Linked Data for Network Science

Paul Groth[1], Yolanda Gil[2]

[1] VU University Amsterdam, De Boelelaan 1081a,
Amsterdam, 1081 HV, The Netherlands
p.t.groth@vu.nl
[2] Information Sciences Institute, University of Southern California,
4676 Admiralty Way, Marina del Rey, CA 90292, USA
gil@isi.edu

Abstract. Network science is an emerging research area focused on developing general network-based approaches for studying phenomena across a range of fields from social science to biology. Techniques from network science include network analysis, network modeling and visualization. A key difficulty facing networks science is data acquisition. Network data must often be mined and converted from non-network sources, which is often a laborious and error prone process. In this paper, we present a simplified approach for extracting networks from Linked Data. These extracted networks can then be analyzed through network analysis algorithms, and the results of these analyses can be published back as Linked Data. The aim is to provide a corpus of well-described networks for use in network science. We describe LinkedDataLens, an implementation of this framework that uses the Wings workflow system to represent multi-step network extraction and analysis processes. Additionally, we describe initial networks that have been extracted and characterized with this framework.

Keywords: knowledge capture, network analysis, linked data, network science

1 Introduction

Network science is a discipline that "aims to develop theoretical and practical approaches and techniques to increase our understanding of natural and man made networks" [7]. It is encompasses a wide variety scientific disciplines ranging from biology, to social science, to physics and math. Common across all these areas is the use of techniques such as network analysis, network modeling and network visualization. A key challenge for network science is data acquisition. While the advent of digital data has made network science possible, for many domains data is often unavailable, incomplete, or has strong biases [7]. In this work, we begin to address this problem through the use of Linked Data [5].

We view Linked Data as a "network of networks." Diverse datasets, such as Geonames and DBpedia, are interlinked into a massive network. Within this Linked Data network, one could identify smaller self-contained subsets represented in turn as networks. For example, one might extract a social network of people who are members of the current US Congress from the larger DBpedia dataset. These

extracted networks could span more than one dataset. For example, one could construct a temporal network of events containing all concerts in a geographical region, which would integrate information from event and geospatial sources. Each of these extracted networks represents a meaningful aspect of some phenomenon, and can be studied and characterized in their own right. For example, using network analysis algorithms we can derive useful summary statistics, detect clusters, and infer new links. The resulting analyses can be seen as metadata of the extracted networks. This metadata can be used to formulate queries to search for networks or entities of interest with particular characteristics. For example, finding whether social networks have parallel network properties to the content networks they are associated [23].

We have implemented this framework in a system called LinkedDataLens. Through it, we generate three kinds of useful artifacts: 1) the extracted networks themselves, 2) their derived characteristics as metadata, and 3) the analytic processes used to derive those characteristics. Since all these artifacts have value for the community, our system exports them as Linked Data. A key aspect of our approach is the use of a workflow system to manage the creation and export of these artifacts. We use computational workflows to represent multi-step network extraction and analysis processes [22]. Workflows represent data analysis routines as workflow components. Workflows also contain links that express the dataflow among these components and reflect the interdependencies that must be managed during their execution. Workflow systems record all execution results together with their provenance. Such systems are often used within the network science community [6].

The major contributions of this work are:

1. A framework for characterizing Linked Data using aggregate measures of its contents. We do this by identifying and extracting meaningful subsets of the data and using network analysis algorithms to derive summary statistics and other metadata of interest.
2. A publicly available open system, LinkedDataLens, that implements this framework. The system will use workflows composed of software components that extract networks from Linked Data, analyze the characteristics of the networks, and generate graphs and visualizations of the results. The workflows will be executed to derive the new metadata and their provenance as represented by the workflows. LinkedDataLens is available open source, so others can extend the system at all levels, from adding components to new functionality to the workflow system. It can be easily installed in a local machine. It is also available as a community web portal at http://linkeddatalens.isi.edu.
3. A new approach to create datasets of interest to network science, published as Linked Data in the form of extracted networks and metadata about their characteristics. The extracted networks and corresponding metadata are published automatically by LinkedDataLens. The system automatically publishes the derived characteristics and their provenance, so anyone using it to run analyses will be exposing useful content to others. Furthermore, this new metadata can be queried to find datasets of interest. Importantly, the resulting networks are readily available to the network science community in a format that they use, to enable cross-pollination and to facilitate sharing.

The paper begins with a description of the general framework that we adopt. We then describe LinkedDataLens as a realization of this framework. After which, specific networks that were created and analyzed with this framework are presented. This is followed by a discussion of related work. We finalize with conclusions.

2 Framework

Our framework addresses a number of challenges to extracting and analyzing networks from Linked Data. First, the networks to be analyzed may not be directly accessible within Linked Data. For example, resources may be connected by multi-hop paths rather than being directly connected in a network by a single relation. Similarly, Linked Data links may be represented by resources rather than by edges in a network. Secondly, most network algorithms do not directly ingest RDF data. Finally, comprehensive metadata and provenance about the extracted networks need to be maintained in order to facilitate search. Our framework consists of the following three steps, which we discuss in more detail below.

1. Pattern-based network extraction from Linked Data
2. Characterization of the extracted networks with statistics through network analysis algorithms
3. Publication of networks as Linked Data with associated statistics and provenance metadata

2.1 Pattern-Based Extraction

Within each dataset that makes up Linked Data, we can extract a wide variety of domain specific networks. Furthermore, we may want to extract networks from across multiple linked data sets. In both cases, the networks we may wish to acquire may span multiple resource paths.

```
PREFIX dailymed:
    <http://www4.wiwiss.fu-berlin.de/dailymed/resource/dailymed/>

SELECT DISTINCT ?n1 ?n2 ?link WHERE {
 ?n1 dailymed:producesDrug ?drug.
 ?drug dailymed:activeIngredient ?link.
 ?n2 dailymed:producesDrug ?drug2.
 ?drug2 dailymed:activeIngredient ?link.
FILTER(?n1 != ?n2)
```

Fig. 1. SPARQL query following a simple triple pattern.

For example, in Figure 1 we see a SPARQL query for the DailyMed dataset that selects the components of a network of competing pharmaceutical companies where competition is defined by selling drugs with the same active ingredient. The network that we would like to construct would have the companies as nodes, and would have links between two nodes indicate competing products. In this case, to derive the links in the network we need to span a resource (some drug) and two RDF properties

(dailymed:producesDrug and dailymed:activeIngredient) to construct the appropriate link.

Even this rather simple network requires creating a view over the original dataset. To facilitate the integration of network extraction with network analysis algorithms, we use a simple pattern-based approach. We define a simple triple pattern specifying the nodes with the network and then link between those nodes. To conform to our pattern, SPARQL queries must use the same variable names (?n1, ?n2, ?link). This simplifies the parsing and construction of networks in the desired format.

We use "select" SPARQL queries instead of "construct" queries as our aim is not to produce new RDF graphs but instead to produce networks in formats that are more amendable to processing by network algorithms. After execution of a SPARQL query, we convert the variable bindings to a weighted network where the weight of each edge in the network is given by the number of occurrences of links between two nodes in the variable bindings.

2.2 Network Characterization

Networks can be characterized using a wide variety of statistical measures. For example, the degree distribution informs us about the connectedness of the network and can be a proxy for identifying the most important nodes within a network. The calculation of betweenness centrality on nodes can help understand whether particular nodes play an important role in connecting the network. Other algorithms identify which nodes provide authoritative information in the network. Simple metrics such as whether a network is connected, its density, how many edges and nodes are also useful points of reference for understanding and comparing networks. See [21] and [4] for definitions and discussion of the aforementioned (and other) network measures. In addition to statistical measures, networks can be characterized through visualizations. Visualizations are often used to be able to identify groupings and associations that are difficult to identify algorithmically. Both visualizations and statistical measures are important tools within network science.

An often overlooked side effect of these analyses is that they can provide useful characterizations of networks to search upon. For example, a science policy analyst may be interested in finding highly dense networks of scientists to study the impact of tight collaboration on productivity. Similarly, an organizational scientist may look for networks that show two dominate organizations to study duopolies. These sorts of use cases provide the motivation for the third step of our framework.

2.3 Publishing Networks

In this step of the framework, we publish the networks along with metadata about those networks. To facilitate the usage of the networks, we publish them in a format (PAJEK) that is widely supported by network tools [3]. The metadata that is associated with the network is published in RDF. In addition to the results of network characterization, we also publish the entire provenance of the both the networks generation and characterization. This additional provenance is important because it

allows us to perform queries over the union of metadata about the network, the query that was used to extract the network as well as the characteristics of the network. In addition, by providing the provenance of the network analysis, users can have greater confidence in the measurements and visualizations generated. Finally, given the completeness of the provenance information provided, others can reuse the same workflow to extract and analyze other networks.

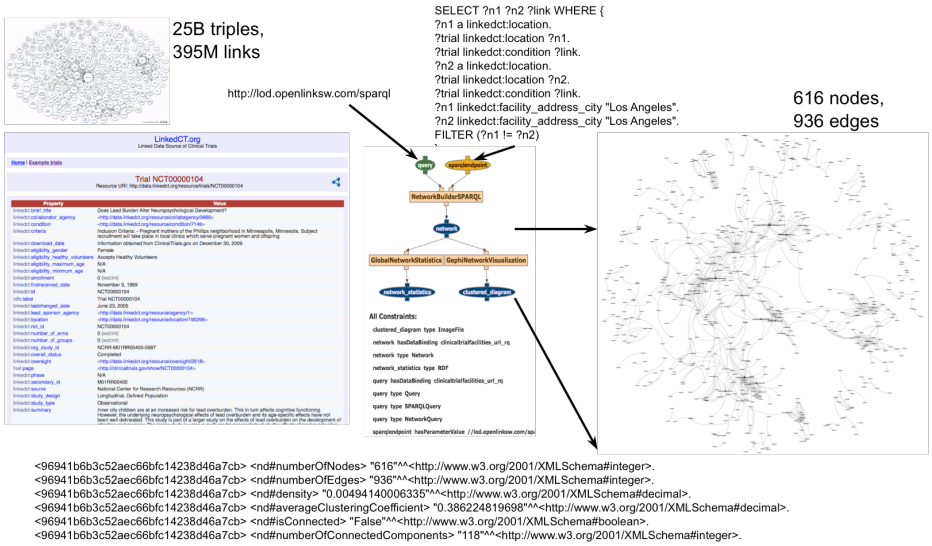
We now discuss our realization of this framework and its application to extract and characterize networks from four different data sets.

3 LinkedDataLens

LinkedDataLens is our realization of the above framework. Figure 2 gives an overview of how LinkedDataLens works. It makes use of workflows to represent network analyses. The inputs to the workflow are typically a query to Linked Data and a location to access it. When workflows are executed, networks of interest are extracted and analyzed. LinkedDataLens takes advantage of the capabilities offered by workflow systems to comprehensively capture the provenance of both the network and its characterization. The results and their provenance are published as linked data. We now discuss the details of the system.

3.1 Representing Network Extraction and Analysis as Computational Workflows

The network extraction and analysis steps are represented as workflows. Using a workflow system provides several key features: 1) it facilitates assembly of workflows from software components; 2) it automatically tracks workflow execution results and their provenance; 3) it enables reuse of workflows for new analyses. LinkedDataLens uses the Wings workflow system [12]. A unique feature of Wings is that it uses semantic representations of workflows to reason about the application requirements and assist users to create complex multi-step workflows. In particular, Wings includes algorithms for automated workflow elaboration [12], provenance and metadata generation [15], and parallel processing of data collections [10]. Wings also provides interactive assistance to create new components and workflows in the science domain [11]. Wings is released as open source software, and uses open semantic web standards such as OWL and RDF, as well as the Pegasus/Condor workflow execution software from the NSF National Middleware Initiative which allows processing datasets of very large scale [9]. The user interface is a web application, so an installation of Wings at a local institution can be accessed remotely by many users to facilitate workflow reuse and data sharing.



<96941b6b3c52aec66bfc14238d46a7cb> <nd#numberOfNodes> "616"^^<http://www.w3.org/2001/XMLSchema#integer>.
<96941b6b3c52aec66bfc14238d46a7cb> <nd#numberOfEdges> "936"^^<http://www.w3.org/2001/XMLSchema#integer>.
<96941b6b3c52aec66bfc14238d46a7cb> <nd#density> "0.00494140006335"^^<http://www.w3.org/2001/XMLSchema#decimal>.
<96941b6b3c52aec66bfc14238d46a7cb> <nd#averageClusteringCoefficient> "0.386224819698"^^<http://www.w3.org/2001/XMLSchema#decimal>.
<96941b6b3c52aec66bfc14238d46a7cb> <nd#isConnected> "False"^^<http://www.w3.org/2001/XMLSchema#boolean>.
<96941b6b3c52aec66bfc14238d46a7cb> <nd#numberOfConnectedComponents> "118"^^<http://www.w3.org/2001/XMLSchema#integer>.

Fig. 2. Overview of how LinkedDataLens works.

Our workflows typically start off with a generic component that is given a patterned query and a SPARQL Endpoint and extracts a network. For other datasets that do not offer an endpoint, as well as to use queries that span several datasets, we use larger aggregators such as the Openlink Linked Open Data LOD Cloud cache[1]. Therefore, we can use queries that aggregate data from different datasets. We are still exploring this capability.

The network extracted is then analyzed and visualized using multiple components. To facilitate interoperability between network components, we have adopted the PAJEK file format as a standard serialization to communicate networks among components. Network statistics are exposed as Linked Data using our own. In future versions of the component we plan to use the ontology defined in [8].

Analysis is performed using components based on the Gephi tookit [2] and a library NetworkX (http://networkx.lanl.gov). The system contains 8 standard network analysis algorithms and 3 visualization components. The system provides a convenient mechanism to wrap any command line tool as a component and define how those components interoperate. Importantly, the underlying implementation details are hidden from the user who can instead focus on constructing a workflow. Once created, a workflow can be applied for other networks.

3.2 Metadata and Provenance

Wings automatically records the provenance of workflow execution results [15]. The provenance includes: the workflow that was executed; links to all input, output

[1] http://lod.openlinksw.com/

and intermediate data; a specification of the software components executed; the bindings of files and parameters to the arguments of the components

This provenance is navigable in the Wings user interface. More importantly, the user can choose to expose this complete provenance as Linked Data. The execution provenance interface and its exported RDF representation can be seen in Figure 3.

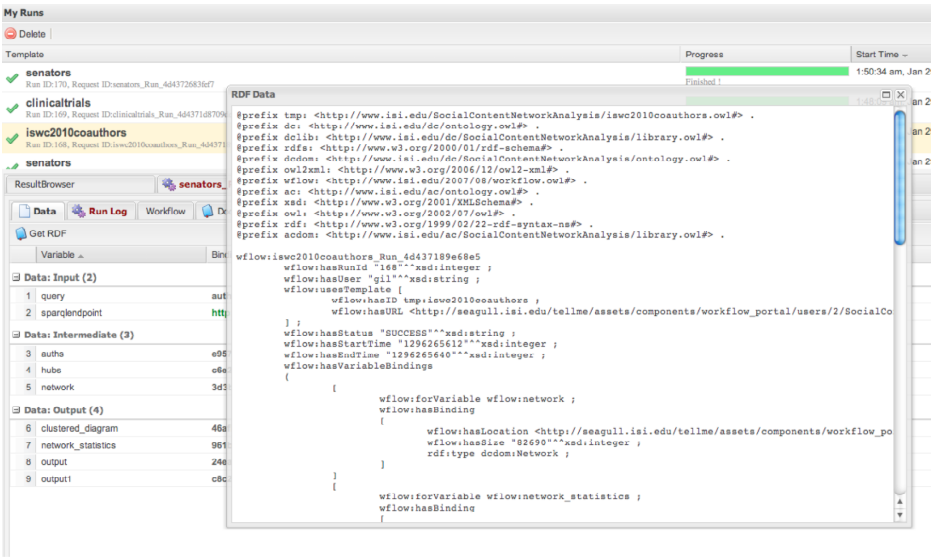


Fig. 3. The provenance records exported by LinkedDataLens.

The provenance exported has a crucial connective role between the network and its characterization. Using the provenance, we can navigate both from the characterizations of the network to the network itself as well as from the network to its characterizations. We aim to export provenance in the forthcoming W3C standard. Additionally, we can query both the results of network characterization and the provenance at the same time.

4 Analysing Subnetworks

In this section, we report on the use of LinkedDataLens to expose four different meaningful sub-networks that were extracted from Linked Data. These four networks are from the following data sets:

- DBpedia – provides access to the structure information contained within wikipedia. DBpedia acts as a focal point of the Web of Data [1].
- LinkedCT – provides a structured representation of clinical trial information interlinked with other Linked Data biomedical data sources [13].
- Drugbank – is a repository of over 5000 FDA approved small molecule and biotech drugs [24]. The Free Universteit Berlin makes available a linked data

version of this database and interlinks it with a biomedical sources such as the aforementioned LinkedCT.

- Semantic Web Dogfood – is a corpus of all information about the main conferences and workshops held in the Semantic Web community. It contains not only information about papers, but also locations, persons, and event organization [19].

We now briefly describe each of the four networks providing links to the workflow provenance, which also contains links to the actual network itself.

The networks are exported as a file in the Pajek format a common format used within the network science community. Note that they could be easily exported in RDF format back to the Web of Data. Going forward, we aim to use our approach to provide a useful corpus of networks to this community.

4.1 US Senators Alumni Network

From DBpedia, we extracted a network of incumbent United States Senators that went to the same university. It could be used to examine whether there is an impact on legislation based on university ties. We used the following SPARQL query to build the network:

```
select DISTINCT ?n1, ?n2, ?link where {
   ?n1 dbpedia-prop:wordnet_type
           wordnet:synset-incumbent-noun-1.
   ?n2 dbpedia-prop:wordnet_type
           wordnet:synset-incumbent-noun-1.
   ?n1 dcterms:subject dbpedia-cat:Living_people.
   ?n2 dcterms:subject dbpedia-cat:Living_people.
   ?n1 dbpedia-owl:almaMater ?link.
   ?n2 dbpedia-owl:almaMater ?link.
   ?n1 dcterms:subject ?state.
   ?state skos:broader
           dbpedia-cat:United_States_Senators.
   ?n2 dcterms:subject ?state2.
   ?state2 skos:broader
           dbpedia-cat:United_States_Senators.
   FILTER(?n1 != ?n2)
}
```

Some characteristics of this network include that it has 17 different connected clusters and that Jim Jeffords (an independent) is the largest hub within it. Its provenance record is available at:

http://seagull.isi.edu/tellme/assets/components/workflow_portal/users/2/SocialContentNetworkAnalysis/run_170.ttl

Figure 4 shows a visualization of this network as a whole and zooming into a portion of it, as well as an excerpt of its characteristics.

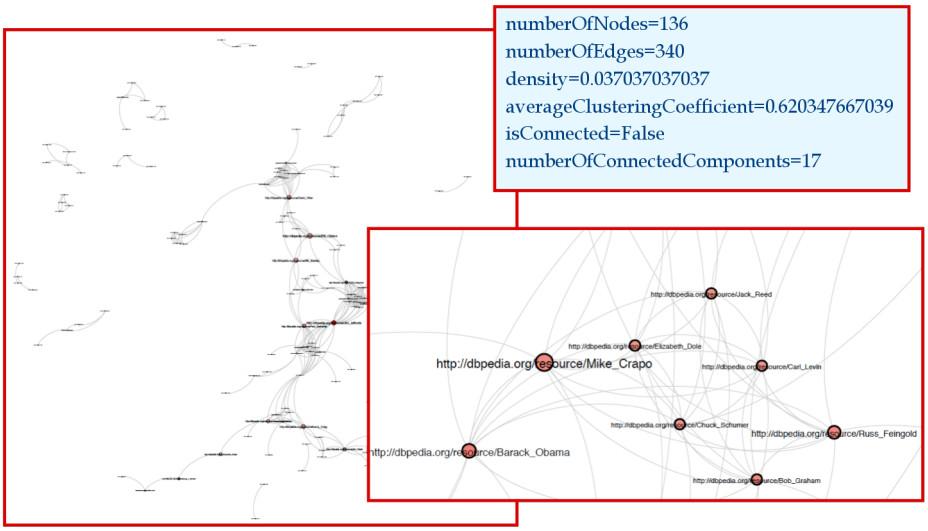


Fig. 4. US Senators that attended the same university.

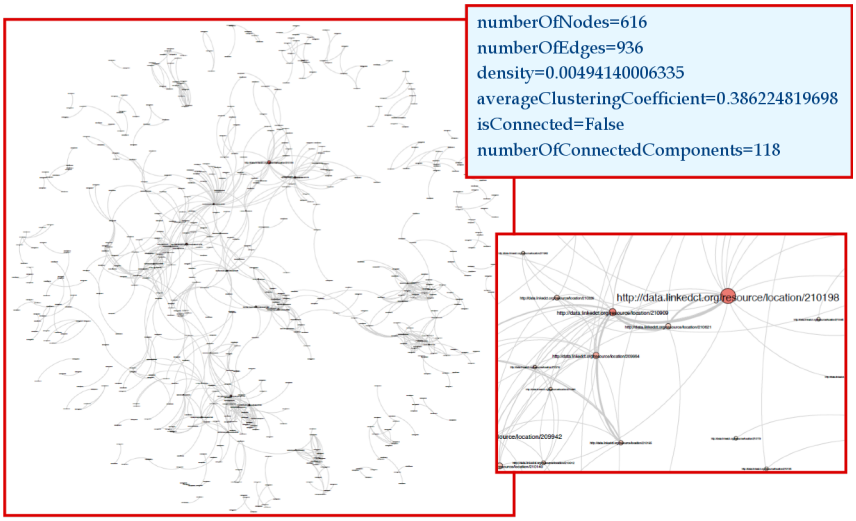


Fig. 5. Facilities in Los Angeles that investigated the same condition in a clinical trial.

4.2 Clinical Trial Facilities in Los Angeles

We extracted from LinkedCT a network of facilities within Los Angeles that have investigated the same condition in a clinical trial. The network represents 616 facilities with 936 connections between those facilities. From this network it is apparent that large universities such as the University of Southern California and UCLA are involved in many clinical trials. However, we also found that

pharmaceutical companies such as GSK are also running a variety of clinical trials in Los Angeles. The extracted network's provenance record is available at:

http://seagull.isi.edu/tellme/assets/components/workflow_portal/users/2/SocialContentNetworkAnalysis/run_178.ttl

Figure 5 shows a visualization of this network as a whole and zooming into a portion of it, as well as an excerpt of its characteristics.

4.3 Competing Pharmaceutical Companies

From the DrugBank dataset, we extracted a network of competing pharmaceutical companies where competition was defined by the selling of drugs with the same active ingredient. The same extraction and analysis workflow for the clinical trials network was used to obtain and describe this network. In this case, the network is highly connected with 17 connected components and 3032 edges for a network with 609 nodes. Instead of using the dataset directly we acquired the network using the LOD Cache endpoint. The provenance record is available at:

http://seagull.isi.edu/tellme/assets/components/workflow_portal/users/2/SocialContentNetworkAnalysis/run_176.ttl

4.4 ISWC 2010 Co-Authors

We extracted the co-author network from the Semantic Web Dogfood corpus for the International Semantic Web Conference 2010. Such co-author networks are often used in the field of scientometrics to analyze a scientific domain. For example, in [18], network analyses over networks from this same corpus were used to determine the importance of members within the Semantic Web academic community. This network has a high-clustering coefficient. This is to be expected as authors cluster together according to papers and most authors do not have more than one or two papers in a single conference. The provenance record is available at:

http://seagull.isi.edu/tellme/assets/components/workflow_portal/users/2/SocialContentNetworkAnalysis/run_177.ttl

4.5 Performance

Table 1 describes the performance of the system as it extracted each of these networks. The system was running in a Quad-Core Intel Xeon 3.6GHz with 3.4GB of RAM. The networks have very different sizes, and the datasets that they were extracted from are very different sizes. The system is able to extract the networks of interest in a very small amount of time.

Table 1. System performance versus different network sizes

Network	Exec. time	Network file size	Nodes	Edges
US Senators	12 sec.	35KB	136	340
Clinical trials	18 sec.	166KB	616	936
Pharmaceuticals	19 sec.	231KB	609	3032
Co-authors 2010	29sec.	81KB	342	579

5 Related Work

There has been some research on combining network analysis and the Semantic Web. In [17], a system, Flink, was presented that allowed network analysis over Semantic Web data. However, unlike our work it did not cater for the republishing of networks with statistics and the creation of analysis pipelines. [Martin et al 09] represent networks in RDF and show how SPARQL can be used for common network queries. Similarly, [8] uses SPARQL to and other semantic web technologies to perform network analysis. Our approach differs from both approaches in that it focuses on constructing analysis pipelines and exposing network metadata not on representing networks themselves.

A variety of social network analysis packages are available where researchers can run algorithms to analyze networks [14]. However, they do not provide a means to compose individual algorithms into a reusable workflow, nor to record provenance of the analytic results.

The SORACS project provides a service-oriented architecture to create workflows for social network analysis [20]. It illustrates the advantages of using workflows to apply heterogeneous software components. SORACS does not extract social networks, and does nor consume nor produce content as Linked Data.

There are data collections that contain social network datasets, such as the Inter-university Consortium for Political and Social Research (http://www.icpsr.umich.edu) and the DataVerse Network Project (http://thedata.org). Those datasets are contributed and described manually by the researchers that collect them. In contrast, our datasets are publicly accessible as web resources and their metadata can be queried programmatically.

5 Conclusion

We presented an approach to extract meaningful networks from Linked Data, characterize them with network analysis algorithms, and export the networks and their characterizations as Linked Data. LinkedDataLens demonstrates that Linked Data can provide a useful substrate for the network science community. In the future, we aim to expand the framework to deal with larger more heterogeneous data sets.

Acknowledgments. We would like to thank Varun Ratnakar for his feedback and assistance with this work. This research was funded in part by the National Science Foundation under grant number IIS-0948429.

References

1. Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. “DBpedia: A Nucleus for a Web of Open Data.” 6th International Semantic Web Conference (ISWC), Busan, Korea, (2007).
2. Bastian M., Heymann S., Jacomy M. “Gephi: an open source software for exploring and manipulating networks.” International AAAI Conference on Weblogs and Social Media, (2009).

3. Batagelj, V., and Mrvar, A. “Pajek. Analysis and visualization of large networks.” In M. Junger, and P. Mutzel (eds), Graph Drawing Software, Springer, (2003).
4. Brandes U. and Erlebach T., “Network Analysis: Methodological Foundations”, Lecture Notes in Computer Science, Vol. 3418, Springer-Verlag, (2005).
5. Bizer, C.; Heath, T.; and Berners-Lee, T. “Linked Data - The Story So Far.” International Journal on Semantic Web and Information Systems, 5(3), (2009).
6. Börner K... Plug-and-play macroscopes. *Commun. ACM* 54, 3 60-69, (2011).
7. Börner, K., Sanyal, S. and Vespignani, A. Network science. Annual Review of Information Science and Technology, 41: 537–607, (2007).
8. Ereteo, G., Buffa, M., Gandon, F., Corby, O. “Analysis of a real online social network using semantic web frameworks.” 8th International Semantic Web Conference (2009).
9. Gil, Y., Ratnakar, V., Deelman, E., Mehta, G. and J. Kim. “Wings for Pegasus: Creating Large-Scale Scientific Applications Using Semantic Representations of Computational Workflows,” Proceedings of the 19th Annual Conference on Innovative Applications of AI (IAAI), (2007).
10. Gil, Y.; Groth, P.; Ratnakar, V.; and Fritz, C. “Expressive Reusable Workflow Templates.” Proceedings of the Fifth IEEE International Conference on e-Science, Oxford, UK, (2009).
11. Gil, Y.; Ratnakar, V.; and Fritz, C. “Assisting Scientists with Complex Data Analysis Tasks through Semantic Workflows.” In Proceedings of the AAAI Fall Symposium on Proactive Assistant Agents, Arlington, VA, (2010).
12. Gil, Y.; Ratnakar, V.; Kim, J.; Gonzalez-Calero, P. A.; Groth, P.; Moody, J.; and Deelman, E. “Wings: Intelligent Workflow-Based Design of Computational Experiments.” IEEE Intelligent Systems, Vol. 26, No. 1, (2011).
13. Hassanzadeh, O., Kementsietsidis A, Lim L, Miller, RJ, and Wang M. “LinkedCT: A Linked Data Space for Clinical Trials,” 2009, http://arxiv.org/abs/0908.0567.
14. Huisman, M. and van Duijn, M.A.J. “A Reader's Guide to SNA Software.” To appear in P.J. Carrington and J. Scott (Eds.) Handbook of Social Network Analysis. SAGE (2010).
15. Kim, J., Deelman, E., Gil, Y., Mehta, G. and V. Ratnakar. “Provenance Trails in the Wings/Pegasus Workflow System,” Concurrency and Computation: Practice and Experience, Special Issue on the First Provenance Challenge, Vol 20, Issue 5, April 2008.
16. Martin, M.S., Gutierrez, C. “Representing, querying and transforming social networks with rdf/sparql.” In: Aroyo, L., Traverso, P., Ciravegna, F. (eds.) Semantic Web: Research and Applications. pp. 293–307 (2009).
17. Mika, P. “Flink: Semantic web technology for the extraction and analysis of social networks.” Journal of Web Semantics 3, 211–223 (2005).
18. Mika, P.; Elfring, T.; and Groenewegen, P. “Application of semantic technology for social network analysis in the sciences,” Scientometrics, vol. 68, pp. 3-27, July (2006).
19. Möller, K., Heath, T., Handschuh, S., Domingue, J.: “Recipes for semantic web dog food: the eswc and iswc metadata projects.” In: ISWC’07/ASWC’07: pp. 802–815. Springer-Verlag, Berlin, Heidelberg (2007)
20. Schmerl, B.; Garlan, D.; Dwivedi, V.; Bigrigg, M.; and Carley, K. “ SORASCS: A Case Study in SOA-based Platform Design for Socio-Cultural Analysis.” In Proceedings of 33rd International Conference on Software Engineering, (2011).
21. Scott, J. “Social Network Analysis: A Handbook. 2nd Ed.”: Sage (2000).
22. Taylor, I., Deelman, E., Gannon, D., Shields, M., (Eds). “Workflows for e-Science”, Springer, (2007).
23. Wang, S.; Groth, P. Measuring the dynamic bi-directional influence between content and social networks. The 9th International Semantic Web Conference 814–829 (2010)
24. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M. “DrugBank: a knowledgebase for drugs, drug actions and drug targets.” Nucleic Acids Research, Jan;36 (Database issue):D901-6, (2008).