CS 343: Graph Data Science
Spring 2024

Homework 2

Due: April 24, 2024 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 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 citation cycles? Citation cycles are a set of papers that cite each other in a circular manner.

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.

4 Rubric

1. Data Model and Loading (20 points)
2. Graph Analytic Queries (30 points)
3. Analysis and Interpretation (30 points)
4. Presentation and Documentation (20 points)