

ECE_5DA04_TP – Big Graph Databases

Lab 1: Neo4j

Garima Gaur

5th December 2024

Submission deadline: Wednesday, 18 December, 2024 at 23:55:00 (Paris time)

1 Lab Organization

- **Neo4j and Cypher recall:** Please find the presentation under **Lab 1 slides** link at Moodle.
- **Neo4j Setup:** Download and install Neo4j Desktop following instructions at <https://neo4j.com/download/>.
- **Cypher Tutorial:** Available at <https://neo4j.com/docs/cypher-manual/current/introduction/>
- **Database creation:** Download the ICIJ dataset dump file (version 5) from <https://offshoreleaks.icij.org/pages/database>. After downloading, open the Neo4j Desktop application and create the database by following the instructions at <https://neo4j.com/docs/desktop-manual/current/operations/create-from-dump/>

2 Assignment

Dataset Overview: We are working with ICIJ Offshore Leaks dataset that captures relationships among people and companies spread across more than 200 countries and territories. These offshore entities are mentioned in the Panama Papers, Pandora Papers, Paradise Papers, Bahamas Leaks, and Offshore Leaks investigations. The data is modeled as a graph database with the schema shown in Figure 2.

Assignment Details: The following sections contain a thematic list of queries defined over the graph database which you have created using the Neo4j dump file of the ICIJ dataset. For this assignment, you have to write and execute these queries on your local machine and later submit the queries in a text file (detailed instructions in Section 3). If you have any questions about the lab or the course material, don't hesitate to ask them during the lab session, via Moodle, or email (don't wait until the last moment).

2.1 Schema Exploration

1. List all the properties of the node that represents SERGEY ROLDUGIN.
(Trivia: Sergey Roldugin is a close friend of Russian President Vladimir Putin.)
2. List all outgoing edges and corresponding neighbors of the node that represents SERGEY ROLDUGIN.
3. List names of any 100 officers who are mentioned in the Panama papers.
4. List all the relationships in the graph along with their frequency, i.e. total number of edges in the graph with that particular edge type.

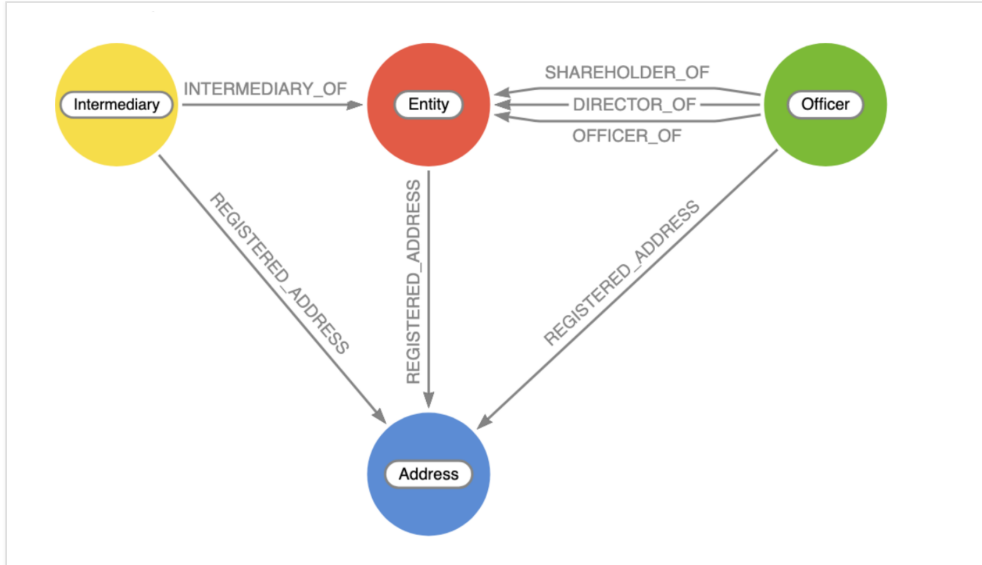


Figure 1: Schema of ICIJ dataset available at <https://offshoreleaks.icij.org/pages/about>

5. List all the node labels along with the frequency.
6. List all edge types along with the labels of the incident nodes. (This is effectively extracting the schema of the data graph)

2.2 Graph Navigational Queries

1. The multiple family members of the current President of Azerbaijan, Ilham Aliyev, are present in the dataset sourced from Offshore leaks and Panama Papers. Find all the paths up to length 4 between all the members of the Aliyev family.
2. Find the shortest paths between all the members of the Aliyev family.

2.3 Interesting Pattern Queries

1. For each individual, find the number of data sources she has appeared in.
2. Find the company that has the most number of off-shore directors.

3 Submission Instructions

You are required to submit a text file containing the Cypher queries. **Strictly follow the formatting guidelines** below while preparing your submission file otherwise your submission might not be graded:

- Each line contains a query ID and the corresponding Cypher query.
- Query ID and the query should be separated by TAB (“\t”).
- The query ID should follow the format *section number_query number*, for example, the first query in Section 2.1 about SERGEY ROLDUGIN will have query ID 2.1.1.

Please name your submission file as `LASTNAME.FIRSTNAME.txt` with first and last name in capital letters. Notice that the submission deadline is Wednesday, 18th December, 2024 at 23:55:00 (Paris time). Kindly adhere to the deadline.