Decomplexifying networks: a tool for RDF/Wikidata to network analysis

Authors: Julie M. Birkholz, Postdoctoral Researcher, WeChangEd, Department of Literary Studies, Ghent University, Ghent, Belgium, @juliebirkholz; & Albert Meroño Peñuela, Postdoctoral Researcher, Faculty of Science, Artificial intelligence
Network Institute, VU Amsterdam, Amsterdam, the Netherlands, @albertmeronyo

Linked Data is a common way to publish structured data in the Humanities (De Boer et al., 2014; Meroño-Peñuela et al., 2017). The non-discriminate model allows users to model knowledge as multimodal graphs- meaning theoretically any related objects can be represented as a graph. Inherently, this affords modeling data as networks and thus the implementation of network analysis. The analysis of networks from RDF is largely done with a pipeline of tools (i.e. Groth & Gil 2011; Gil & Groth 2011). In these workflow approaches, researchers specify SPARQL queries, extract networks and export data as matrices, and implement network analysis tools to investigate graphs. The development of such a pipeline can be a technical adversary, for domain experts (e.g. historians, literary scholars) with (traditionally) limited technical knowledge, but also for researchers with specific expertise in RDF or networks. In addition, in building such pipeline, we lose sight of the hermeneutics of the research objects (Gibbs & Owens 2013). Thus, we argue, there is a need within the DH community, to reduce this RDF-to-network analysis pipeline without creating another domain or research question specific tool and while maintaining oversight over the process from RDF-to graph-to network analysis.

We developed a Jupyter notebook that integrates the Python packages: rdflib with networkx; resulting in a reusable workflow that allows network analyses over RDF data to be more accessible. The full notebook is available at https://github.com/descepolo/rdf-network-analysis/blob/master/rdf-network-analysis.ipynb. A Google Collaboratory version of the notebook, which makes it executable on the Web with no need of local installation, is available at: https://colab.research.google.com/github/descepolo/rdf-network-analysis/blob/master/rdf-network-analysis.ipynb.

The notebook consists of five “cells”, which are actionable code blocks, shown here in Figure 1. The output of all these processes can then be selected and copy-pasted for further reuse in graph processing frameworks or directly in reports or papers. We show this proof concept in a demonstration through the use of examples on Wikidata.

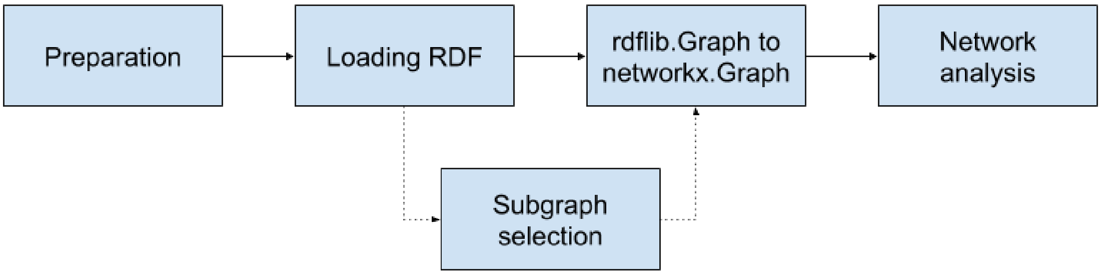

Figure 1. Workflow of the RDF Network Analysis Jupyter notebook.

1. Preparation. The notebook loads the relevant packages - rdfilb and networkx;
2. Loading RDF graphs. The user is prompted to input the full path to an RDF graph. This can be any local or online RDF file
3. Subgraph selection. Users select a specific network in the RDF graph.
4. rdflib.Graph to networkx.Graph. This prepares the graph for networkx analysis.
5. Network analysis. We have selected a standard, non-exhaustive, set of one-mode complete network measures. (See Wasserman & Faust 1994)
6. Following this selection the network analysis is run and the results are printed, as well as a basic visualization which serves for the researcher to confirm the boundaries of the network.

Reference List

De Boer, Victor, Matthias van Rossum, Jurjen Leinenga, and Rik Hoekstra. (2014). "Dutch ships and sailors linked data." In International Semantic Web Conference, pp. 229-244. Springer, Cham.

Meroño-Peñuela, Albert, Ashkan Ashkpour, Christophe Guéret, and Stefan Schlobach. (2017). "CEDAR: the Dutch historical censuses as linked open data." Semantic Web 8, no. 2: 297-310.

Gibbs F., Owens T. (2013). The hermeneutics of data and historical writing. In Nawrotzki K., Dougherty J (eds), Writing History in the Digital Age . Ann Arbor, MI: University of Michigan Press. DOI: http://dx.doi.org/10.3998/dh.12230987.0001.001.

Gil, Y., & Groth, P. (2011). LinkedDataLens: linked data as a network of networks. In Proceedings of the sixth international conference on Knowledge capture (pp. 191-192). ACM.

Groth, P., & Gil, Y. (2011). Linked data for network science. In Proceedings of the First International Conference on Linked Science-Volume 783 (pp. 1-12). CEUR-WS. Org.

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications (Vol. 8). Cambridge University Press.