SEMANTIC DATA MANAGEMENT EXAM

13th of June 2022. *The exam will take **2 hours**. Answer each question in the provided space.* Answers out of such space will not be considered. Further, clearly read the instructions how to answer. Answers not following the format set might not be considered.

Name: …….

QUESTION 1. PROPERTY GRAPHS [3p]

We want to create a property graph modeling job offers and their applicants. The graph needs to store the positions offered (CTO, Director, Manager, Administration, …) by each company. About companies we need to know their name, and also the branch offices they have. Each office is located in a city and has a name. An office may offer a job for a certain position with a proposed salary.

We want to record the persons that apply to a job, their name, email and the city where they live.

The graph must be modeled in such a way that it also allows to efficiently retrieve all people that have applied to a certain position (regardless the company that offers it) besides following the general quality criteria (non-redundancy, maintainability, understandability, …).

a. Draw a sketch of the property graph (**draw the labels and attribute names**) you would propose to model this problem. Justify your decisions in case you add any redundant information, and make explicit any relevant assumptions required to properly interpret your solution.

Use the remaining of this page to provide your solution.

Given the property graph in Figure 1:

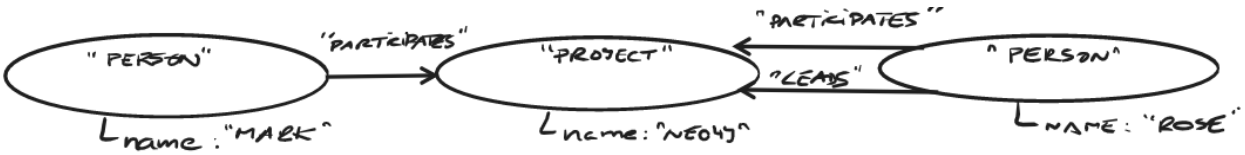


Figure 1.

And given the following Cypher query:

```
MATCH (p1:Person)-->(pr:Project{name:'Neo4j'})<--(p2:Person)
RETURN p1.name, p2.name
```

Answer the following questions:

b. Which is the result of this query on the previous graph?

c. Which would be the result of the same query under homomorphism semantics?

QUESTION 2. DISTRIBUTED GRAPH PROCESSING [3p]

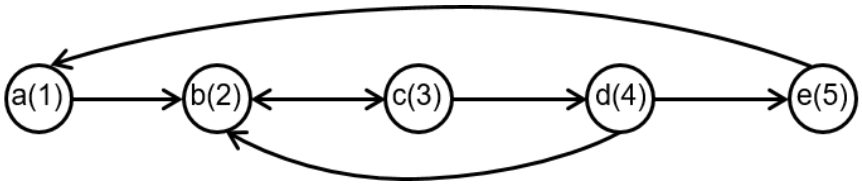


Figure 2.

For the connected graph in Figure 2, we want to compute the maximum value using the **theoretical TLAV framework**. For each vertex in the figure the letter is the identifier of the vertex denoted by v_i and the number in the brackets is the value of the vertex denoted by v_v. Assume the following kernel function:

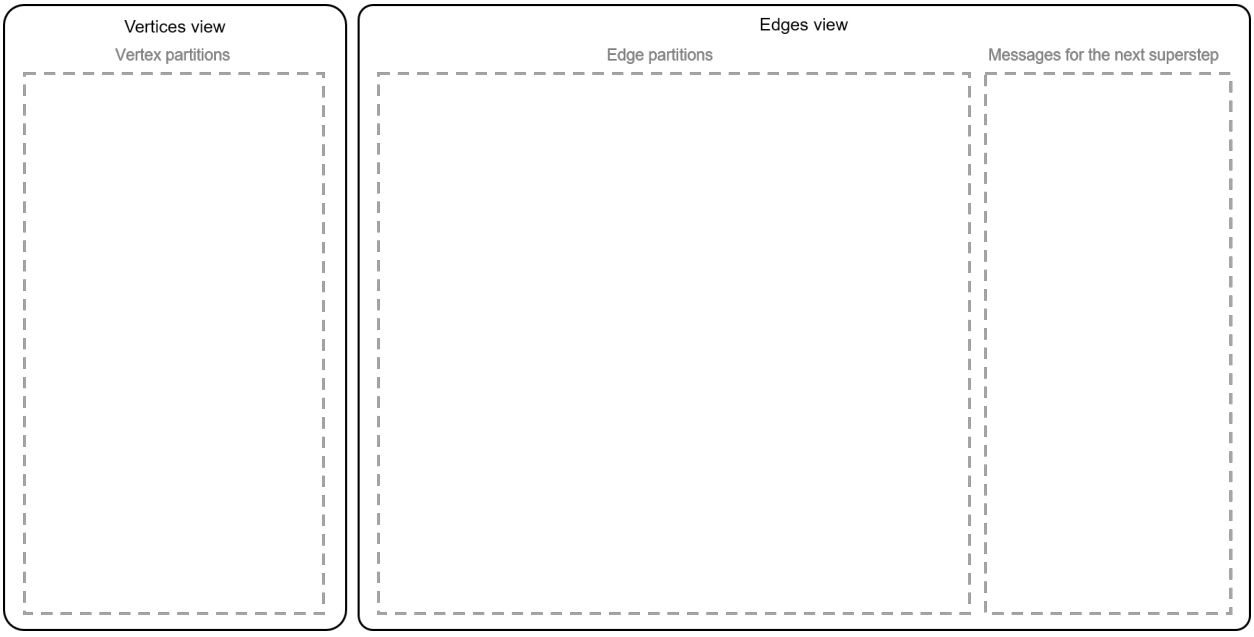
```
maxVal = max(receive(val))
if maxVal>v_v then
  v_v = maxVal
foreach e_vj ∈ E do
   send(v_v,j) //send the new maximum to the neighbouring vertices
```

And the following initialization (assume the default value at each node is 0):

```
foreach v_i ∈ V do
  send v_v to v_i //initial messages sent. Each vertex receives the value shown in Figure 2
```

a. Provide a graph distribution for the vertex and edge views and draw the partitions you consider in the figure below. Consider at least two partitions for the vertex and edge views, respectively.

b. Run the first superstep on top of the graph distribution you proposed and identify all the messages generated on the figure below. Represent the messages to the vertices as $a_{msg}(value)$, where *a* is the node receiving the message and *value* the value received. If applicable, consider combining the messages to reduce communication costs.



QUESTION 3. KNOWLEDGE GRAPHS AND DATA INTEGRATION [4p]

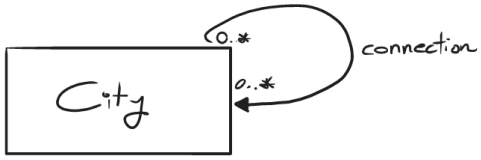


Figure 3.

a. Represent in Description Logics (i.e., correct TBOX axioms) all the constraints in Figure 3.

b. Define a new complex concept (i.e., a correct TBOX axiom) named *ConnectedCity* that represents all the cities with at least one connection to another city.

c. Write the formal semantics for the new concept *ConnectedCity*:

We aim at developing a data integration system to analyse data related to a company tracing mobility of people in a music festival. We aim at developing a **graph-based virtual data integration system** spanning several relevant data sources. Specifically, we choose to implement an **ontology-mediated querying system**. For this exercise, all instances are stored in other formats than graphs. Further, we will only consider two sources:

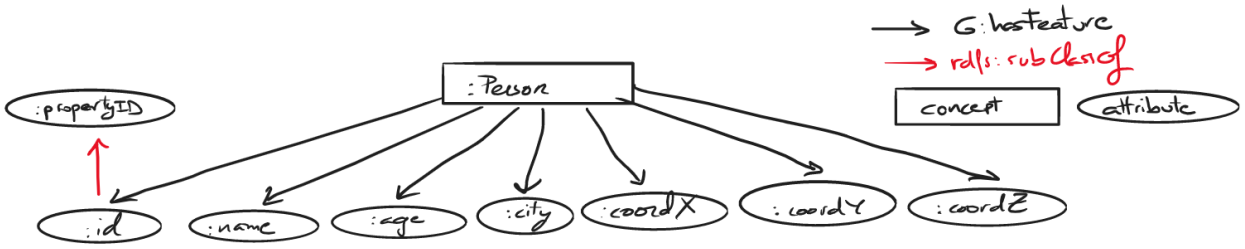
- A PostgreSQL database where the customer information is stored. Wrapper W1 exposes part of this data with the following tabular format:
  W1(idCard, name, age, city)
- A key-value storing traces about customers mobility. Wrapper W2 exposes the following tabular information:
  W2(user_id, coordX, coordY, coordZ)

In the next page, you can see the global (integration) schema generated in the company.

d. First, create the source graphs corresponding to the two wrappers identified. Follow the same notation and format as in the lecturers and draw them **below** the *source graphs* line.

e. Draw the **LAV mappings** between the global and local levels required to make the data integration system work. You can assume the *idCard* and *user_id* attributes from W1 and W2 do join. **Draw the mappings following the notation used in the lectures (i.e., a named graph and its corresponding *owl:sameAs* triples)**. If you can use two different colours for each wrapper, please, do so.

Global graph _____



Source graphs _____