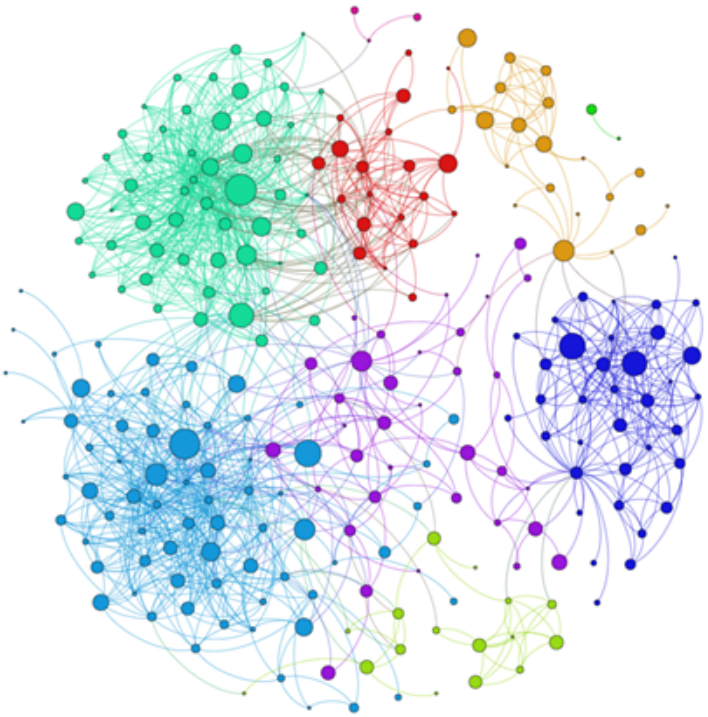
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

BUSINESS ANALYTICS
Master of Science

Homework II in Data Mining Datasets

Neo4j Graph database



Anastasios Theodorou (AM: p2822007)

Konstantina Georgiopoulou (AM: p2822004)

Date: July 4th, 2021

Introduction

In this assignment, we are going to use Neo4j in order to create a database, which contains authors, articles, journals and citations between articles and later we will implement some queries on it.. Neo4j is an open-source, NoSQL, native graph database that provides an ACID-compliant transactional backend for our applications. On this project we use a graph database, because this type of database is very popular to treat data that are connected. In this case, authors write articles which are being published in papers, and articles have citations. Consequently, Neo4j is the best to represent those relationships and easily take valuable information from the database. All the queries are written in Cypher, which is a declarative, SQL-inspired language.

Import the dataset into Neo4j & create graph model

In order to create the graph database, we import 3 ".csv" files, one for articles, one for authors and one for citations. In particular, the dataset contains 29555 articles with id, title, year and abstract, 15420 authors with names, 836 journals with names and 352807 citations among papers. We import the files directed to Neo4j and we create the appropriate relationships to query the database. Following are the commands that we used to import the files to the database and create the schema.

- First, we designed constraints on articles id, journals and authors names in order to be unique

```
CREATE CONSTRAINT ON (a:Article) ASSERT a.id IS UNIQUE;
CREATE CONSTRAINT ON (j:Journal) ASSERT j.jrn IS UNIQUE;
CREATE CONSTRAINT ON (au:Author) ASSERT au.name IS UNIQUE;
```

- Then, we imported the Articles' csv file and created a node:

```
LOAD CSV
FROM "file:///ArticleNodes.csv" AS line
CREATE (n:Article {id: toInteger(line[0]), title: line[1], year:
toInteger(line[2]), abstract: line[4]})
```

- Later, we also created node with the names of journals:

```
LOAD CSV
FROM "file:///ArticleNodes.csv" AS line
WITH line
WHERE line[3] IS NOT NULL
```

```
MERGE (n:Journal {jrn: line[3]})
ON MATCH SET n.id = toInteger(line[0])
```

- We made then a relationship between the 'journal' attribute and the articles

```
LOAD CSV
FROM "file:///ArticleNodes.csv" AS line
MATCH (a:Article), (j:Journal)
WHERE a.id = toInteger(line[0]) AND j.jrn = line[3]
CREATE (a) - [r:PUBLISHED] -> (j)
```

- We designed a node with authors

```
LOAD CSV
FROM "file:///AuthorNodes.csv" AS line
WITH line
WHERE line[1] IS NOT NULL
MERGE (n:Author {name: line[1]})
ON MATCH SET n.id = toInteger(line[0])
```

- We created a relationship between articles and who has written them:

```
LOAD CSV
FROM "file:///AuthorNodes.csv" AS line
MATCH (a1:Author), (a2:Article)
WHERE a1.name = line[1] AND a2.id = toInteger(line[0])
CREATE (a1) -[r:WRITES] -> (a2)
```

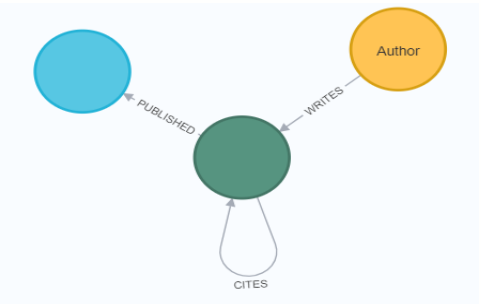
- Finally, we designed new node with articles that cite others:

```
LOAD CSV
FROM "file:///Citations.csv" AS line
FIELDTERMINATOR '\t'
MATCH (a1:Article), (a2:Article)
WHERE a1.id = toInteger(line[0]) AND a2.id = toInteger(line[1])
CREATE (a1) - [r:CITES] -> (a2)
```

Graph Schema

Here, we called the below command in order to visualize the database schema

```
call db.schema.visualization
```



Our graph consists of 3 nodes and 3 relations.

Nodes

Author: The author node contains:

- The name
- The id

Code: Match (n:Author) return properties(n)


```
{
  "name": "Paul S. Aspinwall",
  "id": 1001
}
```


Article: The article node contains:

- The id
- Title
- Year
- Abstract

Code: Match (n:Author) return properties(n)


```
{
  "year": 2000,
  "abstract": "  Various aspects of spaces of chiral
blocks are discussed. In particular-conjectures about the
dimensions of irreducible sub-bundles are reviewed
andtheir relation to symmetry breaking conformal boundary
conditions is outlined.",
  "id": 1005,
  "title": "Bundles of chiral blocks and boundary
conditions in CFT"
}
```


Journal: This node contains:

- The journal name
- The id

Code: Match (n:Author) return properties(n)


```
{
  "jrn": "Class.Quant.Grav.",
  "id": "9912200"
}
```


Edges

Writes: A directed edge from author to the article that he has written

Cites: A directed edge from an article to the articles that it has referenced

Published: A directed edge from the article to the journal that it is published

It is worth mentioning that before importing the '.csv' files we observed that the 'AuthorNodes.csv' file contained the below trash records:

Paper_ID	Author_Name
9503221	11 pages
9405048	115 pages
9411027	13 pages
9201047	18 pages
9705217	2) and C. A. S.
3121	2) and J. Wosiek
9603054	2) and J.A. Helayel-Neto
9709177	2) and Stephen B.
9408151	7 pages
9408150	10 pages

We did not proceed in cleaning the above lines, because they did not correspond to any article. So, they were useless anyway.

Querying the database

Q1: Which are the top 5 authors with the most citations (from other papers). Return author names and number of citations.

Query:

```
MATCH (x:Author) - [w:WRITES] -> (n:Article) <- [c:CITES] - (a:Article)

RETURN x.name AS author, COUNT(c) AS citations

ORDER BY citations DESC

LIMIT 5
```

Result:

	author	citations
1	"Edward Witten"	15681
2	"Ashoke Sen"	7120
3	"Michael R. Douglas"	5577
4	"A.A. Tseytlin"	5288
5	"Joseph Polchinski"	5267

Q2: Which are the top 5 authors with the most collaborations (with different authors). Return author names and number of collaborations.

Query:

```
MATCH (a1:Author)-[r1:WRITES]-> (ar:Article) <- [r2:WRITES] - (a2:Author)

WHERE a1.name <> a2.name
```

```
RETURN a1.name AS author_name, count(distinct(a2.name)) AS counter

ORDER BY counter desc

LIMIT 5
```

Result:

	author_name	counter
1	"C.N. Pope"	50
2	"S. Ferrara"	46
3	"M. Schweda"	46
4	"C. Vafa"	45
5	"H. Lu"	45

Q3: Which is the author who has written the most papers without collaborations? Return author name and number of papers.

Query:

```
MATCH (author:Author) - [w:WRITES] -> (article:Article)

OPTIONAL MATCH (collaborator:Author) - [w1:WRITES] -> (article:Article)

WITH author, COUNT(article) AS articles_count, COUNT(DISTINCT collaborator)  AS
collaborators_count

WHERE collaborators_count = 1

RETURN author.name, articles_count

ORDER BY articles_count DESC

LIMIT 1
```

Result:

	author.name	articles_count
1	"J. Kluson"	18

Q4: Which author published the most papers in 2001? Return author name and number of papers.

Query:

```
MATCH (a:Author) - [r:WRITES] -> (a1:Article) - [r1:PUBLISHED] -> (j:Journal)

WHERE a1.year = 2001

RETURN a.name as author, count(*) as paper_published

ORDER BY paper_published DESC

LIMIT 1
```

Result:

	author	paper_published
1	"Sergei D. Odintsov"	13

Q5: Which is the journal with the most papers about "gravity" (derived only from the paper title) in 1998. Return name of journal and number of papers.

For searching text in Neo4j first we create a fulltext index on the title of the article which will allow us to search for the existence of the keyword "gravity" in all the titles of the journals much more efficiently.

Query:

```
//fulltext index

call db.index.fulltext.createNodeIndex("articleTitle", ["Article"], ["title"])

//create query

 CALL db.index.fulltext.queryNodes("articleTitle", "gravity")

YIELD node AS article

MATCH (author:Author) - [w:WRITES] -> (article:Article)- [p:PUBLISHED]-> (j:Journal)

WHERE article.year = 1998

RETURN j.jrn as title, COUNT(article) AS papers

ORDER BY papers DESC

LIMIT 1
```

Result:

title	papers
"Nucl.Phys."	26

Q6: Which are the top 5 papers with the most citations? Return paper title and number of citations.

Query:

```
MATCH ()-[r:CITES]->(a:Article)

RETURN a.title AS title, count(a) AS counter

ORDER BY counter DESC

LIMIT 5
```

Result:

	title	counter
1	"The Large N Limit of Superconformal Field Theories and Supergravity"	2414
2	"Anti De Sitter Space And Holography"	1775
3	"Gauge Theory Correlators from Non-Critical String Theory"	1641
4	"Monopole Condensation And Confinement In N=2 Supersymmetric Yang-Mills"	1299
5	"M Theory As A Matrix Model: A Conjecture"	1199

Q7: Which were the papers that use "holography" and "anti de sitter" (derived only from the paper abstract). Return authors and title.

And in this case, we create a fulltext index on the abstract of the articles and we search for keywords holography and anti de sitter to exist.

Query:

```
//create fulltext index

CALL db.index.fulltext.createNodeIndex("articleAbstract", ["Article"], [ "abstract"])

//create query

CALL db.index.fulltext.queryNodes("articleAbstract", "holography AND anti de sitter")

 YIELD node AS article

 MATCH (author:Author) - [w:WRITES] -> (article:Article)

 RETURN article.title as title, author.name as authors
```

Result:

Following are some indicative results from 26 records

"title"	"authors"
"Relating Friedmann equation to Cardy formula in universes with"	"Elcio Abdalla"
"Relating Friedmann equation to Cardy formula in universes with"	"Ru-Keng Su"
"Relating Friedmann equation to Cardy formula in universes with"	"Bin Wang"
"A new holographic limit of AdS5 x S5"	"Machiko Hatsuda"
"A new holographic limit of AdS5 x S5"	"Warren Siegel"
"Decomposing Quantum Fields on Branes"	"R. Schaeffer"
"Decomposing Quantum Fields on Branes"	"J. Bros"
"Decomposing Quantum Fields on Branes"	"M. Bertola"
"Decomposing Quantum Fields on Branes"	"U. Moschella"
"Decomposing Quantum Fields on Branes"	"V. Gorini"
"Exploring de Sitter Space and Holography"	"Vijay Balasubramanian"

Q8: Find the shortest path between 'C.N. Pope' and 'M. Schweda' authors (use any type of edges). Return the path and the length of the path. Comment about the type of nodes and edges of the path.

Query:

```
MATCH (a:Author{name:'C.N. Pope'}), (b:Author{name:'M. Schweda'}), p =
shortestPath((a)-[*]-(b))

WHERE a <> b

RETURN a.name as From_Node, [n in nodes(p) | labels(n)] AS ShortestPath_Nodes,
b.name AS To_Node, length(p) as Length

ORDER BY Length ASC
```

Result:

	From_Node	ShortestPath_Nodes	To_Node	Length
1	"C.N. Pope"	[["Author"], ["Article"], ["Journal"], ["Article"], ["Author"]]	"M. Schweda"	4

As we can observe in the above output the system for finding the shortest path between the authors we are interested in, it goes through an article that the specific author ('C.N. Pope') has written, then to a journal that this article has been published, later to another article that was published in the same journal and has been written by 'M. Schweda'. So, the path has length 4 and we see that by default the system is using all the available nodes.

Q9: Run again the previous query (8) but now use only edges between authors and papers. Comment about the type of nodes and edges of the path. Compare the results with query 8.

Query:

```
MATCH (a:Author{name:'C.N. Pope'}), (b:Author{name:'M. Schweda'}), p =
shortestPath((a)-[*]-(b))

WHERE a <> b AND NONE(n in nodes(p) WHERE n : Journal)

RETURN a.name as From_Node, [n in nodes(p) | labels(n)] AS ShortestPath_Nodes,
b.name AS To_Node, length(p) as Length

ORDER BY Length ASC
```

Result:

	From_Node	ShortestPath_Nodes	To_Node	Length
1	"C.N. Pope"	[["Author"], ["Article"], ["Article"], ["Article"], ["Author"]]	"M. Schweda"	4

Here, we can observe that for finding the shortest path between the two authors taken from the previous query, the system goes through an article that the specific author ('C.N. Pope') has written, then to an article that the previous one cites to, later to another article that the previous one cites and

in parallel has been written by 'M. Schweda'. So, the path again has length 4, but now the system uses citations in order to go from one author to the other.

Q10: Find all authors with shortest path lengths > 25 from author 'Edward Witten'. The shortest paths will be calculated only on edges between authors and articles. Return author name, the length and the paper titles for each path.

Query:

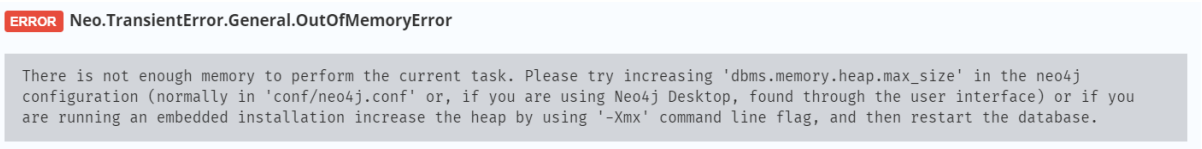
Executing the following query we receive a Memory error.

```
MATCH p = ShortestPath((c:Author{name:'Edward Witten'})-[*]-(f:Author))

Where f<>c AND length(p) > 25 AND NONE(n in nodes(p) WHERE n : Journal)

RETURN f.name as author ,length(p) as length, [n in nodes(p) where n.title is not null |
n.title] as title LIMIT 10
```

Result:



Putting the length path until 15, the query executed with the following results.

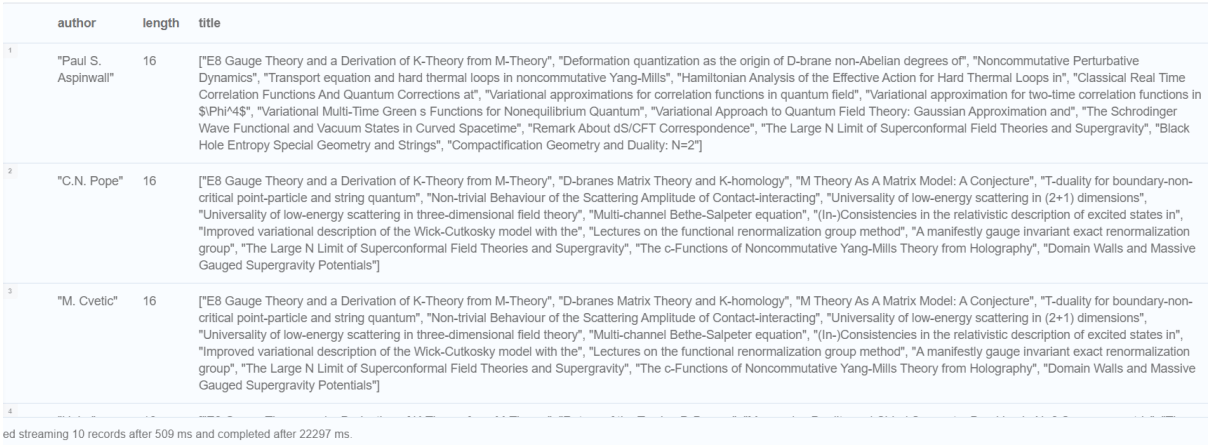
Query:

```
MATCH p = ShortestPath((c:Author{name:'Edward Witten'})-[*]-(f:Author))

Where f<>c AND length(p) > 15 AND NONE(n in nodes(p) WHERE n : Journal)

RETURN f.name as author ,length(p) as length, [n in nodes(p) where n.title is not null |
n.title] as title LIMIT 10
```

Indicative Results:



	author	length	title
1	"Paul S. Aspinwall"	16	["E8 Gauge Theory and a Derivation of K-Theory from M-Theory", "Deformation quantization as the origin of D-brane non-Abelian degrees of", "Noncommutative Perturbative Dynamics", "Transport equation and hard thermal loops in noncommutative Yang-Mills", "Hamiltonian Analysis of the Effective Action for Hard Thermal Loops in", "Classical Real Time Correlation Functions And Quantum Corrections at", "Variational approximations for correlation functions in quantum field", "Variational approximation for two-time correlation functions in Φ^4", "Variational Multi-Time Green s Functions for Nonequilibrium Quantum", "Variational Approach to Quantum Field Theory: Gaussian Approximation and", "The Schrodinger Wave Functional and Vacuum States in Curved Spacetime", "Remark About dS/CFT Correspondence", "The Large N Limit of Superconformal Field Theories and Supergravity", "Black Hole Entropy Special Geometry and Strings", "Compactification Geometry and Duality: N=2"]
2	"C.N. Pope"	16	["E8 Gauge Theory and a Derivation of K-Theory from M-Theory", "D-branes Matrix Theory and K-homology", "M Theory As A Matrix Model: A Conjecture", "T-duality for boundary-non-critical point-particle and string quantum", "Non-trivial Behaviour of the Scattering Amplitude of Contact-interacting", "Universality of low-energy scattering in (2+1) dimensions", "Universality of low-energy scattering in three-dimensional field theory", "Multi-channel Bethe-Salpeter equation", "(In-)Consistencies in the relativistic description of excited states in", "Improved variational description of the Wick-Cutkosky model with the", "Lectures on the functional renormalization group method", "A manifestly gauge invariant exact renormalization group", "The Large N Limit of Superconformal Field Theories and Supergravity", "The c-Functions of Noncommutative Yang-Mills Theory from Holography", "Domain Walls and Massive Gauged Supergravity Potentials"]
3	"M. Cvetic"	16	["E8 Gauge Theory and a Derivation of K-Theory from M-Theory", "D-branes Matrix Theory and K-homology", "M Theory As A Matrix Model: A Conjecture", "T-duality for boundary-non-critical point-particle and string quantum", "Non-trivial Behaviour of the Scattering Amplitude of Contact-interacting", "Universality of low-energy scattering in (2+1) dimensions", "Universality of low-energy scattering in three-dimensional field theory", "Multi-channel Bethe-Salpeter equation", "(In-)Consistencies in the relativistic description of excited states in", "Improved variational description of the Wick-Cutkosky model with the", "Lectures on the functional renormalization group method", "A manifestly gauge invariant exact renormalization group", "The Large N Limit of Superconformal Field Theories and Supergravity", "The c-Functions of Noncommutative Yang-Mills Theory from Holography", "Domain Walls and Massive Gauged Supergravity Potentials"]
4			

ed streaming 10 records after 509 ms and completed after 22297 ms.

Even after trying the below query, which filters the results and accelerates the whole procedure the outcome is null again:

```
MATCH (f:Author), p = ShortestPath((c:Author{name:'Edward
Witten'})-[:AUTHOR*]-(f:Author))

Where f<>c

WITH c.name as fromNode,

f.name as toNode,[n in nodes(p) | n.title] AS SortestPath,

length(p) as Length

WHERE Length >25

RETURN fromNode, toNode,Length, SortestPath

ORDER BY Length DESC
```