CS 343: Graph Data Science
Spring 2025

Homework 2

Due: Sunday, April 6, 2025 at 11:59 PM

1 Introduction

You are provided with a dataset of papers, their authors, and their citations via a GitHub repository (https://github.com/habib-university/cs343-hw2). Your task is to load the data into Neo4j and conduct graph analytics.

2 Question

This assignment is open-ended, requiring you to design your own data model and load the data accordingly. Once the data is loaded, you will utilize graph analytics queries, such as Path Queries, Centralities, and Community Detection, to gain insights from the data. You are expected to present the insights gained, along with the queries used, in a PDF file. You must analyze the data as a data scientist and provide a detailed interpretation of the results.

You may consider the following questions to guide your analysis. However, feel free to explore the data in any way you prefer.

- What are the most influential papers in the dataset?
- What are the most influential authors in the dataset?
- Who are the top authors based on the number of papers they have published?
- Which authors have the highest average number of citations per paper?
- What are the most cited papers in the dataset?
- Are there any papers/authors that act as bridges between different clusters in the network (high betweenness centrality)?
- Do we have disconnected communities in the network? If yes, what are they?
- Do we have communities in the graph? If yes, what are they?
- Do we have citation cycles? Citation cycles are a set of papers that cite each other in a circular manner.

- What is the longest citation chain in the dataset?
- Are there any papers that act as critical connectors between highly cited sub-networks?
- Can you find the most common intermediate authors linking two separate research groups?
- How are different research areas connected through indirect citations?

3 Submission

- This is a **group assignment**. You can form groups of up to 2 members.
- You are required to provide a script for loading the dataset.
- You must submit a PDF file containing your analysis and queries. The maximum word limit for the analysis is **1500 words**. The structure of your document should be as follows:
  - Introduction(150 - 200 words): Briefly introduce your team and provide contribution breakdown.
  - Data Model and Loading(200 - 250 words): Describe your data model and highlight the loading process.
  - Analysis and Interpretation(400 - 650 words): Summarize the insights gained from the queries. You may introduce sub-sections to organize your findings. Also provide queries used to derive these insights.
  - Discussion(400 words): Reflect on the results and discuss the implications of your findings.

4 Rubric

1. Data Model and Loading (20 points): Evaluate correctness, efficiency, and justification of the chosen data model. Ensure proper data ingestion into Neo4j.
2. Graph Analytic Queries (25 points): Assess the variety, correctness, and complexity of queries (e.g., path queries, centralities, community detection).
3. Analysis and Interpretation (35 points): Judge the depth of insights derived, critical thinking, and justification of findings.
4. Presentation and Documentation (20 points): Evaluate clarity, completeness, and structure of the report, including visualizations and code readability.