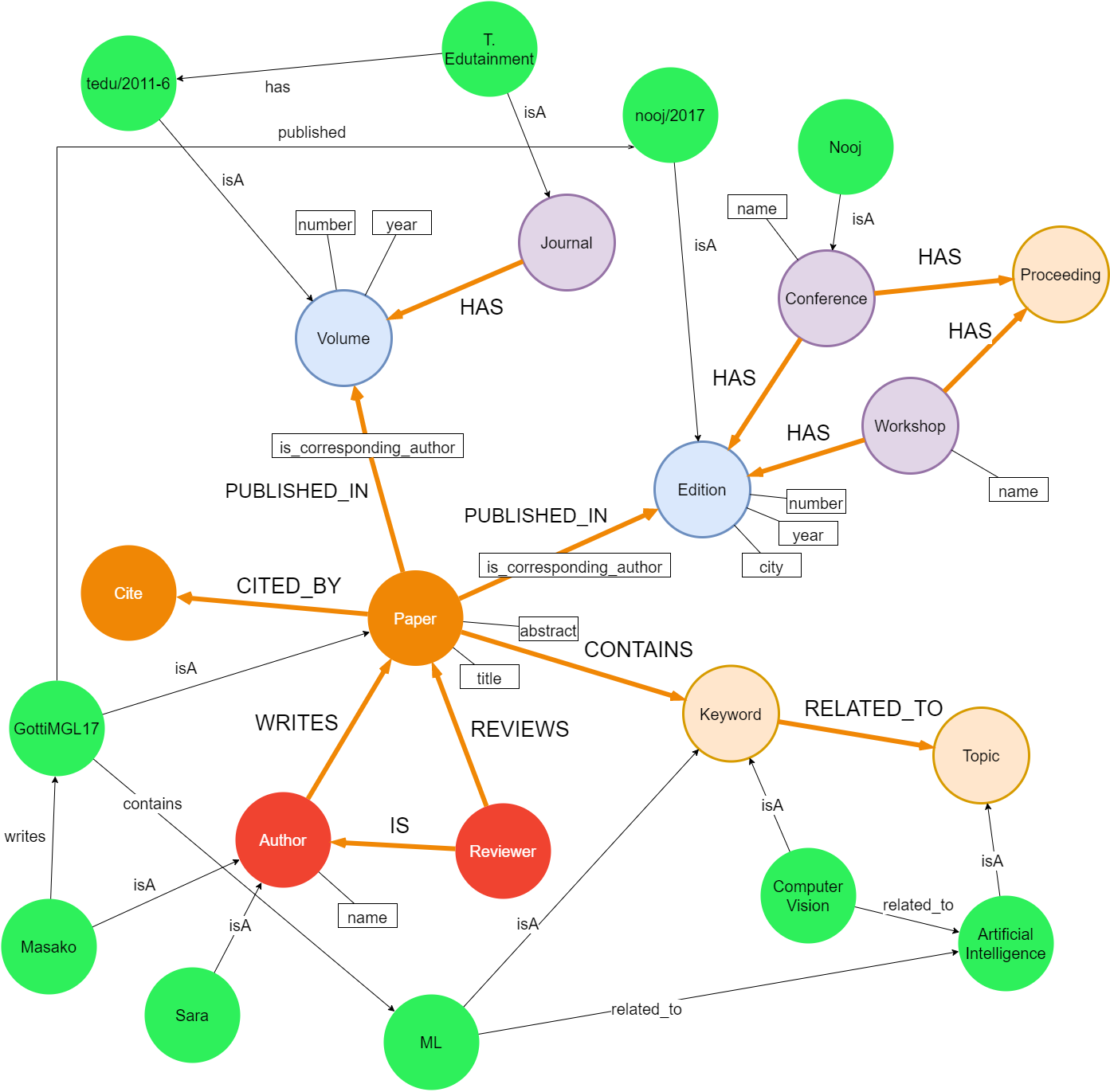
|  |  |
| --- | --- |
|  |  |

# Modeling, Loading, Evolving

## A1. Modeling



In the above graph, the green nodes are representing the data and all the other node are representing the meta-data. We have included all the meta-data and few data in our model to be easier to read. For the same reason, we do not have include all the attributes that we used while loading the real data. The attributes can be found in the loading scripts in Part B.

Regarding our design decisions, we tried to design the model in such a way that the queries in the Part B are well optimised. More specifically, we made the following decision:

* The Edition and the Volume are different nodes from the Conference and the Journal, respectively. We made this choice as most of the queries are related to Conferences (e.g. Find the top 3 most cited papers of each conference.). This means that we do not want to have the Edition as an attribute of the node Conference, as this would affect the efficiency (we would need extra I/Os).
* Similarly, Workshop and Conference are different nodes. If we had both in 1 node, we would need an extra attribute isConference (Boolean) and in this case, we would need to look up in this attribute to specify whether the node is Conference or Workshop.
* The City of the Edition of the Conference is an attribute of the Edition node, as we assume that no further analysis will be done based on the City. If we had queries that engaged the City, it would be better to have it as a node.
* A Paper may have many Keywords and a Keyword may be related to many Topics and vice versa. That is why Keyword and Topic are 2 separate nodes.
* Regarding the Citations, initially we had a self referencing edge from Paper to Paper. However, while loading the data, we realized that the dataset did not provide the information of which Paper cited each paper. The only information available was that a Paper is cited. That is why we have a separated node for Cite, which is linked to the Paper.

## A.2. Instantiating/Loading

Since there was a real dataset available (i.e. DBLP), we decided to use it for our data loading task, in order to keep our data as realistic as possible. As DBLP publishes its data in the form of an XML file with a corresponding DTD, we downloaded the latest one and extracted, transformed and modified the data it contains into CSV files spanning most of our data needs for the proposed model. We used the following python script in GitHub for this purpose:  
<https://github.com/ThomHurks/dblp-to-csv/blob/master/XMLToCSV.py>

The output of running the previous script was a Folder containing 28 csv files some of them representing content data, while others were used solely for storing headers for other CSV files.

Subsequently, and for each needed file, we wrote a python script that extracts the attributes we are interested in our model and generates another intermediate CSV file ready to be bulk loaded in neo4j. The bulk loading scripts were written in conformance to the schema of the intermediate CSV files generated by our Python scripts. This approach helped us keeping the Cypher loading scripts as clean as possible. Following are the remarks on CSV files from DBLP dataset that were used for data extraction:

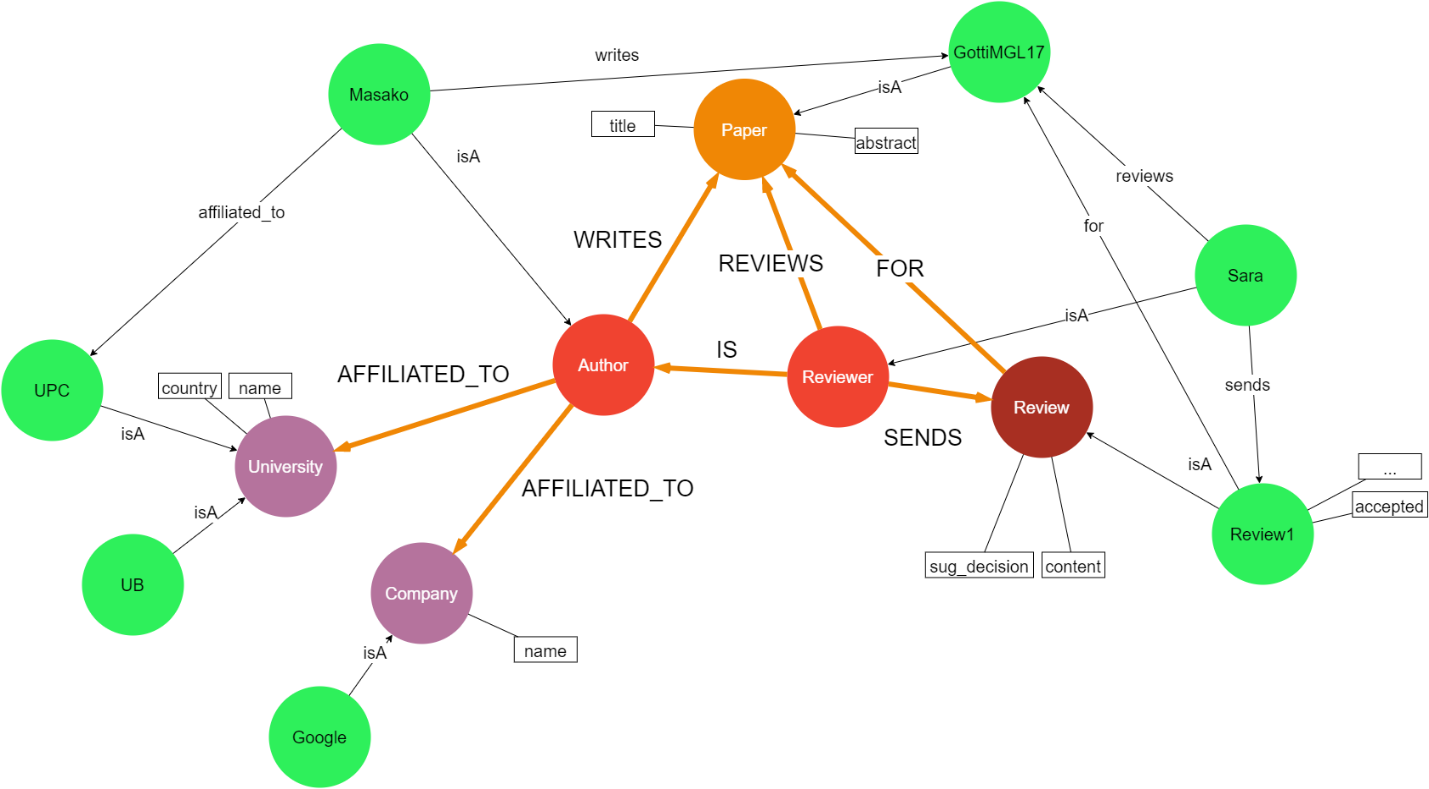
* **output\_proceedings.csv:** From this original file we extracted *Conference.csv (2500 records)*, *Edition.csv (10000 records), Workshop* and *Proceedings.csv* files. The relationship between Edition and Conference was created on the fly following each Edition node’s creation.
* **output\_inproceedings.csv:** From this original file along with its header file, we extracted *Paper(Edition).csv (10000 records)* representing conference papers and their relationships to editions and *Edition\_Paper\_Author.csv*. We also extracted part of *Authors.csv (25000 records)* from this file, spanning the authors who wrote conference papers.
* **output\_article.csv:** From this original file along with its header file, we extracted *Journal.csv (20 records), Volume(Journal).csv (1800 records), Paper(Volume).csv (1800 records)* and part of *Authors.csv.* Taking all this information from this file, helped us store only the journals and its volumes that have published papers which was more meaningful to our model for the next tasks.

As for the missing data from DBLP, we generated the following files:

* **Cite.csv:**  this file contains 45000 artificial citations generated randomly with a Python script. The citation contains only its own identifier, and a key to the paper it cites. As explained in part A, we care only about how many times a certain paper was cites, with no need to keep the source paper.
* **Keyword.csv:** this file contains keywords for 10000 papers, storing two most important keywords for each paper. The assignment of keywords to papers was done randomly in a Python script.
* **Paper\_reviewer.csv:** This file contains the assignment of 1000 reviewers (selected randomly from Authors.csv) to 3000 papers. The assignment rule is that no paper can be assigned more than 3 reviewers, and a reviewer can not be assigned a paper he authored. This was all achieved in a Python script.

**The used dataset can be found** [**here**](https://drive.google.com/drive/folders/1fStYnk5BXQJvGaqZBw-y84i6GgCIkIAw?usp=sharing)**.**

## A.3 Evolving the graph



Regarding our modeling decision for the evolved graph, we would like to make the following notes:

* We decided to separate the Author and Reviewer nodes, are they have different semantics. However, we could also have one node for both, that would have different edges to Paper (WRITES and REVIEWS).
* The Review is connected both to Reviewer and Paper. Another option would be to connect the Review only to its Reviewer, but in this way, we would not be able to identify which Review is for which Paper, as a Reviewer may review more than one Paper.

# Querying

**All the following queries (Part B, C and D) can be also found inside our Application.py file.**

**Query1:** Find the h-indexes of the authors in your graph

MATCH(a:Author)-[:WRITES]->(p:Paper)-[:CITED\_BY]->(c:Citation)

WITH a as authors, p.key as papers, count(c) as number\_of\_citations

ORDER BY number\_of\_citations DESC

WITH authors as authors, collect(number\_of\_citations) as citations\_list

WITH authors as authors, citations\_list AS citations\_list

UNWIND range(0,size(citations\_list)-1) as l\_index

WITH authors as authors,

CASE

WHEN citations\_list[l\_index] >= l\_index+1 THEN l\_index+1

ELSE -1

END AS hindex

WHERE hindex <> -1

RETURN authors.name, max(hindex)

**Query2:** Find the top 3 most cited papers of each conference.

MATCH(c:Conference)-[:HAS]-(e:Edition)<-[:PUBLISHED\_IN]-(p:Paper)-[:CITED\_BY]-(t:Citation)

WITH c as c, p as p, count(t) as cites

ORDER BY c.name ASC, cites DESC

RETURN c.name as Conference, collect(p.title)[..3] as Most3CitedPapers

**Query3**: For each conference find its community: i.e., those authors that have published papers on that conference in, at least, 4 different editions.

MATCH(c:Conference)-[:HAS]->(e:Edition)<-[:PUBLISHED\_IN]-(p:Paper)<-[w:WRITES]-(a:Author)

WITH c as conference, a as author, count(distinct e) as number\_of\_editions

WHERE number\_of\_editions >= 4

RETURN conference.name as Conference, collect(author.name) as Community

**Query 4:** Find the impact factors of the journals in your graph

Match(j:Journal)

OPTIONAL MATCH (j)-[:HAS]->(v:Volume)

OPTIONAL MATCH (v)<-[:PUBLISHED\_IN]-(p:Paper)

OPTIONAL MATCH (p)-[:CITED\_BY]->(c:Citation)

WHERE p.year >= '2018' and p.year <= '2019'

RETURN j.name as Journal, count(DISTINCT p) as Papers18\_19, count(c) as Citations2019, count(c)\*1.0/count(DISTINCT p) as IMPACT\_FACTOR

ORDER BY IMPACT\_FACTOR DESC

# Graph algorithms

**Page Rank**

CALL algo.pageRank.stream(

'MATCH (n:Paper) RETURN id(n) AS id UNION MATCH (n:Citation) RETURN id(n) AS id',

'MATCH (n:Paper)-[:CITED\_BY]->(m:Citation) RETURN id(m) AS source, id(n) AS target',

{graph: 'cypher'})

YIELD nodeId, score RETURN algo.asNode(nodeId).key AS page,score ORDER BY score DESC

The Page Rank algorithms measures the transitive influence or connectivity of nodes. This means that it measures the importance of the nodes, by counting the links to them. In our case, we use this algorithm to count the importance of the Papers, which is determined by the Citations that it has. The more the Cites a Paper has, the biggest its Page Rank will be.

As we explained earlier in the modeling, we modeled the Citation as a separate node and not as a self reference of the Paper node. This affects the result of this algorithm, as Page Rank considers the Page Rank of the Papers that cite other Papers. This means that it works recursively which in our case is impossible as the Citations are not Paper and hence, they are not Cited by other nodes.

**Betweenness Centrality**

CALL algo.betweenness.stream(

'MATCH (p) WHERE ANY(lbl IN ["Paper", "Keyword", "Topic"] WHERE lbl IN LABELS(p)) RETURN id(p) as id',

'MATCH (p1)-[r]->(p2) WHERE TYPE(r) IN ["RELATED\_TO", "CONTAINS"] RETURN id(p1) as source,id(p2) as target', {concurrency:4, graph:'cypher'})

YIELD nodeId, centrality

WITH nodeId AS nodeId, centrality AS centrality

MATCH (k:Keyword) WHERE id(k) = nodeId

RETURN algo.asNode(nodeId).name, centrality ORDER BY centrality desc

We use the Betweenness centrality algorithm to identify Keyword nodes that behave as a bridge from one part of the graph to another. More specifically, we are trying to find Topics that are related to Keywords, in such a way that if we remove a Keyword from the graph, then we will completely lose the connection between Papers and some Topics, that were connected before with only this Keyword and hence, this Keyword was the only bridge between these Papers and these Topics. This information can be useful in our application as we can find Topics that are linked to Papers only via a specific path (Keyword). After that, we can try to find in their content other Keywords that maybe are related to the same Topic in order to link with these as well. For instance, in our database, we created the Keyword {Machine learning} and 3 Topics that are related only to this Keyword. The betweenness centrality of this Keyword is 3, because if we remove it, we will end up with 3 Topic that are not connected to the rest of the graph.

# Recommender

// **Find the database community**

MATCH(k:Keyword)

WHERE k.name in ['data management', 'indexing', 'data modeling', 'big data', 'data processing', 'data storage', 'data querying']

// **Find all the conferences that are related to the database community**

WITH collect(k.name) as db\_community

MATCH(k:Keyword)<-[:CONTAINS]-(p:Paper)-[:PUBLISHED\_IN]->(e)<-[:HAS]-(c)

WHERE k.name in db\_community AND ANY(lbl IN ["Edition", "Volume"] WHERE lbl IN LABELS(e)) AND ANY(lbl IN ["Conference", "Journal"] WHERE lbl IN LABELS(c))

WITH c.name as conference, count(distinct p) as number\_of\_db\_community\_papers

MATCH(k:Keyword)<-[:CONTAINS]-(p:Paper)-[:PUBLISHED\_IN]->(e)<-[:HAS]-(c)

WHERE ANY(lbl IN ["Edition", "Volume"] WHERE lbl IN LABELS(e)) AND ANY(lbl IN ["Conference", "Journal"] WHERE lbl IN LABELS(c))

WITH conference as confjour, LABELS(c)[0] as type, number\_of\_db\_community\_papers as number\_of\_db\_community\_papers , c.name as confjour2, count(distinct p) as number\_of\_papers

WHERE confjour2 = confjour and number\_of\_db\_community\_papers \* 1.0 / number\_of\_papers >= 0.9

WITH confjour as confjour, type as type, number\_of\_db\_community\_papers as number\_of\_db\_community\_papers, number\_of\_papers as number\_of\_papers, number\_of\_db\_community\_papers \* 1.0 / number\_of\_papers as percentage

CREATE (cj:dbCommCJ{ name:confjour, type:type, n\_dbPapers: number\_of\_db\_community\_papers,n\_papers:number\_of\_papers, percentage: percentage })

WITH cj as cj

MATCH (c:Conference {name: cj.name})-[:HAS]->(e:Edition)<-[:PUBLISHED\_IN]-(p:Paper)

CREATE UNIQUE (cj)-[:HAS\_PAPER]->(p)

CREATE UNIQUE (c)-[:IS\_A]->(cj)

RETURN DISTINCT cj.name AS conf\_journal, cj.type AS type, cj.n\_dbPapers As number\_of\_db\_community\_papers, cj.n\_papers AS number\_of\_papers, cj.percentage AS percentage

ORDER BY percentage DESC

In the first query, we identify the Conferences/Journals that belong to a database community and for each one we create dbCommCJ node, in order to include the inferred knowledge to the graph. In addition, we create edges from the dbCommCJ node to the corresponding Papers (:HAS\_PAPER) as well as to the corresponding Conferences/Journals (IS\_A).

// **Find top cited Papers (highest page rank)**

MATCH (cj:dbCommCJ)

WITH collect(cj.name) as dbCommunity

MATCH (cj:dbCommCJ)-[:HAS\_PAPER]->(p:Paper)

OPTIONAL MATCH (p)-[:CITED\_BY]->(c:Citation)

WHERE c.confjour in dbCommunity

WITH p as p,cj.name as conference\_journal, cj.type as type, count(c) as citations

ORDER BY citations DESC Limit 10

CREATE (tp: TopPaper{n\_citations: citations})

CREATE (p)-[:IS\_A]->(tp)

RETURN p.title as paper, conference\_journal, type, citations

Here, we need to mention that we added an attribute to the Citation nodes that represents the Paper by which the Citation is done. Moreover, we create a node for every top Paper and a connection to the corresponding Paper.

// **Create good and guru authors**

CREATE (x:DbReviewer{})

CREATE (y:Guru{})

WITH x as x, y as y

MATCH (a:Author)-[:WRITES]->(p:Paper)-[:IS\_A]->(tp:TopPaper)

WITH a, count(p) as cnt

WITH a.name as Author, CASE WHEN cnt>1 THEN 'YES' ELSE 'NO' END as IS\_GURU

ORDER BY IS\_GURU DESC

CREATE (a)-[:IS\_A]->(x)

WITH IS\_GURU as IS\_GURU, Author as Author

WHERE IS\_GURU = 'YES'

CREATE (a)-[:IS\_A]->( y)

// **Show good and guru authors**

MATCH (a:Author)-[:WRITES]->(p:Paper)-[:IS\_A]->(tp:TopPaper)

WITH a, count(p) as cnt

WITH a.name as TopAuthors, CASE WHEN cnt>1 THEN 'YES' ELSE 'NO' END as IS\_GURU

ORDER BY IS\_GURU DESC

In the last query, we create 2 new node, namely DbReviewer and Guru. Then, we identify all the Authors that can be reviewers of the database community and we link them with the DbReviewer node. We also find the guru Authors and we link them with the Guru node. Finally, we are printing all the top Authors.