Project Report

2019 DataONE Summer Internship Program
Project 6: A Reproducible Network Analysis of the DataONE Linked Open Data graph
Intern: Audrey McCombs
Primary Mentor: Bryce Mecum
Secondary Mentor: Dave Vieglais
Date: August 8, 2019

Introduction

This project was implemented from June 2019 through August 2019 as part of DataONE's summer internship program.

Project description

With over 800,000 datasets accessible through programmatic interfaces, DataONE provides a rich corpus of machine readable metadata that is also expressed as a linked open data (LOD) graph. The goal of this project was to explore the LOD graph of DataONE, provide a network analysis of the graph, and examine how the network differs from the content available through the traditional DataONE Application Programming Interface (API). Questions to be addressed in this project include: How interconnected are data sets and researchers? How many individual authors contributed to how many data sets? Can fields such as keywords be normalized to a small set of controlled vocabularies? How do network analysis measures differ by metadata standard, year of publication, or other facets?

Expected outcomes

- Identification of a set of key metrics or questions to apply to the DataONE linked open data graph
- A reproducible analysis of key metrics or questions in the form of a report that can be re-run periodically to track changes over time

Overview

The overall goal of this project was to conduct an exploratory analysis of the relational structure of datasets in DataONE member repositories. Specifically, we were tasked with creating reproducible code that would take as input a database query from a DataONE member repository, build a network from that query, and output a set of statistics describing important features of the network. Network graphs are made of two types of objects: nodes, and the links connecting nodes. These two types of graph objects were mapped to two types of objects in the database of a DataONE member repository: datasets, and people associated with each dataset. Depending on the type of network we're building, a "person" could be a dataset creator, a dataset contributor, or a record of a visitor to the DataONE website who searched for and/or downloaded a dataset. We use the term "creator" to specifically identify creators associated with a dataset, and the term "user" to generically identify a person who interacted with a dataset in some way.

To meet project expectations, we identified a set of key statistics describing important aspects of the relationships among nodes as captured in topological features of the network. We then created reproducible code to build the network and calculate the relevant statistics. Because the code is reproducible, the networks can be re-built and descriptive statistics re-calculated at regular intervals. Understanding how the topological characteristics of the networks change over time can provide insight into how DataONE's services are being used. Current non-network based metrics provide information about which member repositories are experiencing the most growth and where DataONE has been most successful in increasing the visibility of datasets. A network-based analysis allows DataONE analysts to understand the relational mechanisms that may be causing or contributing to growth and increased visibility. A better understanding of these mechanisms, provided by network topology analysis over time, can help DataONE predict and prepare for future trends in repository use.

An additional feature of this project that may be incorporated into future DataONE queries involves enhancements to DataONE's search engine through a "suggestions" feature. Network analysis can identify communities of datasets that are related to each other through creation, contribution, and download events. Once a researcher finds a dataset in DataONE, an enhanced search engine could potentially suggest other datasets belonging to the same community, thereby improving the researcher's user experience and advancing the ability of the scientific community as a whole to discover, access, and integrate archived datasets into current ecological and earth sciences research.

Methods and Results

The analytical results of this project were developed in R (R Core Team, 2019) using the `igraph` (Csardi & Nepusz, 2006) package. Network visualizations were created in the software package Gephi (Bastian, Heymann & Jacomy, 2009). Graph production is described in detail in the R Markdown document "MakeNetwork.Rmd," available on GitHub[1]. That document also contains the reproducible code for creating the graphs. Data used as input and tables of statistics returned as output are also saved in the repository.

As part of this project we produced three graphs:

1. ADC datasets: The Arctic Data Center[2] repository with datasets as nodes and dataset creators as links. This graph focused on the network structure of the datasets in the repository. Also of interest were the communities of datasets formed by dataset creators.

[1] https://github.com/DataONEorg/lod-graph-analysis/

[2] https://arcticdata.io/

2. ADC creators: The Arctic Data Center repository with dataset creators as nodes and datasets as links. This network focused on the topological characteristics of the network of individual dataset creators. From this graph we identified the individuals who are important to connectivity in the network, and who have created a large number of datasets archived in the repository.

3. DataONE Subset: A subset of all DataONE member repositories with datasets as nodes and users as links. Similar to the ADC datasets graph, the focus for this network was on communities of datasets formed by user interactions. This graph also includes an external dataset attribute—repository name—that we incorporated into the network analysis.

For each graph we produced network visualizations and calculated the graph statistics listed below. (Network statistics are described for a network with datasets as nodes and users as links; for the flipped ADC dataset creators network, substitute "datasets" for "users" and vice versa.)

1. Number of rows in the data table from the original repository database query, typically a .csv file
2. Total number of unique users in the repository
3. Number of users in the repository who only interacted with one dataset
4. Number of users in the repository who interacted with more than one dataset
5. Number of interaction events; that is, the total number of times all users interacted with the datasets in the repository
6. Number of nodes (datasets) in the graph
7. Number of links (users) in the graph
8. Median degree. A node's degree is the number of links connected to that node. Median degree is the median number of links per node over all nodes in the network.
9. Mean degree. Mean number of links per node over all datasets in the network.
10. Max degree. Maximum node degree over all datasets in the network.
11. Number of nodes with degree one; that is, the number of datasets with only one user of the dataset.
12. Network density: how close the network is to complete. This is the ratio of realized edges to the number of possible edges. A complete network has all possible edges.
13. Average shortest path length: The shortest path between any two nodes is the path between those two nodes that passes through the fewest number of other nodes. The length of the shortest path is the number of nodes the path passes through.
14. Network diameter: The longest shortest path on the network.
15. Number of components: The number of components in the network that are isolated from other components.
16. Overall modularity: Modules in a network are groups of links that are more connected among themselves than they are with the rest of the network. Modules are "communities" of nodes in the network. The overall modularity score for a network measures how "clique-y" the network is, as opposed to more evenly connected throughout.

The first five statistics describe aspects of the repository as a whole, independent of the network built from datasets in the repository. Statistics 6 through 16 quantify important aspects of the topology of the network. Tracking how these statistics change over time can provide important information about how the repository is being used by researchers. Details about how to interpret these statistics are included in the specific network descriptions below.

In addition to these repository- and network-level statistics, we also calculated two node-level statistics of interest: 1) degree centrality, which is the degree of the node, and 2) modularity class, which is the community to which the node has been assigned by a community detection algorithm. We ran two community detection algorithms: leading eigenvector clustering and walktrap clustering[3]. Leading eigenvector clustering works with a deterministic algorithm that conducts a spectral decomposition of a matrix related to the adjacency matrix of the network. Walktrap clustering works on a random walk algorithm and is stochastic in nature. Leading eigenvector clustering tends to find fewer communities of larger size than walktrap clustering. For the DataONE Subset network we also included in the table of node attributes the identity of the DataONE repository that hosts the dataset.

The network of ADC datasets

Figure 1 displays a visualization of the network of Arctic Data Center datasets produced in Gephi. In this graph visualization the colors represent modules, or communities of datasets identified by the leading eigenvector community detection algorithm described above. These are groups of datasets that are more connected with each other than they are with the rest of the graph, with connections being determined by creators of datasets in the repository. The interesting topological features of this graph are captured in the network statistics listed in Table 1.

We can interpret these statistics as follows: A database query to the Arctic Data Center repository produced a table of 9,237 unique dataset-creator pairs, with 2,862 unique creators. Of those creators, 2,224 created only one dataset, while 638 created more than one dataset. The repository query captured 6,998 events in which a unique user contributed to a dataset. The network contained 3,790 nodes and 170,719 edges. The median node degree was 35 while the mean degree was about 90, indicating that the degree distribution is strongly skewed with many low values and a few very high values. The maximum number of people who contributed to a single dataset is 372. One hundred and fifty-four datasets only had one creator. The network is only 2.38% complete, which is a low percentage and indicates the network is highly fractured—a characteristic also captured by the high number of separate components (165). For all the shortest paths between two nodes on the network, the average number of nodes the shortest path passes through is about 4, while the maximum number of nodes is 10. The network is highly modular, with a modularity score of 0.789, suggesting communities in the network are easily distinguished from one another.

[3] http://bioconductor.statistik.tu-dortmund.de/cran/web/packages/igraph/igraph.pdf

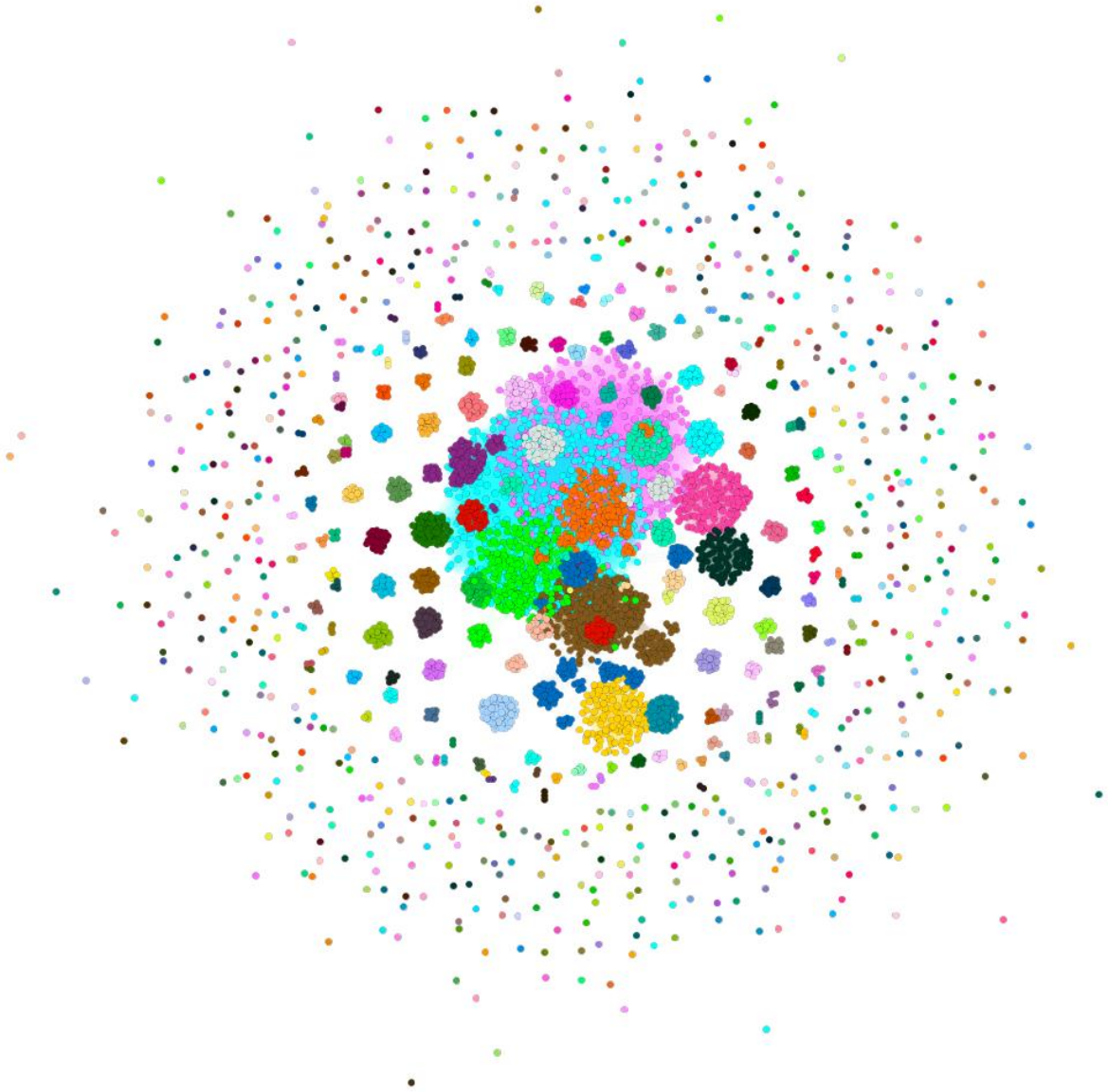

Figure 1: Visualization of the Arctic Data Center network of datasets. Colors indicate communities identified by Gephi's community detection algorithm.

Referring to the visualization above, these statistics capture important characteristics of the network, such as the fact that the network is highly fractured (2.38% complete) and there are many separate components (165 components). Many of the nodes have low degree (the nodes on the periphery of the network), while a relatively few in the middle have a high degree. In this network the communities are easy to distinguish, reflected in the network's high modularity score of 0.789.

Table 1: Statistics calculated for the Arctic Data Center network of datasets. "Users" indicate dataset creators.

Statistic	**Value**
num_rows_csv	9237
num_users	2862
one_dataset_users	2224
mult_dataset_users	638
interaction_events	6998
num_nodes	3790
num_edges	170719
med_degree	35
mean_degree	90.09
max_degree	372
num_degree_one	154
net_density	0.0238
avg_short_path	3.91
net_diameter	10
net_components	165
net_modularity	0.7889

We can examine the central core of the network by visualizing the giant component: the fully-connected component of the graph with the highest number of nodes, as displayed in Figure 2. In this visualization of the giant component, colors represent communities as before (although the colors do not match the colors in the whole-graph visualization above). For the giant component only, the community detection algorithm identified 31 communities of datasets, represented as spurs on the graph. Communities are primarily connected within themselves, however this visualization indicates that there are some connections among communities, especially between the red spur in the upper right and the green spur in the upper left.

The first fifteen rows of the node attributes table for the network of datasets in the Arctic Data Center repository are listed in Table 2.

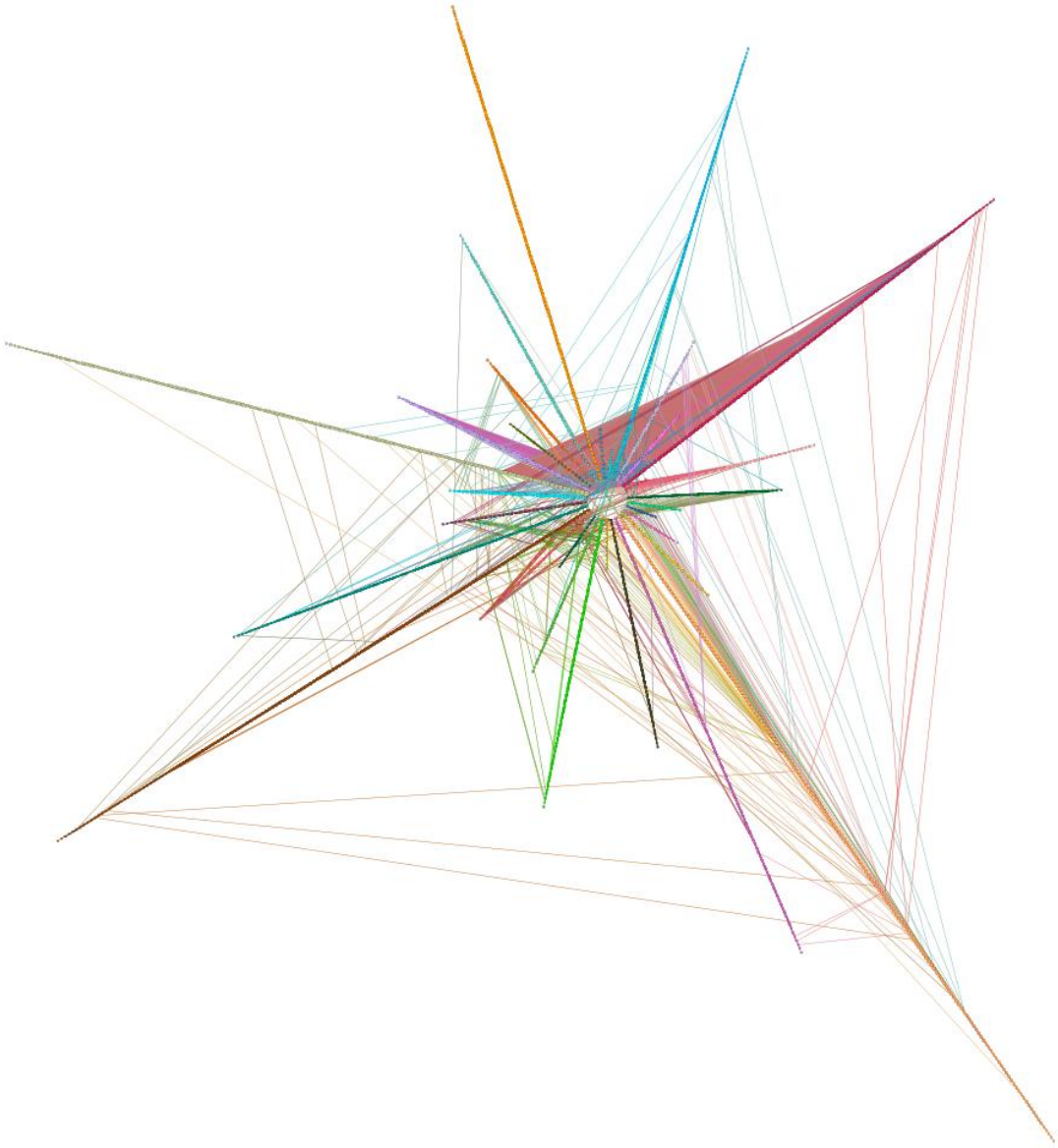

Figure 2: Visualization of the giant component of the Arctic Data Center network of datasets. Colors indicate communities identified by Gephi's community detection algorithm.

Table 2: Table of node attributes for 15 selected datasets in the Arctic Data Center network of datasets.

Dataset ID	Degree	Community identifier (from eigenvector algorithm)	Community identifier (from random walk algorithm)
doi:10.18739/A2000001N	1	1	81
doi:10.18739/A2027H	315	6	43
doi:10.18739/A2028PC7G	109	8	62
doi:10.18739/A2028W	19	9	138
doi:10.18739/A20298	25	167	24
doi:10.18739/A2033B	23	170	79
doi:10.18739/A20531	315	6	43
doi:10.18739/A2057CR7B	9	12	127
doi:10.18739/A2057CR99	109	8	62
doi:10.18739/A2058X	27	13	1
doi:10.18739/A2063C	32	170	25
doi:10.18739/A20819	331	11	82
doi:10.18739/A20B8N	331	11	82
doi:10.18739/A20C0Z	315	6	43
doi:10.18739/A20C1B	315	6	43

The network of ADC creators

The network of creators in the Arctic Data Center repository flips the previous network by setting creators as nodes and datasets as links. The network of ADC creators looks very similar to the network of ADC datasets, with similar topological characteristics. Figure 3 displays a visualization of the network of ADC creators.

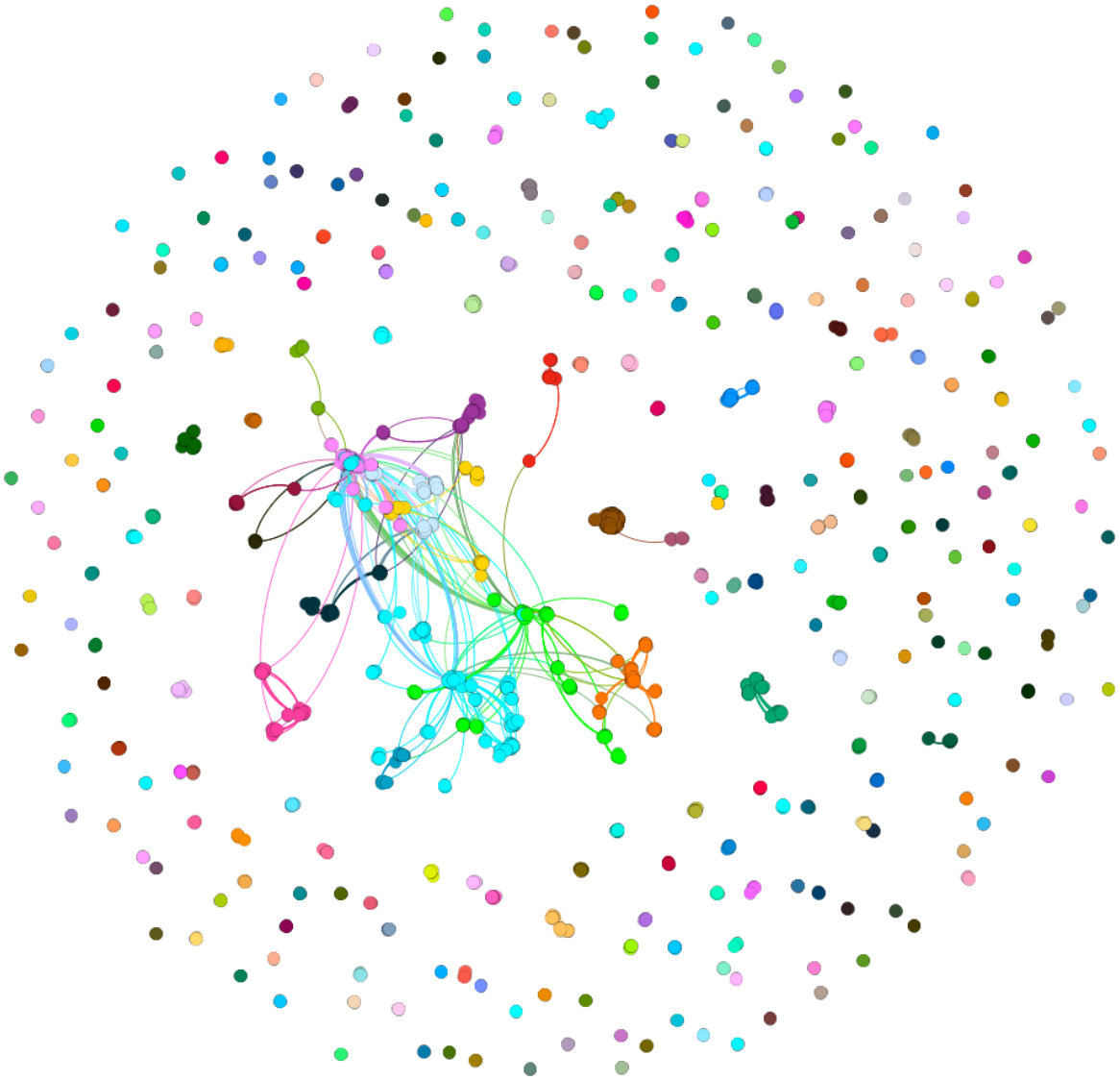

Figure 3: Visualization of the Arctic Data Center network of creators. Colors indicate communities identified by Gephi's community detection algorithm.

This network is smaller than the network of datasets in the Arctic Data Center, but otherwise the two graphs share a similar topology. The statistics for this network are listed in Table 3.

Table 3: Network statistics for the Arctic Data Center network of creators. "Users" in this table indicate dataset creators.

Statistic	**Value**
num_rows_csv	9237
num_datasets	4401
one_user_datasets	2478
mult_user_datasets	1923
interaction_events	6744
num_nodes	2368
num_edges	7530
med_degree	4
mean_degree	6.36
max_degree	618
num_degree_one	355
net_density	0.0027
avg_short_path	3.37
net_diameter	10
net_components	282
net_modularity	0.8132

We interpret these statistics as follows: This graph was produced from the same database query as the previous graph (with datasets as nodes and users as links), so the size of the database query is the same at 9,237 unique dataset-creator pairs, with 4,401 unique datasets in the repository. Of those datasets, 2,478 had only one creator, while 1,923 had multiple creators. The repository query captured 6,744 events in which a dataset was accessed by a creator. The network contained 2,368 nodes and 7,530 edges (two orders of magnitude smaller than the datasets network). The median node degree was 4 while the mean degree was about 6, indicating that the degree distribution is skewed with many low values and a few very high values. The maximum number of datasets created by an individual creator was 618. Three hundred and fifty-five creators only worked on one dataset. The network is only 0.27% complete, lower than the network of datasets by an order of magnitude, with 282 separate components. For all the shortest paths between two nodes on the network, the average number of nodes the shortest path passes through is about 3, while the maximum number of nodes is 10. The network is highly modular, with a modularity score of 0.813, suggesting communities in the network are easily distinguished from one another.

Questions of interest for this graph involve identifying individuals who are important to connectivity in the network, and who have contributed to a large number of datasets. Betweenness centrality is a measure of how important a node is to connectivity in the

network: it is proportional to the number of shortest paths involving that node. The individuals with the highest betweenness centrality on this network are listed in Table 4.

Table 4: Node characteristics for the top 10 dataset creators in the Arctic Data Center network of creators, ranked by betweenness centrality. Creator names shortened to the first three letters of the last name.

Last Name	**ORCID**	**Betweeness centrality**	**Degree**
Ash	https://orcid.org/0000-0002-7894-1519	80475.94	60
One	https://orcid.org/0000-0002-9185-0144	65509.33	618
Nol		63033.33	614
Zin		63033.33	614
Coo		48415.59	50
Wal		40165.21	157
Gre	https://orcid.org/0000-0001-7624-3568	26828.17	37
Sam		25918.95	30
Rac		21048.4	53
Mcg		20173.34	61

Individuals who contributed to a large number of datasets are nodes with the highest degree, as listed in Table 5.

Table 5: Node characteristics for the top 10 dataset creators in the Arctic Data Center network of creators, ranked by degree centrality. Creator names shortened to the first three letters of the last name.

Last Name	**ORCID**	**Degree**	**Betweeness centrality**
One	https://orcid.org/0000-0002-9185-0144	618	65509.33
Nol		614	63033.33
Zin		614	63033.33
Wal		157	40165.21
Moo		106	8811.465
Cop		84	5575.718
Knu		78	4341.398
Eps		76	3250.96
Cha		73	4144.241
Ber		72	3644.142

Table 6 lists the first twenty rows of the node attributes table for the network of creators in the Arctic Data Center repository.

Table 6: First 20 rows of the node attribute table for the Arctic Data Center network of creators.

Creator ID	Degree	Community identifier (from eigenvector algorithm)	Community identifier (from random walk algorithm)
001496a4-2475-47eb-99bf-2406093d744d	4	1	24
00884de9-1f56-48b9-8e0f-cfbf28ff24dd	1	289	31
008f43e8-115c-4e77-8afa-e345ba29702a	4	1	24
0129ac22-92a4-4d38-b201-32b0275145fd	7	289	1
012d9ccb-e332-4ada-9c0f-d3ab34cad14f	4	1	24
014b891b-d294-4e20-82d5-c02ef2c66196	50	289	6
01785d43-fe34-4d06-b05a-a1d6c11948a0	2	235	214
0186359b-1cde-4860-8f71-5a2375e1194c	45	283	6
018dba33-4fd7-4739-a3e1-6081a0627071	3	3	6
01b1f7cd-87da-4acb-a5be-3490d5b3a87d	4	1	24
01b93416-fa0e-45d5-ae10-016a4405ad61	2	231	51
01c5ea7e-a99f-43b9-b90c-cf71eb28d7d1	10	114	257
01d78fed-b47f-46f9-8d6c-8f867f8d3047	4	1	24
01ef51b5-dd82-4976-8dee-c371ca495420	2	94	145
01fbc405-c3e7-488e-9b72-7ed6859f85d3	4	1	24
0256ceaf-678e-43b8-a621-72532791604e	3	263	49
025a206d-dbef-41fe-947e-1b29e5f75ae1	6	283	6
0261d004-187f-4312-9703-c0d84bce2d1f	7	289	81
028bd679-54fb-4093-8c9b-2055650f5e9f	4	1	24
0290c543-7225-49a2-b88f-f286cd784571	1	189	305

The DataONE Subset network

A network for the entire DataONE repository was too large to build with the code developed for this project, so we built a network from a subset of the entire repository. Specifically, we included in the network 23 of the 29 total DataONE member repositories, listed below.

- BCODMO
- EDI
- ESA
- ESS_DIVE
- FEMC
- FIGSHARE_CARY
- GOA
- GRIIDC
- IARC
- IEDA_MGDL
- IEDA_USAP
- mnORC1
- mnUCSB1
- mnUNM1
- NCEI
- NKN
- NMEPSCOR
- ONEShare
- PPBIO
- RW
- SEAD
- TFRI
- UIC

Member repostiories that were removed from the network:

- ARCTIC
- KNB
- LTER
- PANGAEA
- R2R
- TERN

We built a network for the Arctic Data Center as described in the two previous sections of this report, and a network for R2R and TERN could be built in the same way. Networks for KNB, LTER, and PANGAEA were too large for the code we developed as part of this project. We calculated the size of the edge lists for these three networks as: 104,950,604 rows for the KNB repository; 921,952,313 rows for the LTER repository; and 69,724,637 rows for the PANGAEA repository. The size of the edge list for the entire DataONE repository would be approximately 1.2 billion rows. Networks for these repositories should be implemented under a distributed computing paradigm, to be developed at a future date.

To build this network of 23 DataONE member repositories, we queried the entire DataONE database and used session ID's as proxies for user information, assuming that datasets queried together in the same session were related. Both visually and statistically, this network has a very similar topological structure as the networks from the Arctic Data Center: the graph is highly modular (modularity score: 0.640), fractured (272 components), and disconnected (0.54% density).

We produced two visualizations of the network of 23 repositories: the first colors the nodes by modularity class while the second colors the nodes by DataONE repository ID. Figure 4 presents the network with modularity classes identified at the top, and the same network with repository identified at the bottom. A comparison of the two suggests that datasets in particular repositories do form communities based on user interactions.

Network statistics calculated for this network of 23 DataONE member repositories are listed in Table 7.

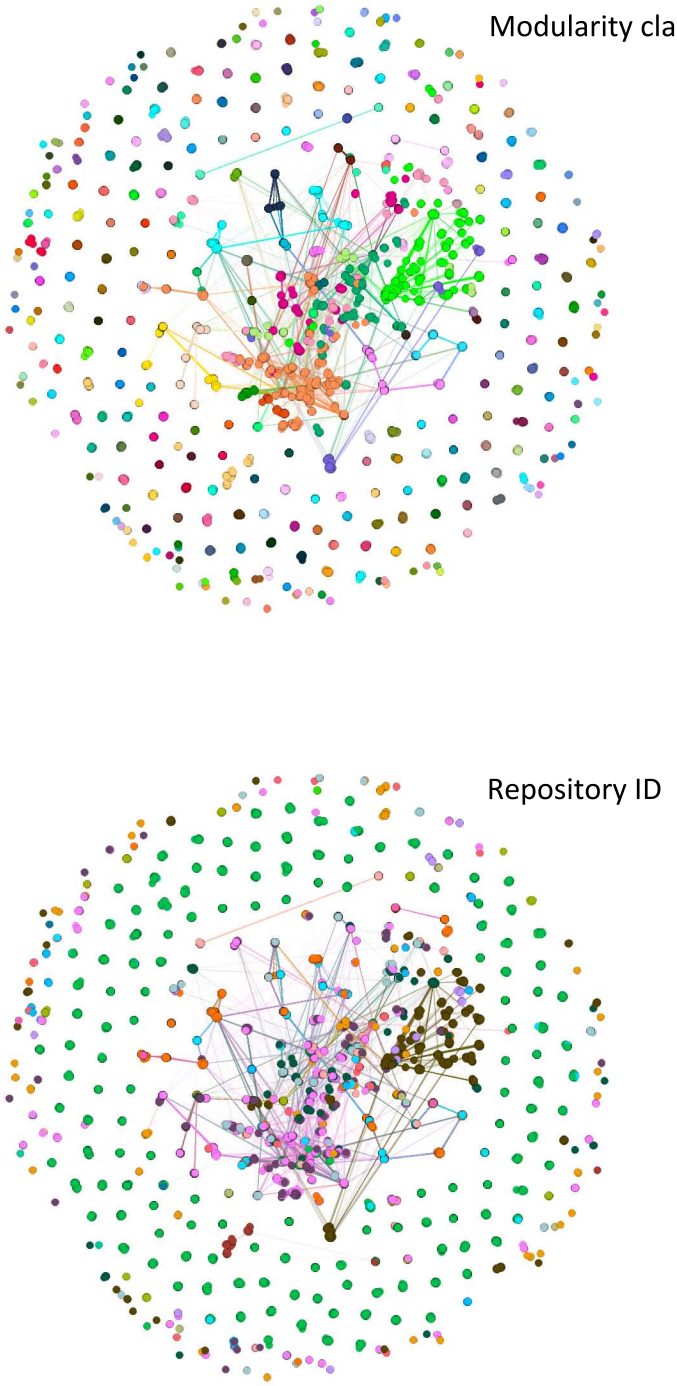

Figure 4: Two visualizations of of the network of 23 DataONE repositories. The top visualization colors nodes by modularity class (as determined by an `igraph` community detection algorithm), while the bottom visualization colors nodes by DataONE repository ID.

Table 7: Network statistics for the network of 23 DataONE member repositories.

Statistic	Value
num_users	6717
one_dataset_users	3132
mult_dataset_users	3585
interaction_events	57181
num_nodes	29454
num_edges	2342733
med_degree	104
mean_degree	159.08
max_degree	890
num_degree_one	222
net_density	0.0054
avg_short_path	5.09
net_diameter	15
net_components	272
net_modularity	0.6399

As would be expected, this is a very large graph, with 29,454 nodes and 2,342,733 edges. We do not recommend constructing a graph with more than 1 million edges with the code developed for this project—the .csv file with the final edge list for this network was approximately 51MB, and too large to load completely in Microsoft Excel (which has a hard limit of just over 1 million rows in a spreadsheet.) R was able to create both the edge list and the graph for this network, but analysis was slow, memory intensive, and caused R to crash several times. Visualizations in Gephi were also difficult.

A sample of the attributes table (Table 8) created for this network suggests that the leading eigenvector community detection algorithm tends to place datasets from different repositories into the same community, whereas communities identified by the random walk algorithm match repository identities more consistently, although not perfectly.

Project challenges and data gaps

The primary challenge we encountered in this project was caused by the sheer size of the networks we were working with. As noted above, networks for several of the DataONE repositories are too large for computing in R using consumer-grade hardware. The R package `sparklyr` (Luraschi et al. 2019) implements distributed computing tools using the Apache Spark analytics engine for large-scale data processing. We note that `sparklyr` also interfaces with `GraphFrames` (Kuo 2018), a graph processing library for Apache Spark that provides network analysis tools similar to `igraph`.

We attempted to use `sparklyr` while building the edge list for the larger repository networks, by saving the edge list in pieces: smaller files saved in distributed memory and accessed as a whole through the Spark interface. We were not able to successfully

Table 8: First 20 rows of the node attribute table for the Arctic Data Center network of creators.

Dataset ID	Degree	Community identifier (from eigenvector algorithm)	Community identifier (from random walk algorithm)	Repository name
114731	287	1	121	NMEPSCOR
114734	70	1	6	IARC
114735	307	1	121	mnUCSB1
114737	156	1	7	NMEPSCOR
114739	499	1	59	NMEPSCOR
114741	326	1	59	NMEPSCOR
114742	369	1	59	NKN
114743	132	1	7	IARC
114744	109	1	238	NMEPSCOR
114749	141	1	7	NMEPSCOR

integrate the `sparklyr` computing tools into our code, however. We believe that using Spark (or a similar big data analytics engine) will allow the construction and analysis of these larger networks, and that integrating our code with a big data analytics engine is an important next step for DataONE. We expect that future analysts will need to coordinate with DataONE's information technologists and network administrators to develop a storage architecture that will allow R's analysis tools to be applied to very large datasets.

A second challenge we encountered on this project was related to linked open data. One of the questions driving this project asked: "Can fields such as keywords be normalized to a small set of controlled vocabularies?" When we examined the keywords associated with the datasets in the Arctic Data Center, we discovered that the keywords were in different formats and used varying vocabularies. For example, the keywords field for one dataset read: "Earth Science > Physical Limnology > temperature Earth Science > Physical Limnology > specific conductance Earth Science > Physical Limnology > dissolved oxygen." The keyword field for a different dataset read: "aquatic arctic fire carbon nitrogen alaska lake fen isotopes." Standardizing the list of keywords, both in terms of content and format, will be time-consuming. But once the keywords are standardized they can be used to define communities in the repository networks, and different community structures defined by different community-assignment techniques can be compared.

Currently, DataONE's training module "Lesson 7: Metadata" includes information on the proper use of keywords in metadata. Continuing to train researchers in the importance of good metadata, standardizing legacy keyword fields, and implementing a standard format for the future will all increase the usefulness of keywords in future network analysis.

Conclusion

The overall goal of this project was to conduct an exploratory analysis of the relational structure of datasets in various DataONE repositories. To meet this goal, we created reproducible code to build and analyze three networks of DataONE datasets. We found the topology across all networks to be very similar, differing primarily in the size of the network as measured by the number of nodes and edges. The code we developed can be re-used so that repository networks can be re-built and re-analyzed over time, to create a time series of network statistics that describe how the relations among datasets, as mediated by users, change over time.

The expected outcomes of the project were met in the following ways:

1. Expected outcome: Identification of a set of key metrics or questions to apply to the DataONE linked open data graph. Methods and results: We identified sixteen network statistics to describe how users mediate the relations among datasets in a repository.

2. Expected outcome: A reproducible analysis of key metrics or questions in the form of a report that can be re-run periodically to track changes over time. Methods and results: We created reproducible code in the form of an R Markdown document stored in a DataONE GitHub repository. That code builds a network from a repository query, calculates the statistics identified in step 1 above, and outputs a table of network statistics that can be tracked over time.

In addition to the expected outcomes, we created the following additional products:

1. Network analyses and visualizations of three specific networks: the Arctic Data Center network of datasets, the Arctic Data Center network of creators, and a network of datasets from a subset of the entire DataONE repository.

2. Code to create an attributes list of community identifiers for datasets in the network. We produced this attributes list for the three networks we created. Future enhancements to DataONE's search engine could incorporate this information into a "suggestions" feature in DataONE Search.

We identified two areas in which this project could be further developed. The first involves integrating big data computing tools into the code we produced so future analysts can work with very large networks (i.e., networks with edge lists of over 1 million rows). The second area for development involves standardizing legacy keyword fields and implementing these standards in future metadata documents, so that keywords can be incorporated into network analysis.

We believe that this project successfully developed the foundational tools for a more nuanced understanding of how DataONE member repositories are used by researchers. A network-based analysis will allow DataONE analysts to better understand the relational mechanisms contributing to growth and increased visibility of scientific datasets. A better

understanding of these mechanisms can help DataONE predict and prepare for future trends in repository use.

References

Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

Csardi G., Nepusz T. (2006). The igraph software package for complex network research. *InterJournal*: Complex Systems, 1695. http://igraph.org

Kuo, K. (2018). graphframes: Interface for 'GraphFrames'. R package version 0.1.2. https://CRAN.R-project.org/package=graphframes

Luraschi, J., Kuo, K., Ushey, K., Allaire, J. J., the Apache Software Foundation. (2019). sparklyr: R Interface to Apache Spark. R package version 1.0.2. https://CRAN.R-project.org/package=sparklyr

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/