Sparql2GraphServer: a Server-side Tool for Extracting Networks from Linked Data for Data Analysis

Petri Leskinen[1][0000−0003−2327−6942] and Eero Hyvönen[1,2][0000−0003−1695−5840] and Jouni Tuominen[1,2][0000−0003−4789−5676]

[1] Semantic Computing Research Group (SeCo), Aalto University, Finland
[2] HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland
http://seco.cs.aalto.fi, http://heldig.fi, firstname.lastname@aalto.fi

Abstract. This paper presents a server-side tool for constructing graph representations from Linked Data for network analysis. The main features for the tool are: 1) Support to read data from a SPARQL endpoint, 2) easy adaptation to virtually any dataset and data model, and 3) easy usage in web portals for data analysis. The application is in use.

Keywords: Linked Open Data · Network Visualization · Data Analysis
Poster paper

1 Introduction

Visualizing a network provides new insights about its structure and the underlying data. A network can be built from Linked Data (LD) publication using SPARQL [3], where extraction patterns [1] allow a researcher to extract various types of connections (*links, edges*) between the *nodes*.

Many portals in the domain of Digital Humanities use network visualizations for data analysis, such as Six Degrees of Francis Bacon[3], Linked Jazz[4], or the co-citation graph and correspondent network in ePistolarium[5]. LodLive[6] allows the user to browse individual triples of a LD database. For front-end visualizations there are libraries such as D3.js[7], Sigma[8], Cytoscape.js[9], and 3D Force-Directed Graph[10] for rendering a network in 3D using WebGL.

This paper introduces an online service, Sparql2GraphServer, for constructing networks based on LD knowledge graphs. The tool facilitates constructing either 1) an egocentric network [7] around *ego*, a specific database ID, or 2) a general

[3] http://sixdegreesoffrancisbacon.com
[4] https://linkedjazz.org/network/
[5] http://ckcc.huygens.knaw.nl/epistolarium
[6] http://en.lodlive.it/
[7] https://d3js.org/
[8] http://sigmajs.org/
[9] https://js.cytoscape.org/
[10] https://github.com/vasturiano/3d-force-graph

network, e.g., a group of nodes with specified properties. The service is used in practise in the biographical semantic portals AcademySampo[11] (in use) [5,6] and LetterSampo[12] (prototype).

2 Using the Tool

Querying an Endpoint The process of using the tool consists of three sequential steps. Firstly, a query for links is performed as a sequence of breadth-first searches until a desired amount of links is achieved. Secondly, the graph is constrained so that if the number of links exceeds the limit, the nodes having the lowest degrees are removed. This aims to retrieving more dense connections e.g. a higher triadic closure instead of start-shaped structures around a few central nodes. Finally, the node metadata is queried calculating simultaneously the predefined network statistics, e.g., the network diameter, average degree, and the total number of edges, nodes and connected components. Example queries using Wikidata are shown in Tables 1 and 2. The link query requests for teacher-student relationships starting from the sources nodes. In the query the placeholder `<ID>` is replaced with a list of source set nodes, initially containing only the *ego* given with query parameter `id`. The link weight is defined as the count of common occupations, fields of work, and Academy memberships between two nodes.

Server Response The JSON-formatted server response is depicted in Table 3. The result field `elements` contains two arrays for `edges` and `nodes`. The API adds the data fields defined in the result clauses in the corresponding SPARQL queries to the result. The node data contains metrics like `degree`, `in_degree`, `pagerank`, and `distance` e.g. the shortest path length from *ego*. Finally, the element `metrics` contains, e.g., network diameter, average degree, and the numbers of nodes, edges, and connected components for the entire network. The response format described is compatible with Cytoscape.js[13], a library for network visualization in a web portal. Furthermore, the application also supports the GraphML[14] format.

Visualization in a Front-End Portal The API response can directly be an input for a cytoscape.js component in a portal. Fig. 1 depicts a network visualization based on the example in tables 1–3. The node and edge labels are extracted from the query results while the visual appearance is configured in the front-side application code. Here the node color on a red-blue palette is based on the path distance from the center node and the sizes are based on the out-degree.

The tool Sparql2GraphServer is based on Flask[15], microframework application written in Python 3.8 using modules NetworkX 2.4[16][2], NumPy[17], RDFLib[18],

[11] https://seco.cs.aalto.fi/projects/yo-matrikkelit/en/
[12] https://seco.cs.aalto.fi/projects/rrl/
[13] https://js.cytoscape.org/
[14] http://graphml.graphdrawing.org/
[15] https://flask.palletsprojects.com/en/2.0.x/
[16] https://networkx.org/
[17] https://numpy.org/
[18] https://rdflib.readthedocs.io

Table 1. Query for links

```
PREFIX rdfs:
  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX wdt:
  <http://www.wikidata.org/prop/direct/>

SELECT DISTINCT ?source ?target ?label
    (COUNT(DISTINCT ?link)+1 AS ?weight)
WHERE {
    VALUES ?source { <ID> }

    # teacher-student relations
    VALUES (?rel ?label) {
        (wdt:P802 "Student")
        (wdt:P185 "Doctoral student")
    }
    ?source ?rel ?target

    # links by common occupations, fields
        of work, or memberships
    OPTIONAL {
        VALUES ?prop
          { wdt:P101 wdt:P463 wdt:P106 }
        ?source ?prop ?link . ?target ?
        prop ?link
    }
} GROUP BY ?source ?target ?label
```

Table 2. Query for nodes

```
PREFIX rdfs:
  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX wdt:
  <http://www.wikidata.org/prop/direct/>

SELECT DISTINCT ?id ?name
WHERE {

    VALUES ?id { <ID_SET> }
    ?id rdfs:label ?name .
    FILTER (LANG(?name) = "en").

} GROUP BY ?id ?name
```

Table 3. Server Response

```
{'elements': {
  'edges': [
    {'data':
      {'source':
        'www.wikidata.org/entity/Q9047',
      'target':
        'www.wikidata.org/entity/Q76510',
      'label':  'doctoral student',
      'weight':   7}
    }, {'data': ... }, ... ],
  'nodes': [
    {'data': {'id':
      'www.wikidata.org/entity/Q9047',
      'name':'Gottfried Wilhelm Leibniz',
      'distance': 0, 'degree':  5, '
       degree_weighted':  22,
      'in_degree': 0, 'out_degree':5,
      'in_degree_weighted': 0,
      'out_degree_weighted':22,
      'pagerank':  0.01214}
    }, {'data': ... }, ... ] },
'metrics':
  {'average_degree': 2.48,
  'diameter': 10,
  'number_connected_components': 1,
  'number_of_edges':  62,
  'number_of_nodes':  50},
'directed':   True
}
```

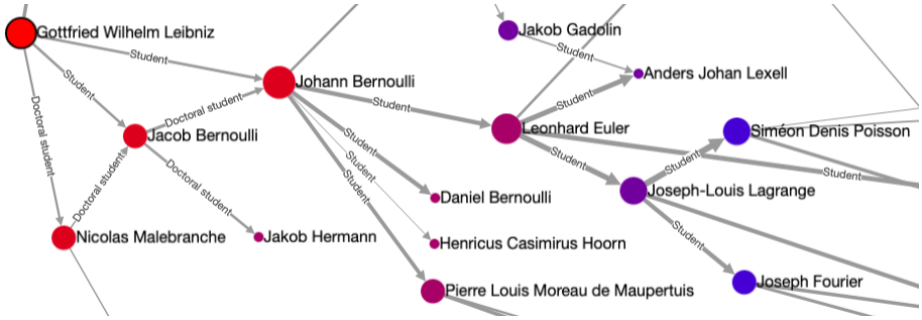

Fig. 1. Social network of mathematician Gottfried Wilhelm Leibniz in Wikidata

and SparqlWrapper[19]. The source code is available in GitHub[20]. The source code repository contains a Dockerfile recipe for running and deploying the application as a container which allows for portability and scaling.

3 Discussion

Integration to Sampo-UI Framework The Sparql2Network API is integrated to the Sampo-UI Framework [4]; the API was developed and is used in the AcademySampo and LetterSampo portals with Sampo-UI. The Sampo-UI framework includes functions for scaling and constraining the numeric result values into, e.g.,

[19] https://rdflib.dev/sparqlwrapper/
[20] https://github.com/SemanticComputing/Sparql2GraphServer

edge widths, node sizes, as well as into RGB or HSL color ranges. For example, this page[21] in AcademySampo shows a connection network around the Finnish composer Jean Sibelius, the page[22] demonstrates using Sparql2Network to create a sociocentric network, possibly filtered by using faceted search.

Evaluation Performing several SPARQL queries to the database might have longer request times, especially in cases of smaller extracted networks. An alternative would be using a single query empowered with, e.g., property paths or nested selection blocks. However, that approach is not guaranteed to perform in feasible time in a more complex case and would require customizing the queries for each specific database. The chosen solution to have a pair of simple and quick queries appears to be more efficient.

Acknowledgements This work is related to the EU project InTaVia: In-Tangible European Heritage[23], and the EU COST action Nexus Linguarum[24] on linguistic data science. CSC – IT Center for Science provided computational resources for the work. Discussions with Mikko Kivelä, Javier Ureña-Carrion, Minna Tamper, and Esko Ikkala are acknowledged.

References

1. Ghawi, R., Pfeffer, J.: Extraction Patterns to Derive Social Networks from Linked Open Data Using SPARQL. Information 11(7), 361 (2020).
2. Hagberg, A.A., Schult, D.A., Swart, P.J.: Exploring network structure, dynamics, and function using networkx. In: Varoquaux, G., Vaught, T., Millman, J. (eds.) Proceedings of the 7th Python in Science Conference. pp. 11 – 15. Pasadena, CA USA (2008).
3. Harris, S., Seaborne, A.: SPARQL 1.1 Query Language. https://www.w3.org/TR/sparql11-query/, accessed: 2021-06-28.
4. Ikkala, E., Hyvönen, E., Rantala, H., Koho, M.: Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces. Semantic Web (2021), http://www.semantic-web-journal.net/, accepted.
5. Leskinen, P., Hyvönen, E.: Linked open data service about historical Finnish academic people in 1640–1899. In: DHN 2020 Digital Humanities in the Nordic Countries. Proceedings of the Digital Humanities in the Nordic Countries 5th Conference. pp. 284–292. CEUR Workshop Proceedings, Vol. 2612 (2020), http://ceur-ws.org/Vol-2612/short14.pdf.
6. Leskinen, P., Hyvönen, E.: Reconciling and using historical person registers as linked open data in the AcademySampo knowledge graph. In: Proceedings of the 20th International Semantic Web Conference (ISWC 2021). Springer (October 2021), https://seco.cs.aalto.fi/publications/2021/leskinen-hyvonen-reconciling-2021.pdf, forthcoming.
7. Marsden, P.V.: Egocentric and sociocentric measures of network centrality. Social networks 24(4), 407–422 (2002).

[21] https://akatemiasampo.fi/en/people/page/p21762/connections
[22] https://akatemiasampo.fi/en/people/faceted-search/network
[23] https://intavia.eu/
[24] https://nexuslinguarum.eu/the-action