Mining Big Datasets – Assignment II

MSc in Business Analytics

Athens University of Economics and Business

July 2021

Students:

Mentakis Dimitrios p2822024

Vlassi Georgia p2822001

Vretteas Stylianos p2822003

Professor: Y. Kotidis

Assistant: I. Filippidou

Dataset

For this assignment we are given a subset of the high energy physics theory citation network, which contains authors, articles, journals and citations between articles. 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.

Below a short description of the files is presented:

- **ArticleNodes.csv**: Contains information regarding the nodes of articles. The properties included are the id of the article, its title, the publishing year, the journal where it was published and an abstract of the article.
- **AuthorNodes.csv**: Contains the id of the article from the previous file and the name of the author(s).
- **Citations.csv**: Contains information about citations between articles, were an articleId,--[Cites]-> articleId.

Observing the data, the following solutions will be implemented for the creation of the graph model. From the first file we will create the Article and the Journal nodes. From the second file, which contains information about the relationship between article and author, we will create the Author nodes and the above relationship named with the alias WRITTEN_BY. The Article and Authors nodes will be connected with the relationship named PUBLISHED_IN. Finally, the third file will be used to describe the citations between the articles, The final relationship is named CITES.

Importing the dataset into Neo4j

To import the files and create our graph model, the Neo4j Desktop is used. Initially, we create a new project named ‘Article-Author-Citations’ and inside this a local database is added with name ‘MyGraph’, as shown in Figure 1.

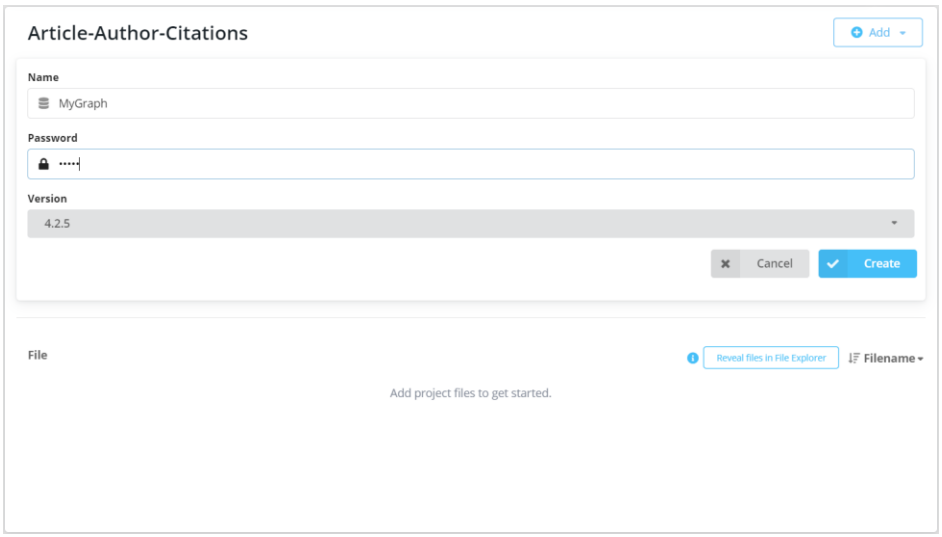


Figure 1 Create new project with a local DBMS

Starting our database, the files mentioned should be imported. In Figure 2 we can see the way to import them.

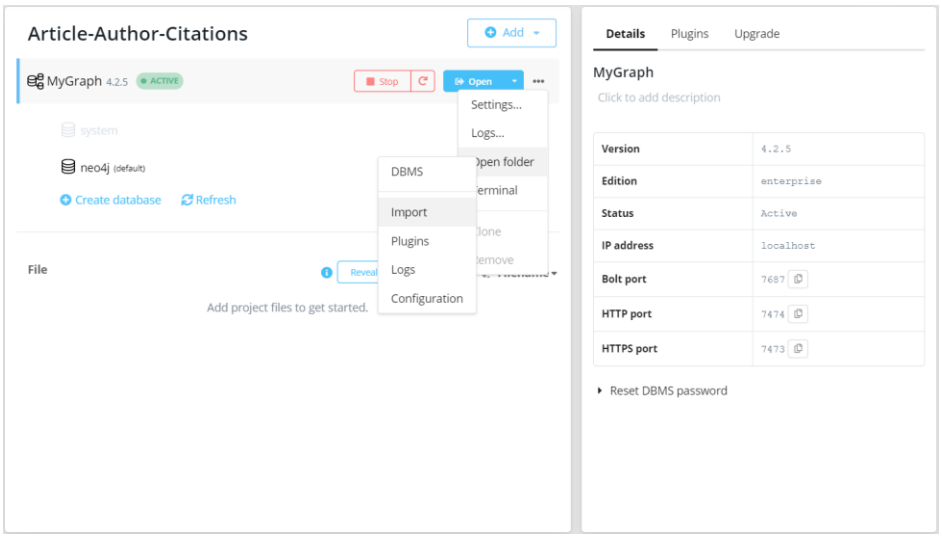
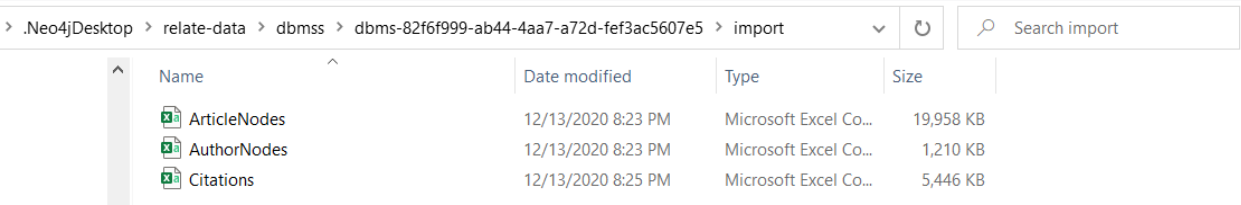


Figure 2 Import the files into the DBMS



After importing all csv files, we open the Neo4j Browser, as shown in Figure 3, in order to create the appropriate entities and relationships. Alongside the relevant labels, types and properties will be assigned.

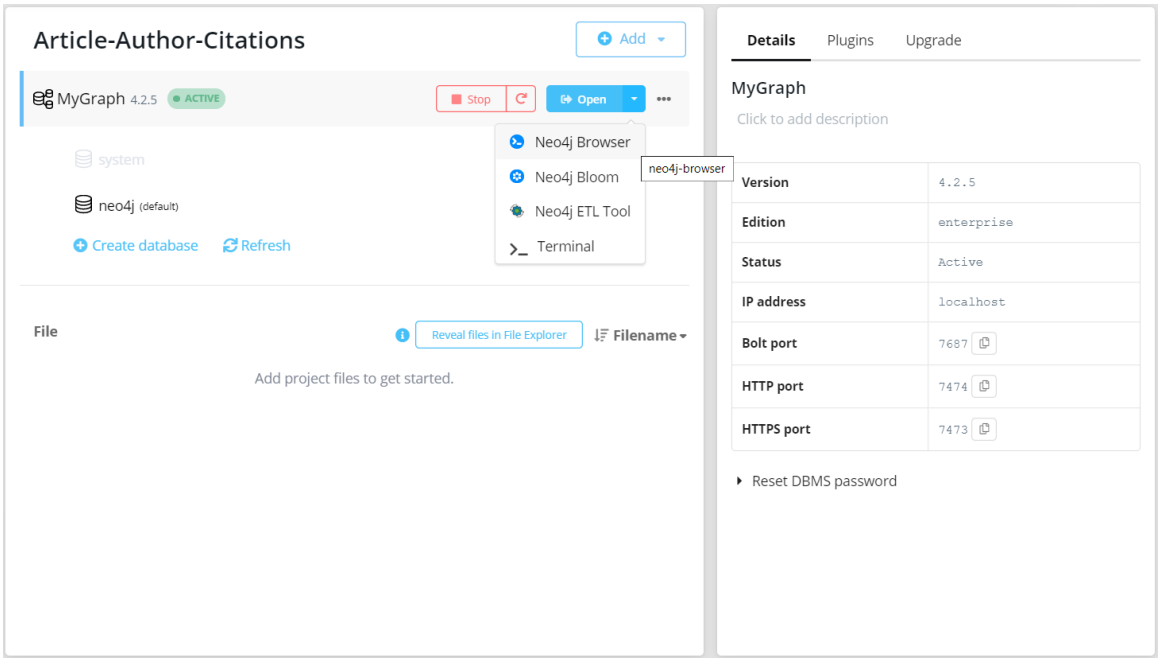


Figure 3 Open the Neo4j Browser

Property graph model

In Figure 4 is represented the required graph model, which contains the nodes with the relationships between them.

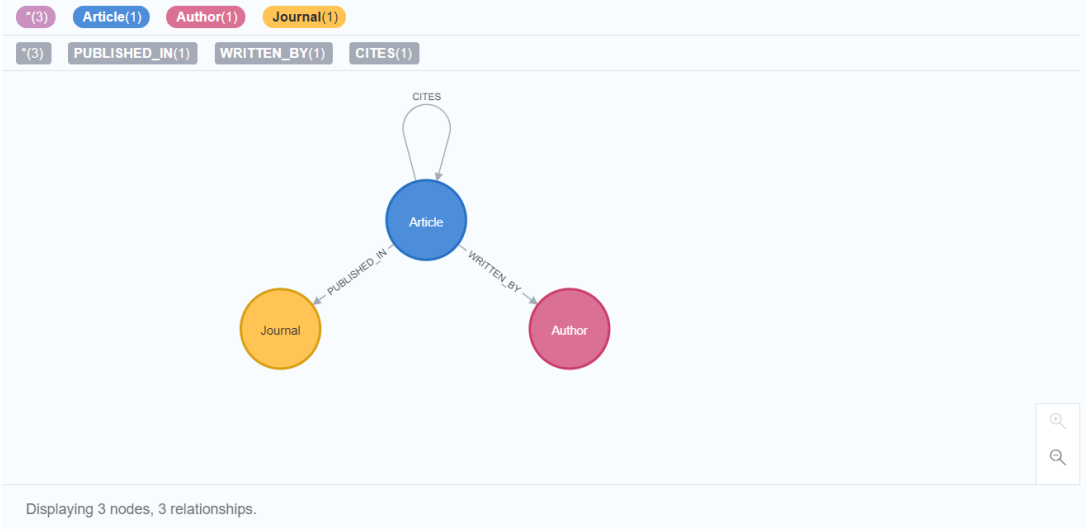


Figure 4 Graph model

To create the Article nodes we initially create a constraint in the id property, so as to have only the distinct articles. After executing the below 29555 nodes are created :

```
neo4j$ CREATE CONSTRAINT ON (article:Article) ASSERT article.id IS UNIQUE
```

```
:auto USING PERIODIC COMMIT 500 LOAD CSV
FROM "file:///ArticleNodes.csv" AS Articles
FIELDTERMINATOR ','
CREATE (a:Article{id:toInteger(Articles[0]), title:Articles[1], year:Articles[2], abstract:Articles[4]})
```

To create the Journal nodes, a constraint should be set to avoid duplicates. Each Journal node match with a specific id of the Article nodes. After the execution of the below 836 nodes are created.

```
neo4j$ CREATE CONSTRAINT ON (j:Journal) ASSERT j.journal IS UNIQUE;
```

```
:auto USING PERIODIC COMMIT 500
LOAD CSV FROM "file:///ArticleNodes.csv" AS journals
WITH journals
WHERE journals[3] IS NOT NULL
MERGE (n:Journal {journal: journals[3]})
ON MATCH SET n.id = toInteger(journals[0])
```

Having created the Article and Journal nodes, we set up the PUBLISHED_IN relationship among them.

```
:auto USING PERIODIC COMMIT 500
LOAD CSV FROM "file:///ArticleNodes.csv" AS row
MATCH (a:Article), (j:Journal)
WHERE a.id = toInteger(row[0]) AND j.journal = row[3]
CREATE (a) - [r:PUBLISHED_IN] → (j)
```

To create the Author node a new constraint is created and from the corresponding file we keep only the attribute with the name of the author. In total, 15420 nodes are created.

```
neo4j$ CREATE CONSTRAINT ON (auth:Author) ASSERT auth.name IS UNIQUE;
```

```
:auto USING PERIODIC COMMIT 500
LOAD CSV FROM "file:///AuthorNodes.csv" AS row
WITH row
WHERE row[1] IS NOT NULL
MERGE (n:Author {name: row[1]})
ON MATCH SET n.id = toInteger(row[0])
```

Having created the Author nodes, we set up the WRITTEN_BY relationship among article and author.

```
:auto USING PERIODIC COMMIT 500
LOAD CSV FROM "file:///AuthorNodes.csv" AS relationships
FIELDTERMINATOR ','
MATCH (article:Article {id: toInteger(relationships[0])})
MATCH (author:Author {name: relationships[1]})
MERGE (article)-[:WRITTEN_BY]→(author)
```

Finally, we create the relationships between articles, where article CITES to another article. There are 352807 relationships. In Figure 5 we can see a subgraph representing all the above executions.

```
:auto USING PERIODIC COMMIT 500
LOAD CSV FROM "file:///Citations.csv" AS citations
FIELDTERMINATOR '\t'
MATCH (article1:Article {id: toInteger(citations[0])})
MATCH (article2:Article {id: toInteger(citations[1])})
MERGE (article1)-[:CITES]→(article2)
```

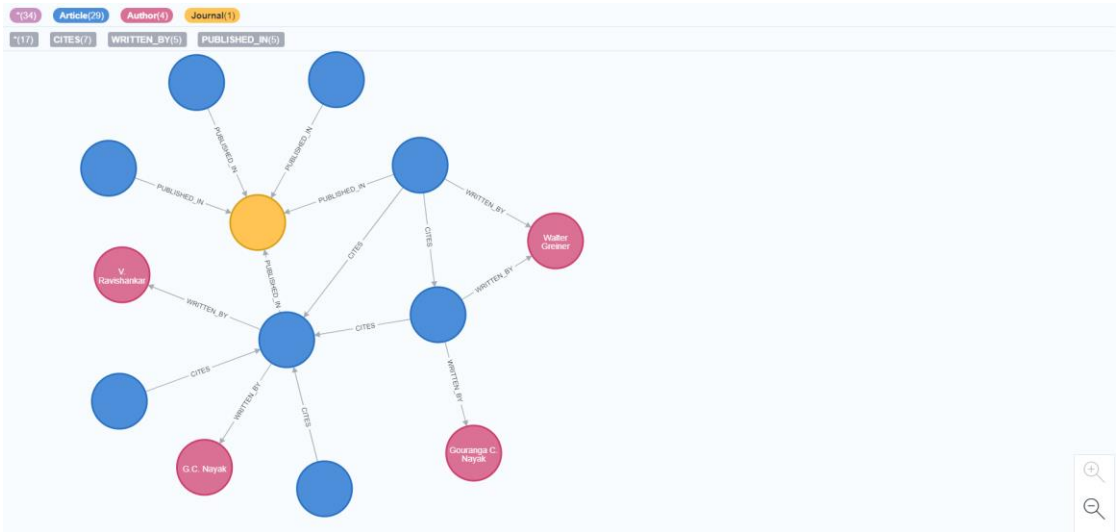


Figure 5 Subgraph of the model

Querying the database

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

Query

```
MATCH (ar1:Article)-[r:CITES]->(ar2:Article),(ar2)-[:WRITTEN_BY]->(au:Author)
RETURN au.name AS Author, COUNT(r) AS Number_of_Citations
ORDER BY Number_of_Citations
DESC LIMIT 5
```

Results

"Author"	"Number_of_Citations"
"Edward Witten"	15681
"Ashoke Sen"	7120
"Michael R. Douglas"	5577
"A.A. Tseytlin"	5288
"Joseph Polchinski"	5267

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

Query

```
MATCH(ar:Article)-[:WRITTEN_BY]->(au:Author), (ar)-[:WRITTEN_BY]->(au2:Author)
WHERE au.name<>au2.name
RETURN au.name AS Author, COUNT(DISTINCT au2) AS Number_of_Collaborations
ORDER BY Number_of_Collaborations
DESC LIMIT 5
```

Results

"Author"	"Number_of_Collaborations"
"C.N. Pope"	50
"M. Schweda"	46
"S. Ferrara"	46
"H. Lu"	45
"C. Vafa"	45

3. Which is the author who has wrote the most papers without collaborations. Return author name and number of papers.

Query

```
MATCH (ar:Article)-[w:WRITTEN_BY]->(au:Author)
MATCH (ar)-[w2:WRITTEN_BY]->(au2:Author)
WITH au, COUNT(ar) AS  Number_of_Papers, COUNT(DISTINCT au2) as
Collaborators_counter
WHERE Collaborators_counter = 1
RETURN au.name AS Author, Number_of_Papers
ORDER BY Number_of_Papers
DESC LIMIT 1
```

Results

"Author"	"Number_of_Papers"
"J. Kluson"	18

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

Query

```
MATCH (ar:Article)-[w:WRITTEN_BY]->(au:Author), (ar)-[p:PUBLISHED_IN]-
>(j:Journal)
WHERE ar.year = '2001'
RETURN au.name AS Author, COUNT(ar) AS Number_of_Papers
ORDER BY Number_of_Papers
DESC LIMIT 1
```

Results

"Author"	"Number_of_Papers"
"Sergei D. Odintsov"	13

5. 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.

Query

```
MATCH (ar:Article)-[p:PUBLISHED_IN]->(j:Journal)
WHERE ar.year ='1998'
AND toLower(ar.title)CONTAINS"gravity"
RETURN j.journal AS Journal, COUNT(ar) AS Number_of_Papers
ORDER BY Number_of_Papers
DESC LIMIT 1
```

Results

"Journal"	"Number_of_Papers"
"Nucl.Phys."	34

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

Query

```
MATCH (ar1:Article)-[r:CITES]->(ar2:Article)
RETURN ar2.title AS Paper_Title, COUNT(r) AS Number_of_Citations
ORDER BY Number_of_Citations
DESC LIMIT 5
```

Results

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

7. Which were the papers that use “holography” and “anti de sitter” (derived only from the paper abstract). Return authors and title.

Query

```
MATCH (ar:Article)-[w:WRITTEN_BY]->(au:Author)
WHERE toLower(ar.abstract) CONTAINS "holography"
AND toLower(ar.abstract) CONTAINS "anti de sitter"
RETURN au.name AS Author, ar.title AS Title
```

Results

"Author"	"Title"
"Bin Wang"	"Relating Friedmann equation to Cardy formula in universes with"
"Ru-Keng Su"	"Relating Friedmann equation to Cardy formula in universes with"
"Elcio Abdalla"	"Relating Friedmann equation to Cardy formula in universes with"
"Seungjoon Hyun"	"Background geometry of DLCQ M theory on a p-torus and holography"
"Youngjai Kiem"	"Background geometry of DLCQ M theory on a p-torus and holography"

8. 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 p = shortestPath((au:Author{name:'C.N. Pope'})-[*]-(au2:Author{name:'M. Schweda'}))
RETURN [n in nodes(p)] AS Path, length(p) AS Path_Length
```

Results

"Path"	"Path_Length"
[{"name":"C.N. Pope","id":9910252},{"id":9910252,"abstract":" We construct the complete and explicit non-linear Kaluza-Klein Ansatz forderiving the bosonic sector of the standard N=4 SO(4) gauged four-dimensionalsupergravity from the reduction of D=11 supergravity on S^7. This provides away of interpreting all bosonic solutions of the four-dimensional gauged theorys exact solutions in eleven-dimensional supergravity. We discuss certainlimiting forms of the Kaluza-Klein reduction- and compare them with relatedforms in the Freedman-Schwarz N=4 SU(2)xSU(2) gauged theory. This leads us tothe result that the Freedman-Schwarz model is in fact a singular limiting caseof the standard SO(4) gauged supergravity. We show that in this limit- ourAnsatz for getting the SO(4) gauged theory as an S^7 reduction from D=11 indeedreduces to an S^3 x S^3 reduction from D=10- which makes contact with previousresults in the literature. We also show that there is no distinction to be madebetween having equal or unequal values for the gauge coupling constants g and$\tilde g$ of the two SU(2) gauge-group factors in the standard N=4 SO(4)gauged supergravity- whilst by contrast the ratio of g to $\tilde g$ is anon-trivial parameter of the Freedman-Schwarz model.","title":"Four-dimensional N=4 SO(4) Gauged Supergravity from D=11","year":"1999"},{"journal":"Nucl.Phys.","id":9912285},{"id":9904204,"abstract":" We study the ultraviolet and the infrared behavior of 2D topologicalBF-Theory coupled to vector and scalar fields. This model is equivalent to 2Dgravity coupled to topological matter. Using techniques of the algebraicrenormalization program we show that this model is anomaly free and ultravioletas well as infrared finite at all orders of perturbation theory.","title":"Finiteness of 2D Topological BF-Theory with Matter Coupling","year":"1999"},{"name":"M. Schweda","id":9911127}]	4

We observe that the shortest path between authors ‘C.N. Pope’ and ‘M. Schweda’ has length equal to 4. This path consists of nodes and edges (relationships) of all types. In Figure 6 below we visualize this path and we conclude that the authors have written two different articles, which have been published in the same journal.

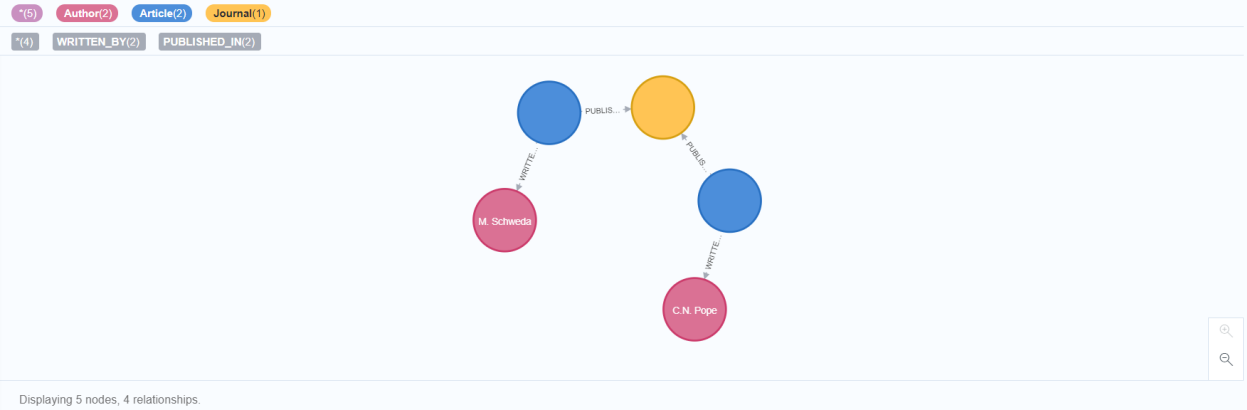


Figure 6 Shortest path between authors ‘C.N. Pope’ and ‘M. Schweda’

9. 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 p = shortestPath((au:Author{name:'C.N. Pope'})-[w:WRITTEN_BY*]-
(au2:Author{name:'M. Schweda'}))
RETURN [n in nodes(p)] AS Path, length(p) AS Path_Length
```

Results

"Path"	"Path_Length"
[{"name":"C.N. Pope","id":9910252},{"id":304132,"abstract":" We obtain a solution to eleven-dimensional supergravity that consists of M2-branes embedded in a dielectric distribution of M5-branes. Contrary tonormal expectations- this solution has maximal supersymmetry for a branesolution (i.e. sixteen supercharges). While the solution is constructed usinggauged supergravity in four dimensions- the complete eleven-dimensionalsolution is given. In particular- we obtain the Killing spinors explicitly- andwe find that they are characterised by a duality rotation of the standardDirichlet projection matrix for M2-branes.","title":"A Dielectric Flow Solution with Maximal Supersymmetry","year":"2003"},{"name":"N.P. Warner","id":9911240},{"id":9811228,"abstract":" We show how certain F^4 couplings in eight dimensions can be computed usingthe mirror map and K3 data. They perfectly match with the correspondingheterotic one-loop couplings- and therefore this amounts to a successful testof the conjectured duality between the heterotic string on T^2 and F-theory on K3. The underlying quantum geometry appears to be a 5-fold- consisting of ahyperk ahler 4-fold fibered over a IP^1 base. The natural candidate for thisfiber is the symmetric product Sym^2(K3). We are lead to this structure byanalyzing the implications of higher powers of E_2 in the relevant Borcherdscounting functions- and in particular the appropriate generalizations of thePicard-Fuchs equations for the K3.","title":"Quartic Gauge Couplings from K3 Geometry","year":"1998"},{"name":"W. Lerche","id":9910207},{"id":9412198,"abstract":" We show that the BRST structure of the topological string is encoded in the``small $N=4$ superconformal algebra- enabling us to obtain- in a non-trivialway- the string theory from hamiltonian reduction of $A(1|1)$. This leads tothe important conclusion that not only ordinary string theories- buttopological strings as well- can be obtained- or even defined- by hamiltonianreduction from WZW models. Using two different gradations- we find either thestandard $N=2$ minimal models coupled to topological gravity- or an embeddingof the bosonic string into the topological string. We also comment briefly onthe generalization to super Lie algebras $A(n|n)$.","title":"Topological Strings from WZW Models","year":"1994"},{"name":"K. Landsteiner","id":9911124},{"id":9909166,"abstract":" We establish the existence of the topological vector supersymmetry in the sixdimensional topological field theory for two-form fields introduced by Baulieuand West. We investigate the relation of these symmetries to the twistoperation for the (2-0) supersymmetry and comment on their resemblance to theanalogous symmetries in topological Yang-Mills theory.","title":"Remarks on Topological SUSY in six dimensional TQFTs","year":"1999"},{"name":"M. Schweda","id":9911127}]	8

We observe that the shortest path between authors ‘C.N. Pope’ and ‘M. Schweda’ has length equal to 8. Although the authors are the same with the previous query, in this one we take into

consideration only the relationship WRITTEN_BY, because we are interested in only edges and nodes between authors and articles. Comparing the results with the previous, we observe that no journal node is included in the shortest path and in the distance between ‘C.N. Pope’ and ‘M. Schweda’ there are three more authors.

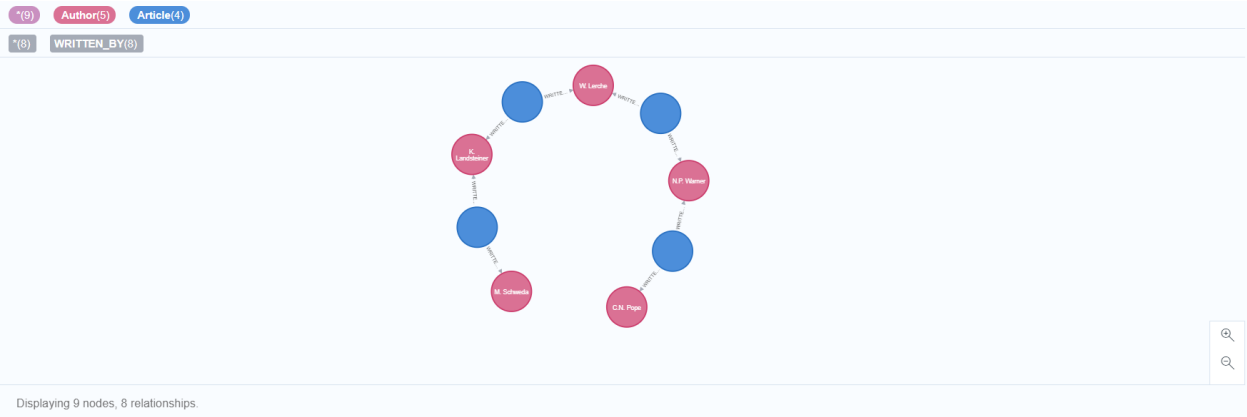


Figure 7 Figure 6 Shortest path between authors ‘C.N. Pope’ and ‘M. Schweda’ without Journal nodes

10. 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

```
MATCH p = ShortestPath((au:Author{name:'Edward Witten'})-[w:WRITTEN_BY*]-(au2:Author))
WHERE au<>au2
AND length(p) > 25
AND NONE(n in nodes(p) WHERE n:Journal)
RETURN au2.name as Author_Name, length(p) AS Path_Length, [n in nodes(p) WHERE n.title IS NOT NULL| n.title] AS Paper_Title
```

Results

After several attempts we conclude that the above query is the correct one, as it calculates all the authors that have shortest path greater than length 25 with the author ‘Edward Witten’. From this calculation, the relationship among journal are omitted. Unfortunately, although the query is executable, it does not completed.


MATCH p = ShortestPath((au:Author{name:'Edward Witten'})-[w:WRITTEN_BY*]-(au2:Author))
WHERE au<>au2
AND length(p) > 25
AND NONE(n in nodes(p) WHERE n:Journal)
RETURN au2.name as Author_Name, length(p) AS Path_Length, [n in nodes(p) | n.title] AS Paper_Title
Table
Code