Big Graph Processing Systems 2025

TP Component 2: Analysis

Christopher Spinrath

General Information

The goal of this component is to analyse the property graph of the International Consortium of Investigative Journalists, for short the ICIJ property graph. (For this course, the property graph was slightly modified: some properties were deleted, some property types were changed, and 200 problematic nodes/edges were removed.)

This component consists of ten tasks. A complete solution for each task consists of

- an openCypher query,
- the number of answers it returns,
- an explanation of the query, and
- (for some tasks) an interpretation of the query result.

For the query you can gain up to 2 points, for the number of answers 0.5 point, and for the explanation/interpretation 1-1.5 points. The total amount of points achievable with this component is 36 points.

Important

- **New:** Do not return lists as values in a column, except if explicitly ask for.
- Do not use `CALL ... YIELD` clauses to call procedures. They are Neo4j specific and are not part of openCypher/GQL.
- Make sure that all queries are free of syntax errors by running them, even if you make only small changes. A query with syntax errors will be graded with 0 points.
- A good explanation helps someone familiar with basic knowledge on openCypher to understand the (key) idea(s) of a query. "Reading" or repeating the query in natural language is, for example, not very helpful.
- Solve the tasks yourself, on your own. **Plagiarism is fraud.**

Task 1 **(3.5 points)**

a) Write an openCypher query that returns all distinct source ids of nodes and how often they occur.

b) How many answers does your query return?

c) Explain concisely how you came up with your query.

Task 2 **(3.5 points)**

a) Write an openCypher query that returns, for each pair (A, B) of different source IDs, the number of edges from nodes having source ID A to nodes having source ID B.
The query does **not** have to return pairs for which the there are no edges.

b) How many answers does your query return?

c) Explain concisely how you came up with your query.

Task 3 **(3.5 points)**

a) We say that an officer is **currently managing** an entity, if the officer is an officer of the entity and the entity is **not** labelled as struck-off.

Write an openCypher query that returns the names of the ten officers who currently manage the most entities, together with the number of entities they currently manage, in descending order.

b) How many answers does your query return?

c) Explain concisely how you came up with your query.

Task 4 **(3.5 points)**

a) Write an openCypher query that returns, for each year in which an entity or other identity was incorporated, the number of entities (or other other identities) were incorporated, ordered by year in ascending order.

The query should **filter out** rows, which are clearly wrong due to data in the property graph being "wrong" (typos, etc.).

Do not use string operations (instead lookup how to use the Date value type).

b) How many answers does your query return?

c) Explain concisely how you came up with your query.

Task 5 **(3.5 points)**

a) Write an openCypher query that returns, for each country, the country together with the number of entities operating in this country and were incorporated by the service provider "Mossack Fonseca" in 2006.

The query does not have to return rows for countries for which the number of entities is zero.

b) How many answers does your query return?

c) Explain concisely how you came up with your query.

Task 6 **(3.5 points)**

a) Write an openCypher query that returns, for each country, the number of intermediaries who have a registered address in this country, ordered by the number in descending order.

b) How many answers does your query return?

c) Explain concisely how you came up with your query.

Task 7 **(4 points)**

a) Write an openCypher query to **validate** the following constraint:
Every node labelled `StruckOff` has an inactivation or a struck-off date.

b) How many answers does your query return?

c) Explain concisely how you came up with your query.

d) State whether the constraint is satisfied in the ICIJ property graph and explain how you derived this from the result of your query. If the constraint is not satisfied, propose how the property graph could be changed to satisfy it. (Do not write a query, just explain your idea informally.)

Task 8 **(4 points)**

a) Write an openCypher query to **validate** the following constraint:
No node labelled with `StruckOff` has the status "Active".

b) How many answers does your query return?

c) Explain concisely how you came up with your query.

d) State whether the constraint is satisfied in the ICIJ property graph and explain how you derived this from the result of your query. If the constraint is not satisfied, propose how the property graph could be changed to satisfy it. (Do not write a query, just explain your idea informally.)

Task 9 **(3.5 points)**

a) Write an openCypher query to **validate** the following constraint.
For every entity, for which a registration date is known, the registration date and the incorporation date are identical.
Hint: The result of comparisons in Neo4j might be unexpected if one of the values being compared is null.

b) How many answers does your query return?

c) Explain concisely how you came up with your query.

d) State whether the constraint is satisfied in the ICIJ property graph and explain how you derived this from the result of your query.

Task 10 **(3.5 points)**

a) Write an openCypher query to **validate** the following constraint.
If an officer has an "officer of" relationship with a struck-off entity, and the relationship has an end date property, then the relationship has ended before (or at the same time) the entity was struck-off or/and deactivated.

b) How many answers does your query return?

c) Explain concisely how you came up with your query.