CS 343: Graph Data Science
Spring 2024

Homework 1

Due: March 4, 2024 at 09:00 AM at LMS

1 Introduction

You have to develop a graph property model for the given problem statements and provide Cypher queries to answer the questions. The model should be designed to capture the relationships between the entities in the problem statement, and the Cypher queries should be implemented to extract insights from the graph property model. The assignment is divided into two sections, each focusing on a different problem statement. The first section is about bibliographic data analysis, and the second section is about a university course registration system. The assignment is designed to assess your ability to design graph property models, implement Cypher queries, and derive insights from the data. The rubric for the assignment is provided at the end of the document.

2 Submission

This is a **group assignment**. You can form groups of up to 2 members. The deadline for submission is March 4, 2024 at 09:00 AM at LMS. Late submissions will not be accepted. Submit a single PDF file that contains the following sections for each questions along with the list of members.

1. **Model:** A graph property model that captures the relationships between the entities in the problem statement.
2. **Rationale:** Provide a rationale for the design choices made in the graph property model, explaining the considerations behind the clarity, completeness, and scalability of the model. Limit your response to 250 words.
3. **Queries:** Cypher queries that answer the questions mentioned in the question.

3 Questions

3.1 Bibliographic Data Analysis

You are a data scientist working for a renowned academic institution's research department. Your institution maintains a vast bibliographic database spanning multiple disciplines, containing information on research papers, authors, journals, and citations. The database encompasses decades of scholarly work, providing a rich source of data for academic analysis and insights generation.

The bibliographic data includes details such as:

- Research Papers: Each paper entry in the database includes metadata such as title, authors, publication year, abstract, keywords, and citation information.
- Authors: The database records information about authors, including their names, affiliations, and publication history.
- Journals: Information about journals, including titles, publishers, impact factors, and publication guidelines, is stored in the database.
- Citations: The database tracks citations between papers, indicating which papers cite or are cited by others, forming citation networks.

Your task as a data scientist is to use this wealth of bibliographic data to extract valuable insights into research trends, author collaborations, journal impact, and emerging topics within various academic fields. By designing a graph property model that accurately represents the relationships between papers, authors, journals, and citations, you aim to facilitate comprehensive analysis and visualization of the scholarly landscape. This model will enable researchers to uncover hidden patterns, track the evolution of research areas, and make informed decisions in their academic pursuits.

Also provide Cypher queries against your model to answer the following questions:

1. Identify top 10 authors who have collaborated on multiple papers based on the strength of their collaboration based on the number of joint publications.
2. Identify the top 10 most cited papers in the database and the number of citations they have received.
3. Identify the top 10 most influential authors based on the number of citations their papers have received.
4. Identify the top 10 authors with the most publications in the database.
5. Identify the top 10 pairs of papers that are most frequently co-cited together.
6. Calculate the diameter of the citation network, which represents the longest shortest path between any two papers in the network.
7. Calculate the degree distribution of the citation network, which represents the number of citations each paper has received.
8. Calculate the degree distribution of the co-authorship network, which represents the number of co-authors each author has collaborated with.
9. Identify the top 10 pairs of keywords that most frequently co-occur in the same paper.
10. Calculate the average number of citations per year for top 10 cited papesr in the database.

3.2 University Course Registration System

As part of a university's administrative team responsible for managing the course registration system. The system contains data about students, the courses they enroll in, and the prerequisites required for each course. Your objective is to develop a graph property model that captures the relationships between students, courses, and prerequisites, enabling efficient course planning and scheduling. This model can help identify course sequences that students must follow based on prerequisite constraints, identify requirements for a major or minor, and provide insights into students' academic pathways and progressions.

Also provide Cypher queries against your model to answer the following questions:

1. Identify the prerequisites for a course with course code 'CS101'.
2. List number of courses must be taken to fullfill requirements for each major.
3. List number of courses must be taken to fullfill requirements for each minor.
4. List number of courses must be taken before enrolling the course with course code 'CS431'.
5. Identify the courses that are taken together by at least 10 students. Return the courses and the number of students enrolled in each course.
6. Identify the courses that are offered in the same semester, indicating potential course scheduling patterns. Return the semester and the courses offered in that semester.
7. Identify the top 10 most popular courses based on the number of students enrolled.
8. Identify the courses which are taken together by students of Computer Science and Electrical Engineering major.
9. Identify the sequence of courses that a student with id "1214" has taken, indicating the order in which they have completed the courses.
10. List courses with the number of prerequisites assigned, indicating the complexity of the course requirements.

3.3 KSE 100 Companies

You are provided with a CSV file containing an edge list representing relationships between KSE 100 [1] companies based on the number of shared board members. Each row in the CSV file indicates the number of board members shared between two companies. For example, if row 1 indicates "Company A, Company B, 3", it means that Company A and Company B share 3 board members.

You have to perform the following tasks:

- **Model:** Design a graph property modl wthat captures the relationships between companies based on the number of shared board members. Also provide your rationale for the design choices made in the graph property model.

[1] This data was retrieved in the year 2016 for educational purposes. The data was curated under the supervision of Dr. Shahram Azhar.

- Provide cypher queries to answer the following questions:

    1. Provide import statement to load the data into the graph database.
    2. Determine the number of companies in the dataset.
    3. Determine the number of edges in the dataset.
    4. Identify the companies that share the highest number of board members.
    5. Determine the average number of shared board members between all pairs of companies.
    6. Find the length of the maximum path between any two companies.
    7. Determine companies have the highest number of board member associations with other companies along with the total number of board members they have.
    8. A triangle can be formed if a company shared board members with two other companies such that two companies also share board members with each other. Determine the number of companies that form a triangle.
    9. As the CEO of *Nestle Milkpak Ltd* you want to connect to someone who is connected to *Shell Pakistan Ltd*. Find the number of people you need to connect to reach *Shell Pakistan Ltd*.

4 Rubric

	Total Points	Model	Rationale	Queries
Q1	30	15	5	10
Q2	30	15	5	10
Q3	40	10	5	25

- **Model:** The graph property model accurately captures the relationships between the entities in the problem statement. The model incorporates all relevant entities, attributes, and relationships necessary for efficient course registration management. The model is scalable and adaptable to accommodate potential changes or expansions in the university's course catalog and academic programs.
- **Rationale:** The rationale for the design choices made in the graph property model is clearly explained and justified.
- **Queries:** The Cypher queries are correctly implemented and provide accurate insights into the data.