Applied Data Science

COMS30050/COMS30051/COMSM0055/COMSM0056 (2021-22 TB-2)

Lab 3: Exploring Graph Database Using Neo4j

Week #3 Lab builds on the lectures on Data Management and involves exploring Paradise Papers graph database using Neo4j Sandbox.

List of queries to perform during the lab:

1. Let us examine the overall size and shape of the Paradise Papers dataset. Write a query to find count of nodes of each type.

2. Let us now examine each node type individually. Write queries to retrieve one record of each node type and examine the attributes.

   Hint: (Sample answer) MATCH (o:Officer) RETURN o LIMIT 1

3. Cypher supports filtering similar to SQL using a WHERE clause. Write a query to find a family name from your country that shows up in the Paradise Papers.

4. Cypher supports aggregation operators similar to SQL. In Query:1, we examined the nodes and their size via counts. Other than counts, we can also find averages and maximums. Each node in the graph database is connected to other nodes. We now examine the degree distributions. Write a query to compute degree distribution including average and maximum to get an idea of how connected different pieces of the graph are.

   Hint: Use “size((n)--()) AS degree”

5. We notice that the data contains information on thousands of people or companies who play a role in an offshore company. Let us now examine the various relationship types. Write a query to show the count for each relationship type present in the dataset.

   Hint: We specify a relationship using ()-[r]->()

6. Let us now examine intermediaries. Write a query to find top 20 intermediaries sorted based on the number of connections to other entities.

   How many of these intermediaries have over 100 connections?

7. Now let us explore these intermediaries. Write a query to retrieve a list of entities registered by these intermediaries.

8. Examine the entities in a similar manner as we examined the intermediaries. Which entities are linked to most intermediaries?

9. Which officers are linked to most entities?

10. Let us now examine the distribution of the addresses by countries. Write a query to count addresses grouped by countries and list the top 10 countries.

    Hint: Address has an attribute country_codes

11. Let us also look at the jurisdiction information in the data. List entities based on the jurisdiction description.

    Hint: Entity has an attribute jurisdiction_description

12. What are the most popular offshore jurisdictions, by country of residence of the beneficiary or officer?

13. What are the most common offshore jurisdictions for officers with addresses in the United Kingdom?

14. ICIJ reported that Wilbur Ross, who served as the US Secretary of Commerce from 2017 to 2021, has connections to offshore companies. What are the jurisdictions of Ross' connected entities? What are the intermediaries linked to Wilbur Ross?

15. The Duchy of Lancaster – Queen Elizabeth II's private estate and portfolio also appears in the Paradise Papers dataset. Write a Cypher query for all two-degree connections to the Duchy:

16. Name of Rex Tillerson, who served as the U.S. Secretary of State, also appears in the Paradise Papers. Write a query to find the shortest path between the Duchy of Lancaster and Rex Tillerson.

    Hint: Try shortestPath and allShortestPaths functions available in Neo4j

17. How is Wilbur Ross connected to the Queen?

18. Are there any other duchies listed in the Paradise Papers?

If the lab tasks interest you, you can explore and examine connections of the people and organizations listed on the Wikipedia page about the Paradise Papers:
https://en.wikipedia.org/wiki/List_of_people_and_organisations_named_in_the_Paradise_Papers