Capstone Project 1: Displaying Large Organizational Networks – Final Report

The Challenge of Communicating Large Organizational Networks

Networks can be a powerful tool for understanding, analyzing, and modeling the complex interactions that go on between organizations. They can even be deployed to quickly communicate broad information about the patterns that those interactions can take. But, most real-world networks are large, particularly those involved in making and implementing public policy, and those same network maps that can quickly communicate structures can just as easily hide both structure and detail once the map becomes large. This undermines their effectiveness as a communication tool.

That effectiveness can be reclaimed if the network is presented in the best way. One way that has shown signs of effectiveness is to break the network into subnetworks and only display a targeted portion of the network at any one time. But, this, too, presents a problem: an analyst is required who can take requests or anticipate what subnetworks will be needed, then build and supply those separate subnetworks. Instead, what is needed, is a tool that will allow anyone who accesses the tool to explore the network on their own, only asking for greater analysis when their interests go beyond what can be seen in the tool. In this project, I will build this tool using empirical data gathered from Vermont's portion of the Lake Champlain Basin. The tool can now be accessed online: https://wmirecon.shinyapps.io/vgn-v5_10oct17/.

Data

Context

In 2014 and 2015, researcher with the Vermont Experimental Program to Stimulate Competitive Research (Vermont EPSCoR),[1] a branch of NSF EPSCoR, gathered two editions of an organizational network survey which gathered data on the networks that make and implement water quality policy in Vermont's portion of the Lake Champlain Basin (LCB). During that time, Vermont was the focus of actions by the federal Environmental Protection Agency (EPA). The Clean Water Act (CWA) requires jurisdictions to ensure that its waters are sufficiently free of pollution. Thresholds for how much of different pollutants can be handled in a waterbody are analyzed and set by the EPA through a Total Maximum Daily Load (TMDL). As the name indicates, the document reports the maximum amount of a pollutant that the waterbody can receive daily and remain clean. It assesses the natural processes in the waterbody and measures how much pollutant the waterbody can process in a day. Vermont has never met its obligations

[1] I personally led this effort, leading the drafting of the survey instrument, outreach to survey respondents, and data preparation, cleaning, and analysis.

under the CWA in the LCB and, several years before the survey occurred, legal action overturned Vermont's previous Phosphorus TMDL for the LCB. The legal challenge indicated that the previous TMDL overestimated the amount of pollution that the lake could process and failed to provide sufficient assurances that the targets could be met. This action forced state policy and how the CWA requirements could be met to the forefront, making water quality management a central policy issue in Vermont.

Organization Data

The survey was executed as an online survey, with organizational representatives selected for their organization's participation and approached through personal outreach. Since these data were gathered with federal research dollars the data is required to be made available and, in accordance with that requirement, a de-identified version, which masks organizational names and does not include the names or contact information for organizational contacts, is available for free distribution, along with a full codebook.[2] Table 1 lists the four organizational groups into which we sorted our respondents and reports the rate at organizations responded, by group. These groupings provide a means to reduce the burden on survey respondents and some information about the structure of system, showing that organizations fall into four broad categories, including two for broad-based organizations (governmental programs and NGOs) and two for smaller, geographically-focused organizations (the two watersheds).

Table 1: Survey Response Rates

Organizational Group	Number of Contacts		Completed Responses		Response Rate (%)		Observation Rate (%)[3]	
	2015	2014	2015	2014	2015	2014	2015	2014
Governmental Programs	53	56	30	26	56.6	46.4	81.6	71.75
Regional Actors and NGOs	51	50	24	26	47.1	52.0	72.5	73.47
Winooski Watershed	52	52	29	11	55.8	21.2	80.9	38.16
Missisquoi Watershed	34	40	12	12	35.3	30.0	58.8	51.54
Total	**190**	**198**	**95**	**75**	**50.0**	**37.9**	**75.1**	**60.26**

A total of 206 organizations were identified as participating in the water quality management in Vermont's portion of the LCB. For each organization, several characteristics, or node attributes, were also recorded, including a measure of the organization's budget and staff size, the

[2] Data are found here: https://github.com/wmirecon/Water_Quality_Governance_Networks

[3] Observation Rate records the percentage of non-directional network links that the survey was able to observe by obtaining a response from at least one of the two organizations involved in each link. See: Scheinert, S., Koliba, C., Hurley, S., Coleman, S., and Zia, A, 2015, The shape of watershed governance: Locating the boundaries of multiplex networks. *Complexity, Governance & Networks*, 2(1), 65-82. doi: 10.7654/15-CGN25.

organization's sector (such as public, private, or non-profit), the organization's geographic jurisdiction (Vermont, New York, Quebec, USA, Canada, etc.), and jurisdictional level (municipal, regional, state/province, federal, international), using numeric labels. Table 2 presents of slice of this data.

Table 2. Organization Data Example: First 15 Organizations

org.name	group	capacity	sector	jurisdiction	juris.level	winooski	missisquoi
g321601	Government	3	2	1	6	Yes	Yes
g121623	Government	1	2	1	6	Yes	Yes
g321603	Government	3	2	1	6	Yes	Yes
g321602	Government	3	2	1	6	Yes	Yes
g321605	Government	3	2	1	6	Yes	No
g321200	Government	3	2	1	2	Yes	Yes
n166600	NGO	1	6	6	6	No	No
g321604	Government	3	2	1	6	Yes	Yes
w241201	Winooski	2	4	1	2	Yes	No
n451800	NGO	4	5	1	8	No	No
m241201	Missisquoi	2	4	1	2	No	Yes
n454701	NGO	4	5	4	7	Yes	Yes
w261401	Winooski	2	6	1	4	Yes	No
w231402	Winooski	2	3	1	4	Yes	No
w161401	Winooski	1	6	1	4	Yes	No

Organization names, as seen in Table 2, are anonymous identifies and fully explained in the codebook. The *winooski* and m*issisquoi* variables identify whether or not that organization is active in those two basins while the organization's grouping is identified in the *group* variable. From our interactions with stakeholders while studying this network and building the survey, we learned that stakeholders are most interested in the sector attribute, which does closely link to the group categories. Table three takes a closer look at the breakdown of organizations by sector, which will also be used later in the network maps that are presented below.

Table 3. Sector Data: Sector Frequencies

Sector	Frequency
Local gov't agencies	57
NGO/Non-Profit	56
State/Provincial agencies	49
Profit/For-Profit	23
Regional gov't agencies	12
Federal agencies	4
Research Inst.	4
Int'l gov't agencies	1

Link Data

Each respondent was presented with a list of all the organizations in the survey. Organizations were presented in the survey by group, reducing the burden on respondents by only presenting more limited lists. Respondents indicated which organization's the respondent's organization shared information, provided technical assistance, collaborated or coordinated on projects, provided reports of their operations, and shared financial resources. We derived five different functional subnetworks, one from each of these types of interactions. After processing, network data were stored in five different "edgelists",[4] which record the initiating ("from") and receiving ("to") organizations for each network link, or "edge." Table 4 provides a slice of the edgelist for the information sharing functional subnetwork:

Table 4. Edgelist Example: First 10 links in the information sharing edgelist

from	to
n281600	n281800
n281600	n266500
n281600	n266500
n281600	n266500
n281600	n281600
n281600	n351700
n281600	n454800
n281600	n464800
n281600	n261607
n281600	w157600

The first questions to ask are how many organizations are in the network and how many links does it have. This gives a basic concept of what this network looks like. The number of links must depend on the number of organizations, since links can only exist between organizations and, whether or not a link is has some value is not part of counting the number of links. The density is, out of the total possible number of links that could exist, how many do exist, expressed as a percentage:

$$D = 2\frac{linkCount}{n(n-1)}$$

where n is the node count and links are non-directional.[5] In these networks, we're using a common organization list, so the number of nodes will be the same in each subnetwork,

[4] Processing was carried out using R prior to this project. It was repeated using Python and can be viewed here:

- R version: https://github.com/wmirecon/Visualizing_Gov_Nets/blob/master/data%20prep%20r%20code.R
- Python version: https://github.com/wmirecon/Visualizing_Gov_Nets/blob/master/ipython%20script%20prep%2014aug17.ipynb
- Data wrangling report, which explains the process in detail: https://github.com/wmirecon/Visualizing_Gov_Nets/blob/master/data%20wrangling%20report.pdf

[5] If the links are directional, then the value is not multiplied by 2.

specifically, 206 nodes for the 206 organizations. Table 5 reports the node and link counts and the density for each functional subnetwork.

Table 5. Functional Subnetwork Statistics

Subnetwork	Node Count	Link Count	Density
Information Sharing	206	3185	0.151
Technical Assistance	206	2331	0.110
Reporting	206	1047	0.050
Financial Resource Sharing	206	1323	0.063
Project Coordination & Collaboration	206	2092	0.099

Displaying Large Networks in Static Images[6]

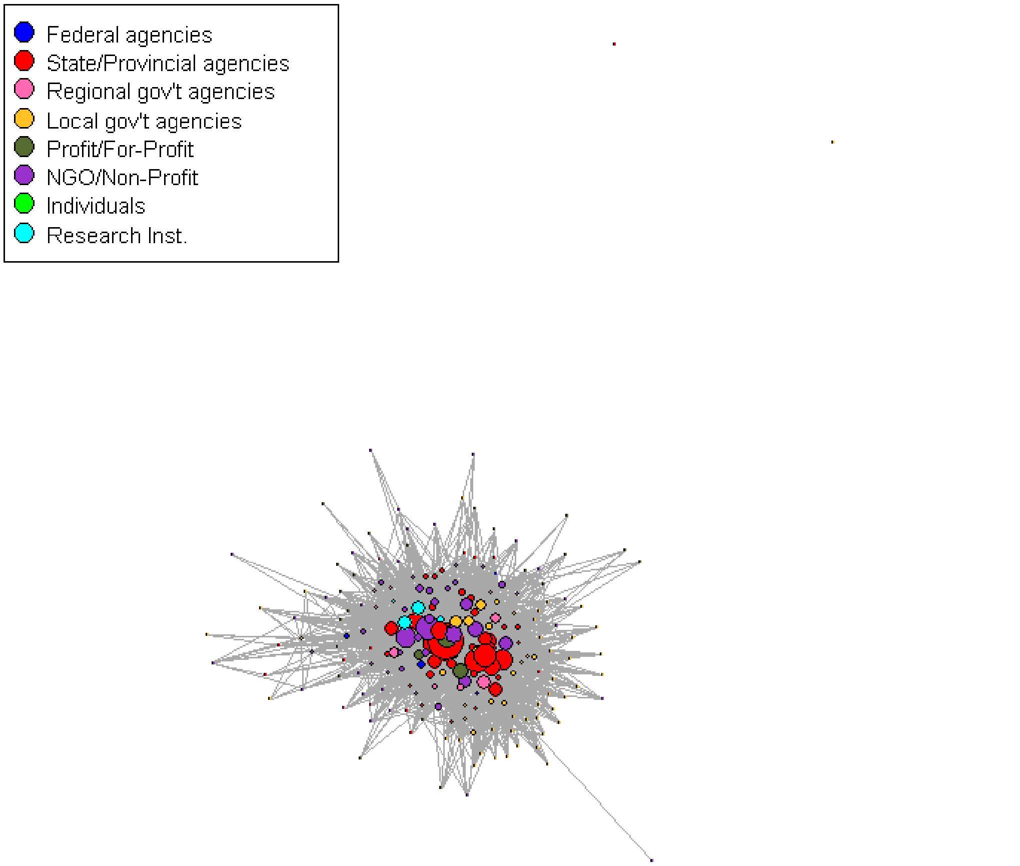
Often the fastest and most effective way to look at network data, and indeed to quickly see many of the characteristics of the network, is to look at that network's map. The node and link counts tell us that we have large networks, though very sparse networks. Even the densest has only about 15% of the total possible links. But this information is minimal, and only truly useful to analysts. Maps are a better means of communicating with non-specialists.

Large, Unreadable Maps

Network analysts have a number of different creative names for network maps that are too large or two dense to convey any meaningful information. Some prefer “Persian rugs” and others prefer “hairballs.” Figure 1 presents an example; the figure includes a static map of the information subnetwork. Information is frequently the largest and densest network in this context. It is the easiest task for organizations to engage in and the one with both the lowest cost and fewest risks. We see this here. There are many well-connected nodes, coming from a wide range of sectors. It is also clear that the network has a strong center of particularly well-connected organizations, with a periphery of many more less-well connected organizations. We can see that this center is dominated by state agencies and NGOs, but it is very difficult to see much beyond this, and certainly not without organization names included. In short, a map like this can tell us about some of the structures of the network but little about the details while it was those very details that many of the stakeholders in Vermont asked about when they first saw these network data.

[6] A milestone report was previously prepared and lays the discussion on the limits of network maps in detail. It can be viewed here: https://github.com/wmirecon/Visualizing_Gov_Nets/blob/master/Capstone%201%20-%20Visualizing%20Governance%20Networks%20Milestone%20Report.ipynb

Figure 1. Information Sharing Functional Subnetwork



There are three alternate approaches to addressing this issue:

1. Make bigger maps
2. Use interactive mapping
3. Make network slices

We could, of course, make physically larger maps. But there is a limit to how big these maps can be made, especially without printing them. And, indeed, they would have to be very large to be of use. Interactive mapping can allow us to focus on parts of the whole by zooming in, providing only certain labels, and allowing for nodes to be manipulated. But it has limits, particularly that a network this big will be slow and cumbersome to manipulate. Also, it would fall on the analysts to be present for each stakeholder's individual exploration. Network slices allow for a more focused look on specific pieces, whether their focused on a substantively-based piece of a network or the local neighborhood of a specific node. The former of these focuses, the substantively-based pieces, is something that can be anticipated, as there is only a limited selection of these pieces.

Smaller, Readable Maps

The survey that collected these data, along with asking about interactions, asked about organizations' use of a range of policy tools and involvement in a range of policy domains, action arenas, and geographic areas. Policy tools are the different mechanisms that policy makers write into policy so that that policy will have an impact once implemented. Regulations are the best known, but they also include measures like cost sharing, grants, loans and guarantees, permitting, and public information. The survey collected data on 12 such tools, with the full list included in the data's codebook. Domains are policy issues that around which policy is made. In the survey, six domains, such as agricultural land management, river corridors, and development were included. Action arenas are the spaces in which organizations interact. Many are physical spaces, though this is not necessary. This dataset includes another 12 action arenas, including several different legislative committees, regional planning commissions, the US EPA-initiated Total Maximum Daily Load (TMDL) revision and implementation planning, Vermont's tactical basin planning, and the Lake Champlain Basin Program. For demonstration purposes, we will look at a map of one of these: the EPA-initiated TMDL, shown in Figure 2, as this is a key map for stakeholders' interest in the drafting of the new TMDL and its associated implementation plan.

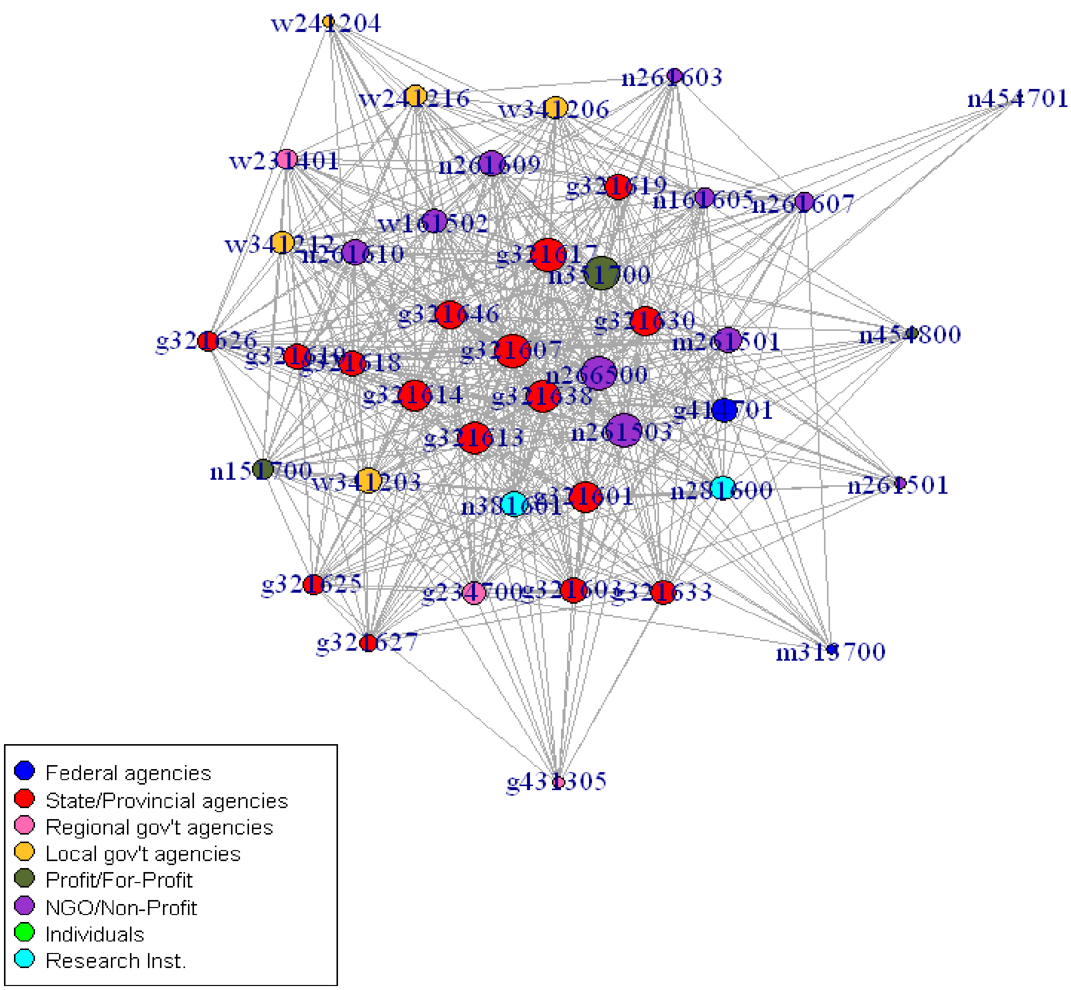


Figure 2. EPA Initiated-TMDL Action Arena Subnetwork Map

This is a much more readable map. It is still dense and still has problems with node label overlaps, but it is far easier for a non-specialist to look at this map and learn something about how specific organizations are interacting and embedded in this network.

Application and Features

But this is still not a map that a non-specialist can explore. That requires a different approach: an interactive map. With 12 policy tools, 12 action arenas, and 6 policy domains, there are almost 1,200 combinations that stakeholder may find interesting, for each functional subnetwork, and not considering any focus organizations. If stakeholders want to know more about the specific network neighborhoods for any of the 206 organizations, then the number of possible networks explodes. What is needed, instead, is for the tool to allow stakeholders to manipulate the network themselves and explore where they each want to know more. For this, an interactive map is needed and it needs to be hosted in a widely-available space where stakeholders can reach the interactive mapping application. Any other method will swamp the analyst in responding to stakeholders' separate requests for unique maps, many of which will still be difficult to read.

This application has been completed for this network dataset and is available for viewing at: https://wmirecon.shinyapps.io/vgn-v5_10oct17/. It asks for a user to select which functional subnetwork they would like to display. Once this is selected, the application produces a network map. Nodes display a color based on their sector and, when hovered over, give the percentage of other nodes with whom they share a connection. That is, the map displays the organization's name and its normalized degree centrality measure, which is measured as the number of links a node has divided by the maximum possible number of links it could have. Figure 3 displays a screen shot of the information sharing network, as displayed in the application.

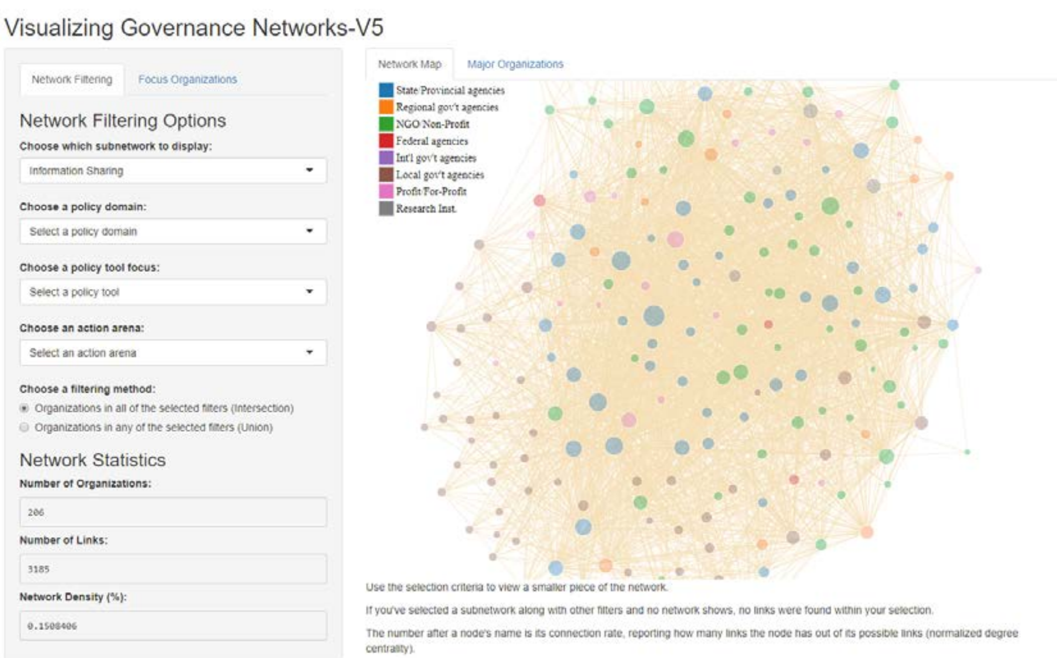


Figure 3. Application Display with the Information Sharing Functional Subnetwork

Network Filtering

Along the left side of the application is the main control panel. As seen in Figure 3, the map is still very large, and so it is difficult to see much. The dynamic nature of the map does allow the user to zoom in on the map, see details on individual organizations, and pull within the map to see how the map responds to changing layouts. But more detail is still needed to see whole structures. The left side panel contains the filtering controls for this purpose. Whatever domains, tools, or any combination of domain, tool, and arena, or action arenas the user would like to see can be chosen here. The application responds and produces a new map. Figure 4 gives an example of the application's display, using the same TMDL action arena as Figure 2.

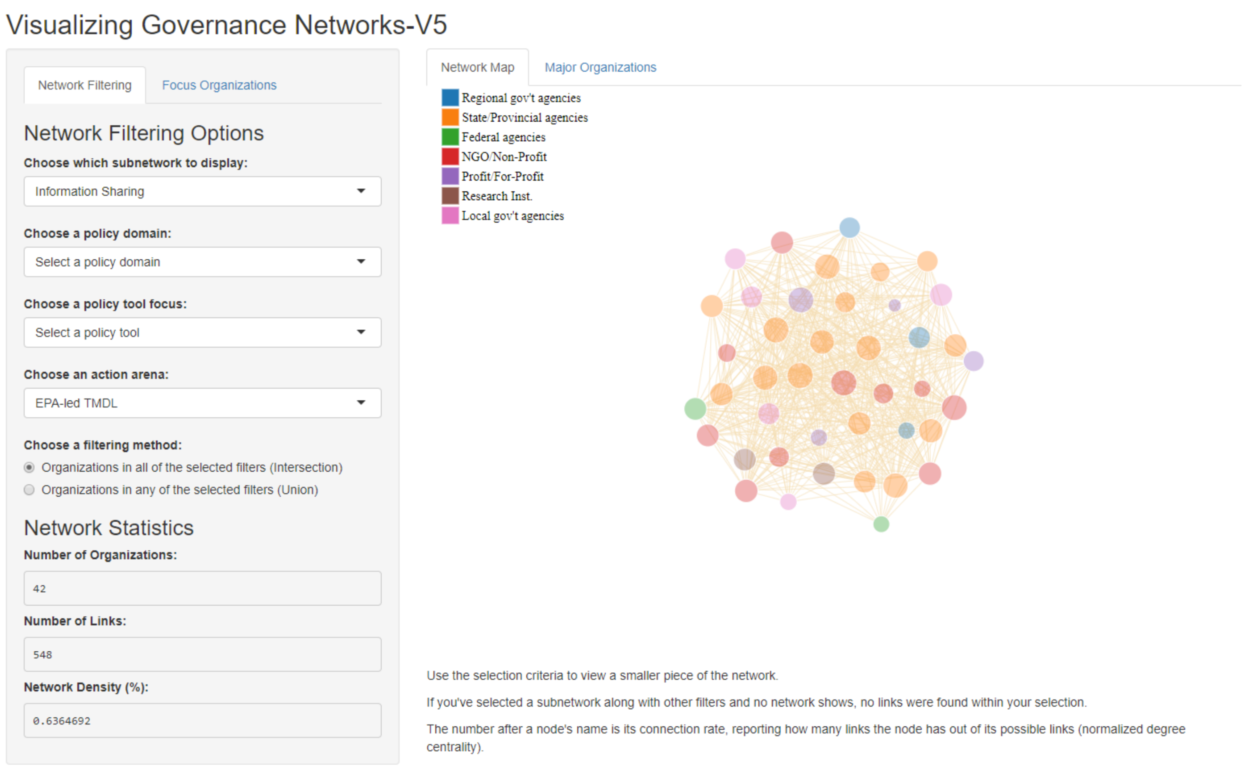


Figure 4. EPA-Led TMDL Action Arena Subnetwork Displayed in the Application

Focus Organization Selection

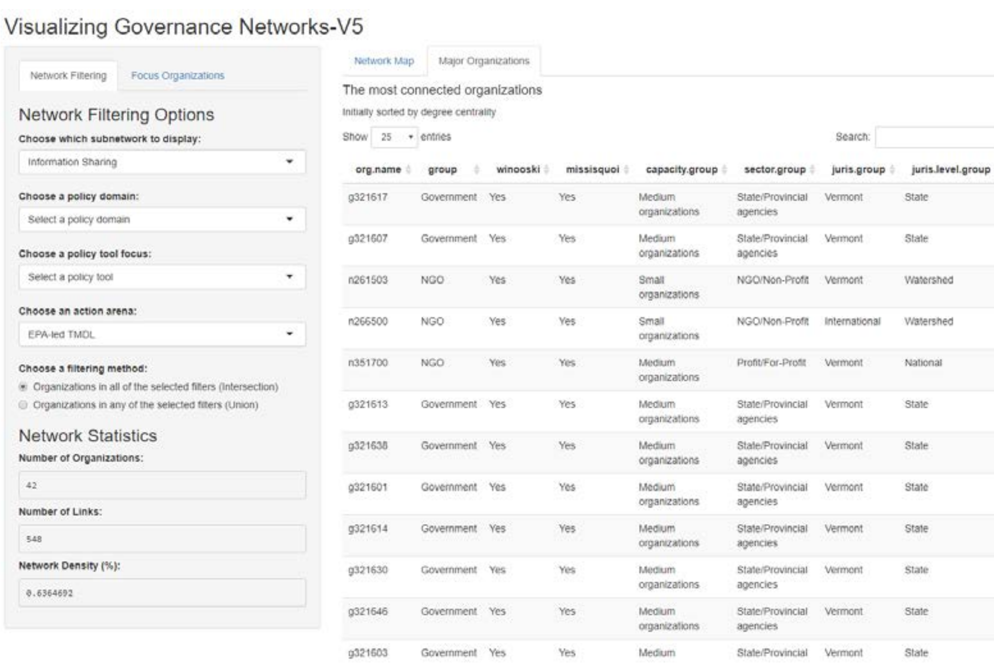
Along with the subnetwork filtering for domains, tools, and action arenas, the application allows users to select one or more focus organizations. When this is done, the map displays what network analysts refer to as an "egonetwork," the network organizations that all connect to the focus organization, including the focus organization. By selecting the 'Focus Organization' tab on the left side panel, users view a list of all 206 organizations. Users can select specific organizations as a focus and see their egonetwork. This network is drawn from the subnetwork filtering so that users see the egonetwork from within their selected subnetwork filters. If more than one organization is selected, then the user will see the egonetworks for all of the selected organizations.

Network Statistics Reporting

In the lower left-hand corner, the application reports some basic network-level statistics for the network that is currently being displayed. Specifically, it provides the number of organizations in the network, the number of links in the network, and the density. This output provides the user with some quick measurements of the network they are looking at and so helps by providing some additional information about the size and cohesiveness of the network.

Organizational Detail Reporting

Finally, under the "Major Organizations" tab, the application reports detailed information for all of the organizations present in the user's selected network. This includes all of the organization's attributes: what group it is in, whether or not it works in the Winooski and Missisquoi watersheds, and what capacity, sector, jurisdiction, and jurisdictional level. Figure 5 presents an image of this output. Additionally, the application reports all the organization's centrality scores. These scores are the way that analysts identify the most important nodes in a network. The application reports three centrality measures: degree, betweenness, and eigenvector. All are normalized for comparison across networks. Degree, as mentioned above, is based on the number of links that a node has. Organizations with a high degree can directly reach many other organizations. Eigenvector is similar to degree, but factors in the degree of the node's neighbors. Organizations with a high eigenvector can quickly reach a large portion of the network, since not only is that organization well-connected, but so are its neighbors. Betweenness is based on flows through the network. There is a shortest path from every organization in the network to every other organization. Betweenness counts how many of those paths flow through each organization. Organizations with high betweenness are strategically located to monitor and control or influence the flows of information and resources through the network. The list is initially sorted by degree centrality and can be re-sorted by the user.



org.name	group	winooski	missisquoi	capacity.group	sector.group	juris.group	juris.level.group
g321617	Government	Yes	Yes	Medium organizations	State/Provincial agencies	Vermont	State
g321607	Government	Yes	Yes	Medium organizations	State/Provincial agencies	Vermont	State
n261503	NGO	Yes	Yes	Small organizations	NGO/Non-Profit	Vermont	Watershed
n266500	NGO	Yes	Yes	Small organizations	NGO/Non-Profit	International	Watershed
n351700	NGO	Yes	Yes	Medium organizations	Profit/For-Profit	Vermont	National
g321613	Government	Yes	Yes	Medium organizations	State/Provincial agencies	Vermont	State
g321638	Government	Yes	Yes	Medium organizations	State/Provincial agencies	Vermont	State
g321601	Government	Yes	Yes	Medium organizations	State/Provincial agencies	Vermont	State
g321614	Government	Yes	Yes	Medium organizations	State/Provincial agencies	Vermont	State
g321630	Government	Yes	Yes	Medium organizations	State/Provincial agencies	Vermont	State
g321646	Government	Yes	Yes	Medium organizations	State/Provincial agencies	Vermont	State
g321603	Government	Yes	Yes	Medium	State/Provincial	Vermont	State

Figure 5. Detail of Major Organizations Output

Conclusions and Recommendations

Network images are inherently communication tools. Their efficacy as an analytic tool is limited but, as a picture, they can quickly communicate a considerable amount of information. But, network maps lose their strength in communication as the network gets larger. Still, many users will want to know the information contained in that map. In Vermont's LCB, state policy relies heavily on partnerships and so stakeholders are actively interested in learning about and building their network for implementing water quality policy. This tool shows them that network and makes it accessible to everyone. It allows a user to explore, to find questions about the network when they may not know the question in advance and to pin-point parts of the network for display when the user does know which part of the network they want to learn about.

This application aids users with a range of different. Anyone who engages with network analysis can benefit from the approach to interactive network mapping that this project proposes. The immediate beneficiaries will be those working with the governance networks in the LCB, including the researchers who can better communicate their findings and the stakeholders who are embedded in the network. They will be better able to understand the context in which they work, allowing them to make better strategic decisions for their organizations, as well as to potentially diagnose and address situations where the network is impeding its own efforts to improve water quality in the LCB. The application does not support firm conclusions about which policies should be applied or about the nature of networked governance.[7] This application is designed to be a communications tool, not an analytic tool. The application's openness and accessibility demonstrates the benefits of using this kind of modeling approach in other contexts.

[7] Other published works based on these data have explored these questions.