.

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.

Data Acquisition and Initial Filtering: In accordance with previous Machine Learning (ML) survey literature [5, 6, 15], the filtering process compiled a subset of Large Language Model (LLM) publications by searching for an interpretable set of keywords. The filtering process is performed on the data based on keywords related to a specific area of interest within the broader fields of CS and Stat. In this case, we focus on research pertaining to large language models (LLMs) and ChatGPT.

Keyword-Based Refinement: As the primary goal was to characterize temporal trends, a broad set of terms was selected intentionally, aiming to capture relevant publications predating modern instruction-tuned, chat-style LLMs to address research questions regarding changes over time more effectively. Thus, some terms are included such as "language model" and "BERT," which have been used for an extended period, alongside more recently introduced terms like "large language model." The comprehensive keyword list encompasses "language model", "foundation model", "BERT", "XL-Net", "GPT-2", "GPT-3", "GPT-4", "GPT-Neo", "GPT-J", "ChatGPT", "PaLM", "LLaMA", encompassing 16,979 publications since 2018 that contain at least one of these keywords in their title or abstract.

The decision to include "foundation models" was made despite the potential ambiguity with vision models. However, many publications involved language to some extent (e.g., large vision-language models, LVLMs), warranting its inclusion to reflect the rising interest in multimodal approaches. Beyond the "language model" and "foundation model," the selection of specific model keywords was guided by referencing Wikipedia's page for LLMs [7]. Some models were excluded that are less popular in case of searching with terms, such as (e.g., Chinchilla, LaMDA, Galactica) or models with numerous false positives (e.g., OPT, Claude, BLOOM are unrelated to Natural Language Processing (NLP)).

Final data Preprocessing for the Co-Authorship Network Construction: In order to create two different networks, we divided the data into two subsets, one subset of papers where authors are from industry organizations such as Google, Microsoft, Samsung, OpenAI, etc., and industry research groups such as Google DeepMind, Microsoft AI research, etc. Another subset comprises authors from academia from academic institutions in different countries. Considering this, we prepare data for two different Co-authorship networks, one from the authors of the industry and another from the authors of academia. Thus 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

[3]https://www.kaggle.com
[4]https://github.com/rmovva/LLM-publication-patterns-public
[5]https://blog.arxiv.org/2020/08/05/leveraging-machine-learning-to-fuel-new-discoveries-with-the-arxiv-dataset
[6]https://pypi.org/project/pdftotext
[7]https://en.wikipedia.org/wiki/Large_language_model

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.

Academic/Industry Subset Creation: In the context of constructing a co-authorship network, the data preparation phase involved collecting and organizing a dataset comprising academic and industry papers. Each entry in this dataset was meticulously annotated with several critical pieces of information, including the identities of the authors, the publication date, the research domains the paper contributes to, and a binary classification distinguishing between industry and academic research.

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. We clean up the author data to extract individual author names and eliminate any unnecessary whitespace.

Unique Identifier Generation: To address the challenges of author disambiguation and encapsulate the breadth of an author's collaborative endeavors, we introduced a novel method for generating unique identifiers. This method involves aggregating the names of an author's co-authors and the domains associated with their collective publications. Specifically, for each author under consideration, we traverse the dataset of papers to extract and compile a list of associated domains for papers where the author has contributed. Notably, the author's name is excluded from their own list of co-authors to ensure the uniqueness and relevance of the identifier. The final identifier is crafted by concatenating the author's name, and associated domains, thereby creating a comprehensive and unique representation of each author's academic and collaborative footprint.

2.3 Network Analysis

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

2.4 Co-Authorship Network Analysis

Co-author Network Construction: With the dataset prepared, the next phase involved the actual construction of the co-authorship network. We created subsets of all the selected papers, (a) the papers that are published by authors affiliated with academic institutions (see Figure 1 (a)), and (b) the papers that are published by authors affiliated with industry institutions (see Figure 1 (b)). From these subsets, we created two Co-authorship networks. The networks were initialized as undirected graphs, reflecting the bidirectional nature of co-authorship relationships. In this model, the directionality of collaboration is not considered, with the primary focus being on the existence of a collaborative link between any two authors.

Node & Edge Addition: Leveraging the unique identifiers, we proceed to construct the co-authorship networks. This process begins with the initialization of a graph structure, where each node represents an individual author. For each paper in the dataset, we identify the authors and their unique identifiers. Each author was added to the graph as a node, and we connected two authors with an edge between them if and only if they co-authored a paper. This process was iteratively applied across all entries in the dataset, cumulatively building a comprehensive representation of the collaboration network. Further, the node attributes included the author's original name and potentially other metadata derived from the dataset, such as their predominant research domain or affiliation. 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 approach ensures that the network graph accurately reflects the collaborative relationships across the entire dataset. 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. Recognizing the distinct nature of collaborations within academia and industry, our methodology extends to the development of separate networks for each domain. By segregating the dataset into academic and industry papers, we construct two parallel networks, enabling focused analysis of the unique characteristics and dynamics of collaboration within each sector.

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 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.

Visualization: After creating the co-authorship networks, we generate the visual representation of the graph that the network represents. Using the *Gephi & 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.

Analysis and Observations: The Figure 1 represents the two Co-authorship networks. Furthermore, Figure 2 and 3 present some general statistics such as (a) degree distributions, (b) connected component distributions, (c) betweenness centrality distributions and (d) eigenvector centrality distributions of Academic Co-Authorship Network and Industry Co-Authorship Network. From the figure, we can observe the following:

- The number of nodes and edges in the Academic Co-Authorship network (Figure 1 (a)) are much higher than the Industry Co-Authorship Network (Figure 1 (b))
- Although the Node degree distribution in the Academic Co-Authorship network is comparatively skewed than the Industry Co-Authorship Network. This indicates authors in

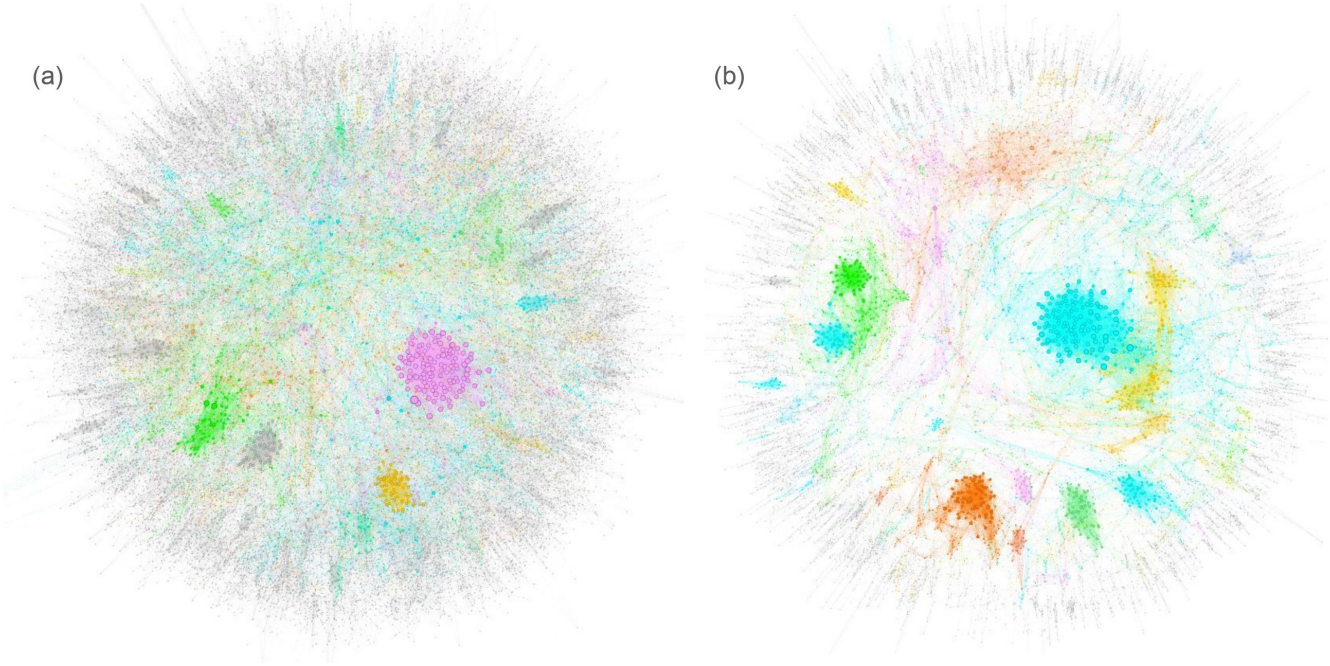

Figure 1: Figure (a) denotes the Academia Co-authorship network, and (b) denotes the Industry Co-authorship Network.

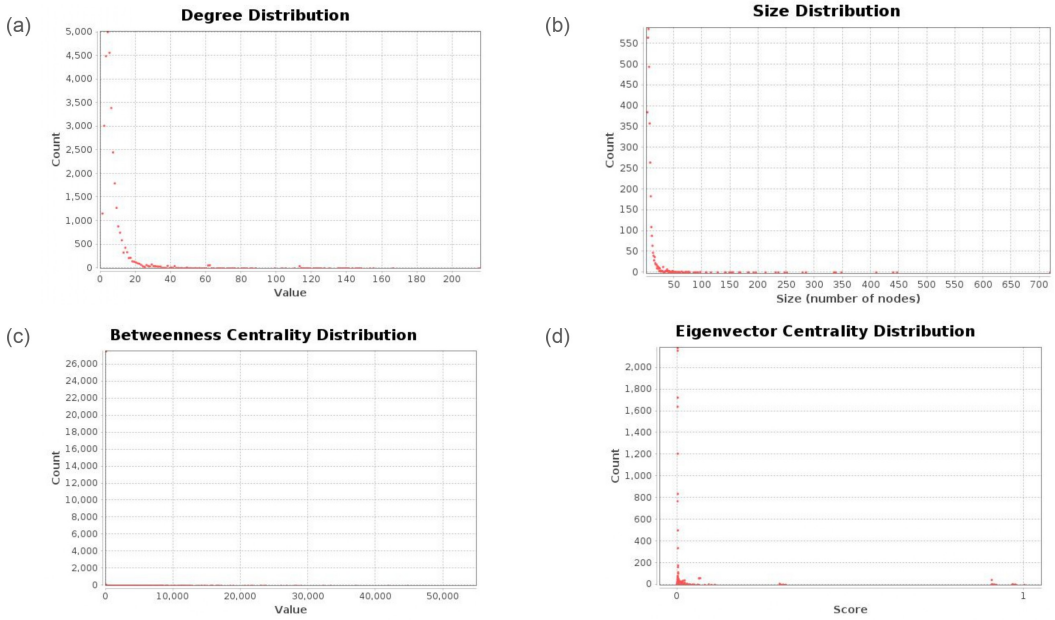

Figure 2: Network analysis of Academia Co-Authorship Network

the Industry are more likely to collaborate with each other than in Academic institutions

- In contrast, the Academic Co-Authorship network has a greater number of connected components (see Figure 2 (b)) in comparison to authors with industry affiliations (see Figure 3 (b))

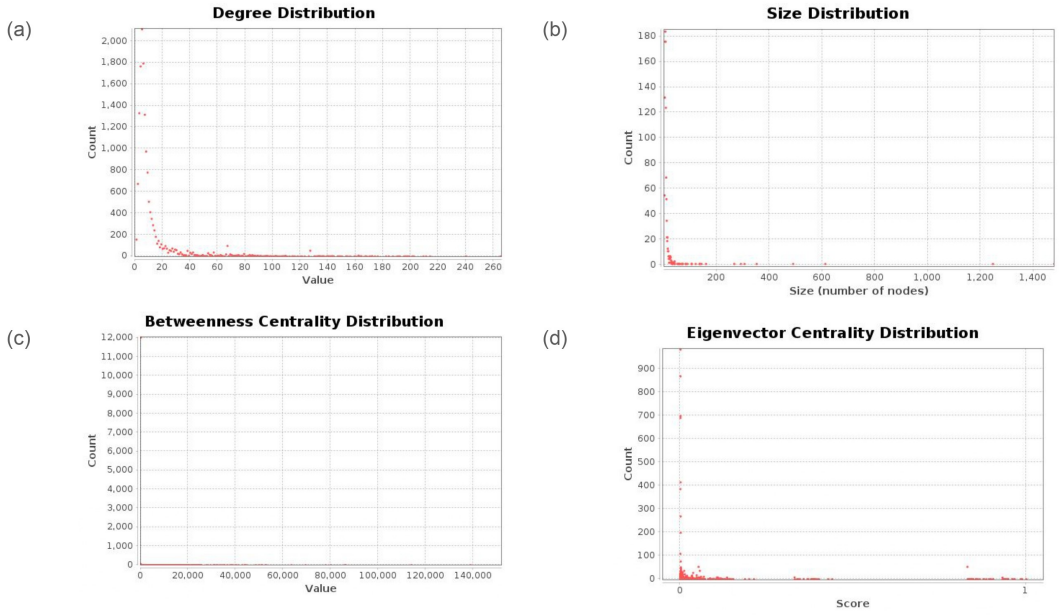

Figure 3: Network analysis of Industry Co-authorship Network

Table 1: General Statistics of Academic and Industry Co-Authorship Network

Network Statistics	Network	
	Academic Co-Authorship Network	Industry Co-Authorship Network
# of Nodes	32799	15084
# of Edges	120345	89464
# of Connected Components	3547	1196
Average Node Degree	7.338	11.862
Average Path Length	3.437	3.811
Network Diameter	13	13

2.5 Affiliation Network Analysis

Data Preprocessing: The initial phase of constructing the affiliation network involved a thorough preprocessing of the dataset. Each entry in the dataset, representing a research paper, was analyzed to extract information relevant to the network construction, particularly the affiliations of the authors to various institutions. A key aspect of this phase was the transformation of research domains into abbreviated institutional names, facilitating a standardized representation of institutions across the dataset.

Institution Identification: For each paper, the domains associated with the research were mapped to their corresponding institutions using a predefined dictionary. This mapping ensured that each institution was consistently identified across the dataset, regardless of the variability in domain names or the specificity of research areas.

Paper Count and Collaboration Tracking: Two critical metrics were tracked during the preprocessing stage: the count of papers contributed by each institution and the collaboration instances between pairs of institutions. The paper count provided a quantitative measure of the research output of each institution, while the collaboration count highlighted the frequency and strength of partnerships between institutions. These metrics were aggregated using "defaultdict" structures, allowing for efficient accumulation of counts throughout the dataset processing.

Affiliation Network Construction: With the preprocessing complete, the construction of the affiliation network commenced. The network was modeled as an undirected graph, where nodes represented institutions, and edges signified collaborations between them. This model aptly captured the bidirectional nature of collaboration, where the partnership is mutually beneficial and not hierarchically structured.

Node & Edge Addition: Nodes were added to the graph for each institution identified in the dataset, with attributes assigned based on the paper count and the nature of the institution (industry or academic). The differentiation between industry and academia was made based on the domain associated with the institution, adding a layer of contextual information to the network. Edges were introduced between pairs of institutions to represent their collaborations. The weight of each edge was determined by the collaboration count, offering a tangible measure of the partnership's intensity. This weighting provided a nuanced view of the network, highlighting not just the existence of collaborations but also their relative significance.

Visualization and Observations: The network is visualized using *PyVis* and *Gephi* library. Node sizes were scaled according to the paper count, and edge thicknesses were adjusted based on the collaboration count, providing a visual representation of the institution's research output and collaborative relationships. The Figure 4 depicts the Author Affiliation network. In the network, labels were selectively applied to nodes with larger paper counts to maintain

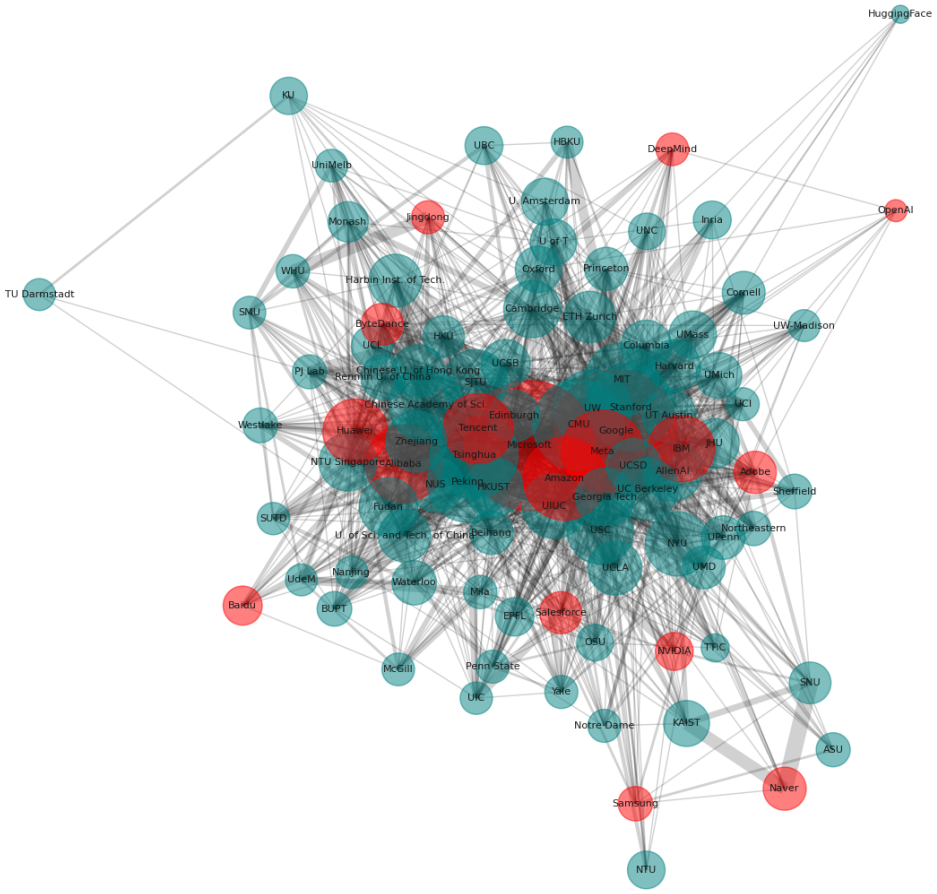

Figure 4: Author Affiliation Network

readability and focus on significant contributors within the network. This selective labeling strategy helped reduce visual clutter while ensuring that key players remained highlighted.

Some interesting observation from the author affiliation network in Figure 4 is listed below:

- Most big tech companies like Microsoft, Google, Meta, and Amazon from the West and Tencent, Huawei, and ByteDance from Asia are leading the Arxiv publication related to large language models
- Top-ranked academic institutions, such as Carnegie Mellon University, Stanford, UCLA, UIUC, MIT are leading publishers from the US, and Peking University, Tsinghua University, KAIST, NTU, and NUS are leading in LLM-related research publications in Asia

3 EVALUATION

3.1 Important Observations from the Arxiv Dataset

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 [8], seaborn [18], plotly[9], etc. Below we present some interesting observations from our network analysis.

Industry-Academia Collaboration in Large Language Model Research: The network in Figure 5 illustrates the collaboration between industry giants and academic institutions worldwide in LLM research, with a significant concentration of these collaborative efforts stemming from both American and Chinese entities. Notably, American tech companies like Microsoft, Amazon, and Google, and smaller specialized entities like AllenAI, along with prestigious universities such as MIT, Stanford, UC Berkeley, and CMU, underscore the strong focus on LLM research in the United States. Equally prominent are Chinese companies like Alibaba, Huawei,

[8] https://matplotlib.org
[9] https://plotly.com

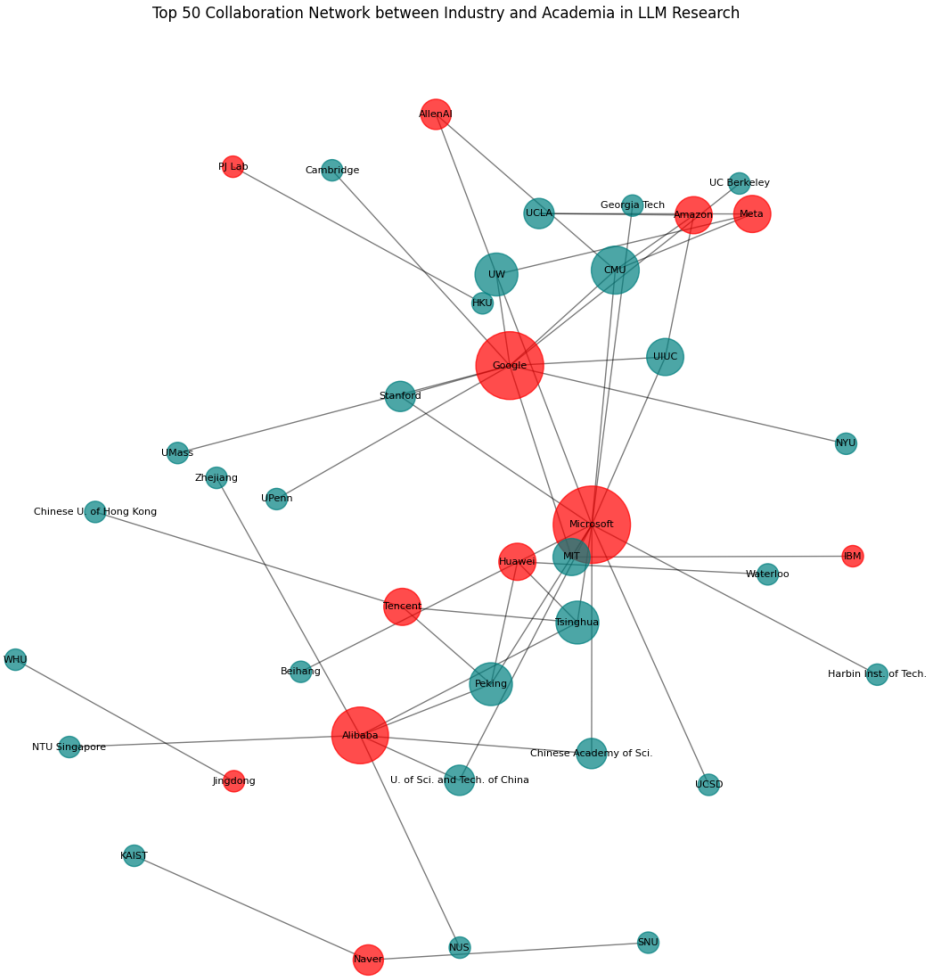

Figure 5: Top 50 Most Common Industry-Academia Collaboration Network

and Tencent, and educational institutions like Tsinghua University and Peking University, which signals a robust investment in LLM research within China.

The geographical distribution of these nodes implies a strong East-West dynamic in LLM research, with notable collaborations between U.S. and Chinese entities that could be indicative of cross-pollination in knowledge and technology despite any geopolitical tensions. Moreover, the presence of European universities such as Cambridge and Asian universities like the National University of Singapore (NUS) reflects a more global interest and contribution to the field. However, the central clustering of large industry nodes in the network indicates these companies' pivotal roles in driving research and potentially shaping the direction of partnerships and academic focus. This network also suggests a potential flow of talent and resources between these significant nodes, which might create regional poles of LLM expertise and innovation.

Leading the Charge in LLM Innovation: A Five-Year Snapshot: Figure 6 represents the top 50 institutions by the number of LLM papers published from 2018 to 2023, differentiating between industry and academia. Microsoft leads with a significant margin, publishing 729 papers, followed by Google and Carnegie Mellon University (CMU) with 654 and 359 papers respectively. Tsinghua University is the top academic contributor with 332 papers, closely followed by Stanford University. There's a notable presence of other tech companies such as Amazon and Meta, as well as a strong showing from other prestigious academic institutions like MIT, Peking University, and the University of Washington (UW). The data points suggest that industry giants are major contributors to LLM research, with academic institutions also making significant contributions. This indicates a vibrant and competitive field of study with substantial output from both sectors.

Academic vs. Industry LLM Research Trends (2018-2023): The line graph depicts a trend in the inclusion of LLM-related terms in research publications from 2018 to 2023, contrasting academic and industry contributions. Academic institutions show a striking upward trajectory, with a steep increase particularly noticeable from 2021 onwards, reflecting a growing academic interest and perhaps an expansion of LLM-related courses or fields of study.

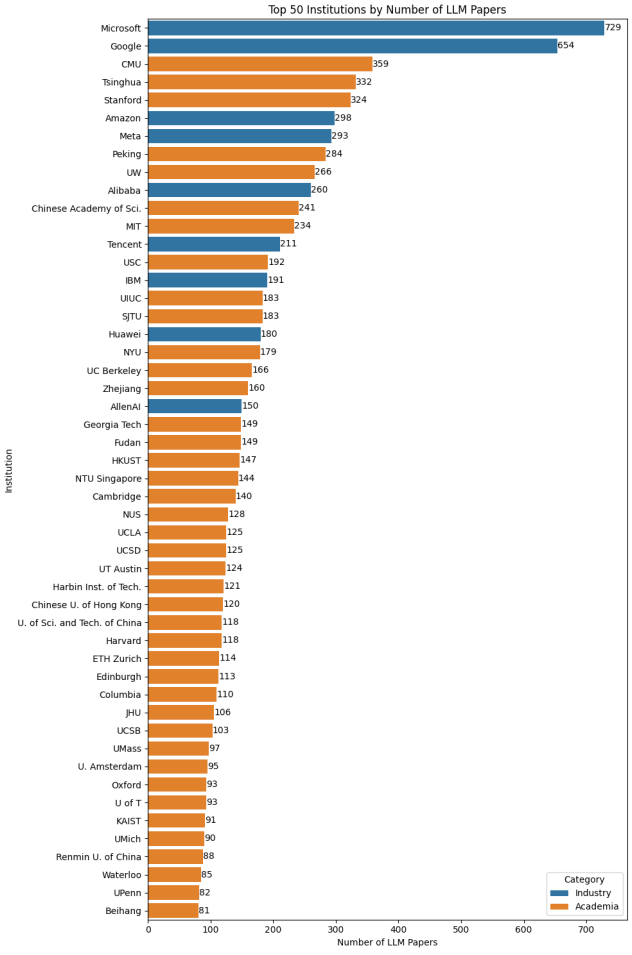

Figure 6: A breakdown of the most prolific contributors to Large Language Model research over 2018-2023, showcasing Microsoft's dominance in the industry and Tsinghua University's prominence in academia.

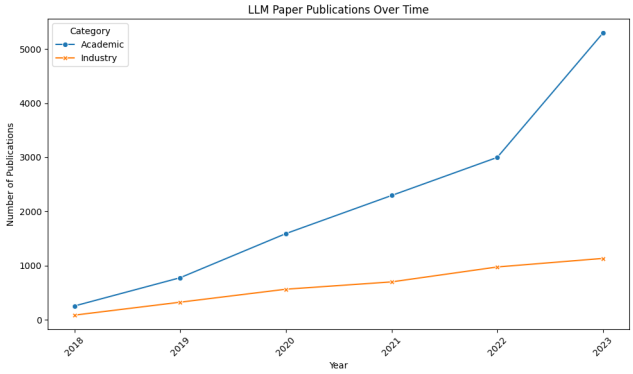

Figure 7: A clear divergence in publication growth rates with academia exponentially outpacing the industry in LLM research terms inclusion.

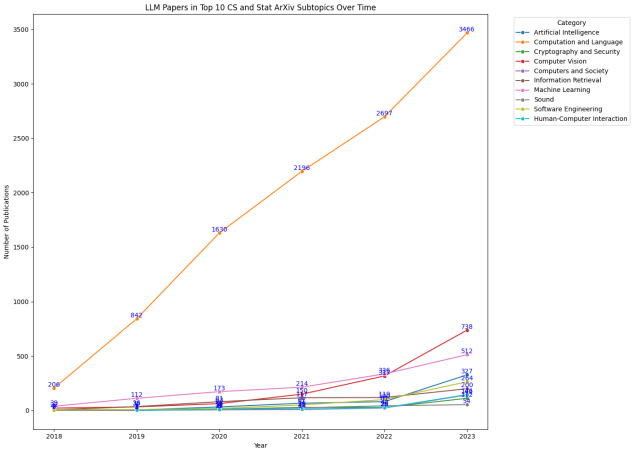

Figure 8: Dominating the discourse: 'Computation and Language' subcategory records an unprecedented rise in LLM publications, charting a robust trajectory and reflecting the burgeoning focus on language processing technologies.

By 2023, academic publications including LLM terms have risen to over 5000. In contrast, industry publications demonstrate a modest yet consistent growth over the same period, with the line flattening slightly toward 2023, suggesting a steady but more cautious or targeted approach to LLM research publication. The divergence between the two sectors may reflect academia's focus on exploring the theoretical and foundational aspects of LLMs, while industry may prioritize applied research with direct commercial applications.

The Linguistic Leap: LLM Research Trajectory on arXiv: Figure 8 is a line graph that displays the year-on-year growth of LLM papers in various subcategories of computer science and statistics on arXiv from 2018 to 2023, with "Computation and Language" experiencing a phenomenal ascent. The subtopic witnessed a 442.1% increase in publications from 2018 to 2019, followed by a 308.7% rise from 2019 to 2020. The growth rate then began to decelerate, posting a 93.6% increase from 2020 to 2021, a 65.5% increase from 2021 to 2022, and a 28.5% rise from 2022 to 2023. Despite the slowing growth rate in recent years, the absolute number of publications in "Computation and Language" vastly outnumbers other subtopics, indicating a significant and sustained focus in this area. Meanwhile, subtopics such as "Artificial Intelligence," "Machine Learning," "Cryptography and Security," and "Computer Vision" have also seen growth, but none as exponential as "Computation and Language," which is indicative of the burgeoning interest and research output in language processing technologies over the past five years.

The Interdisciplinary Reach of LLM: Expansion Beyond Core Tech Fields: Figure 9 depicts a chart that illustrates a significant year-on-year growth in the number of LLM papers published in the top 10 non-CS/Stat arXiv subtopics from 2018 to 2023. One subtopic, which is not specified by name but is represented by an orange line, shows remarkable growth. Starting from 14 publications in 2018, it

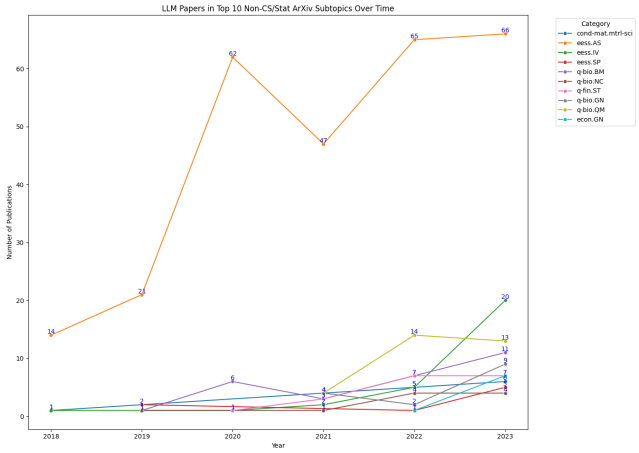

Figure 9: Tracing the ripple effect of LLM research into diverse scientific domains, highlighting notable growth in specific non-CS/Stat subtopics on arXiv from 2018 to 2023.

spiked to 62 in 2020, then dipped to 47 in 2021, before rising again to 66 by 2023. The fluctuations indicate that while there is interest in LLM across diverse scientific fields, it may be subject to varying research trends or the maturation of the field. The subtopics represented in the graph could include areas like *Electrical Engineering and Systems Science (eess.IV, eess.AS)*, and various subfields of *quantitative biology (q-bio)*, each showing a trend of increasing publications, although none as dramatically as the leading subtopic. This expansion beyond the traditional realms of CS and statistics underscores the interdisciplinary impact of LLM technologies.

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] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

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

[5] Lizhou Fan, Lingyao Li, Zihui Ma, Sanggyu Lee, Huizi Yu, and Libby Hemphill. 2023. A Bibliometric Review of Large Language Models Research from 2017 to 2023. arXiv:2304.02020 [cs.DL]

[6] Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 1905–1925. https://doi.org/10.18653/v1/2021.acl-long.149

[7] 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

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

[9] 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

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

[11] 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

[12] 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

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

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

[15] Kenny Peng, Arunesh Mathur, and Arvind Narayanan. 2021. Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers. arXiv:2108.02922 [cs.LG]

[16] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. arXiv:1802.05365 [cs.CL]

[17] 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]

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