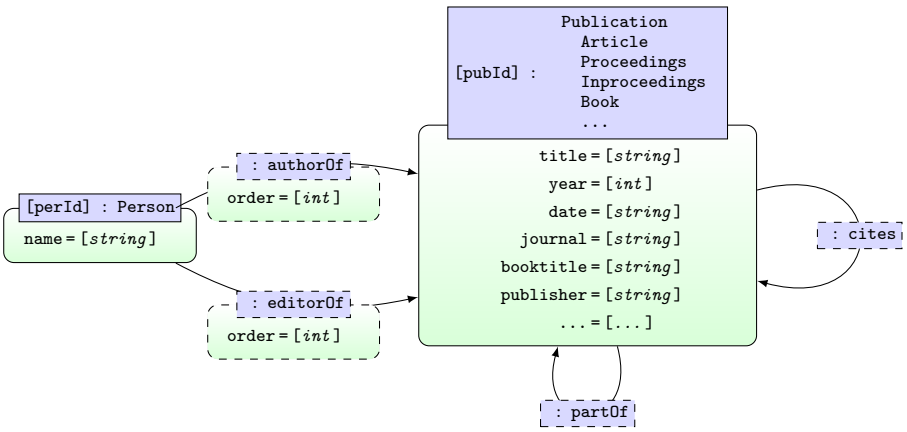
Lab 11 – Neo4j

CC5212-1 – June 26, 2024

Today we will query some graph data with Neo4j and the Cypher query language. More specifically, we will do queries over a bibliographical database of Computer Science papers called DBLP (see an example on this page for Sebastián Ferrada: `https://dblp.uni-trier.de/pers/hd/f/Ferrada:Sebasti=aacute=n`), which we've converted from the raw XML dump to a property graph with the following structure:



There are two main types of nodes (`Person` and `Publication`); the `Publication` node type also has several sub-types, including `Article` (a journal paper), `Proceedings` (a collection of conference papers), `Inproceedings` (a conference paper), `Book` (a monograph), and so forth (there are more we will not use, like `PhDThesis`, etc.). There are four types of relationships: `authorOf` (relating people to journal papers, conference papers, books, etc.), `editorOf` (relating people to proceedings), `cites` (relating publications to publications) and `partOf` (relating conference papers to the proceedings they are part of). Nodes have some attributes, including the name of the person; the title, year, date, journal (in case of journal paper), booktitle (in case of conference paper), publisher, ..., of the publication. The order attribute on the `authorOf` and `editorOf` relations indicate first author/editor (`order = 1`), second author/editor (`order = 2`), and so forth. The resulting DBLP graph has 8,561,079 nodes and 33,937,673 relations. The DBLP data are quite complete for papers and authors of academic Computer Science publications. However, the citations are unfortunately very incomplete; there are 225,148 such relations in the data, which are a small fraction of reality!

We have installed Neo4j on the master machine only (which the community edition supports). Sometimes the load of malformed queries can kill Neo4j. We will post in the forum some instructions to get it started again if necessary.

So let's open up and start to look at Neo4j

- First we get into the Cypher shell and load the database: log into the cluster as `uhadoop` and run `cypher-shell`. The username is `neo4j` and the password will be on the board.
- We can try a quick test query: `MATCH (n) RETURN count(*);`. Note the semi-colon at the end! What does the query return?[1]
- Okay, we will try some quick queries for some nodes:
  - Query for and return 10 nodes (`LIMIT 10`) of any type.
  - Query for and return 10 `:Person` nodes.
  - Query for and return the names of 10 `:Person` nodes.
  - Query for and return `:Person` nodes with the name `"Sebastián Ferrada"`.

[1] Neo4j indexes the value directly so it doesn't actually have to count all the nodes. You can try a similar query to count relationships but unfortunately it will take around 5 minutes, on a good day.

  - Query for and return any node with the name `"Sebastián Ferrada"` (...)

- Ouch, that last one took a while. The graph is quite large and we don't have an index for that last query, so Neo4j will end up scanning all nodes looking for one whose name matches the given value. To see the available indexes run `CALL db.indexes;`. While there is an index on `:Person(name)`, the last query did not restrict the node to be of type `:Person` so that index will not be used.

- We should keep in mind that person names are not unique!
  - Return the count of persons with the name `"Jorge Pérez"`.
  - Return the person IDs (`perId`) of persons named `"Jorge Pérez"` or `"Jorge Perez"`.
  - Return the count of persons whose name starts with `"Alber"`.

- Okay, let us try some queries on relationships to see what Claudio does in his office all day:
  - Find the titles of all publications authored by `"Claudio Gutiérrez"`.
  - Find the titles and year of journal articles authored by `"Claudio Gutiérrez"`, along with Claudio's position in the author list, ordered ascending by year and in case of a tie on year, by title, ascending.
  - Find the titles and year of journal articles where `"Claudio Gutiérrez"` is first author.
  - Find the titles and year of journal articles where `"Claudio Gutiérrez"` is not the first author.
  - Find the unique names of all coauthors of `"Claudio Gutiérrez"` on journal articles.
  - Count the unique coauthors of `"Claudio Gutiérrez"` on journal articles.
  - Count the unique coauthors of `"Claudio Gutiérrez"` on journal articles per year (return year and count).
  - Return the unique names of all coauthors of `"Claudio Gutiérrez"` that have worked with him on at least two journal articles (hint: it is easier not to use `count`).

- Next up, let's move towards running some path queries:
  - Return a shortest undirected path from `"Claudio Gutiérrez"` to `"Tim Berners-Lee"`.
  - Return the shortest undirected path length from `"Claudio Gutiérrez"` to `"Tim Berners-Lee"`.
  - Return all undirected shortest paths from `"Claudio Gutiérrez"` to `"Tim Berners-Lee"`.
  - Find all paths of length between 2 and 4 (inclusive) from `"Claudio Gutiérrez"` to `"Juan Alvarez"`.

- To exit the `cypher-shell` console, type `:exit`.

- SUBMIT to u-cursos a text file with all commands entered into the console.