# Capstone Project 1: Displaying Large Organizational Networks

### Project Description and Problem Statement

Despite their effectiveness as tools for studying complex systems, many large networks are difficult to describe, understand, and visualize. Even when the networks are not densely connected, their maps may be dense, with nodes, links, and labels all overlapping and obscuring each other. Network statistics help an analyst to understand such a large network, but can be difficult to use to communicate to non-specialists. And, indeed, communicating these networks to non-specialists is a key task of many network analysts. The primary goal of this project is to develop and demonstrate a tool that allows not just analysts but non-specialists to engage with a large organizational network and seek answers to their own questions about the network, its structures, and their organization’s own role in the network.

Network datasets are difficult to gather and gathering enough data to statistically study overall network structures is extremely difficult, as hundreds, if not thousands, of separate networks must be collected, each being individually difficult to collect. More effectively communicating networks on their structures can aid in this task. Having improved results that can use the statistical language that many even outside analytic professions have begun to recognize would also help by being more understandable.

Addressing the challenge in gathering complete network data from a different approach, network sampling and extrapolation from an observed sample of a network to the unobserved portions of the network is a continual challenge. The locations of network links are not independent observations, but dependent on the presence of other network links, making traditional statistical forecasting unreliable. Network analysts have developed a technique called Exponential Random Graph Models (ERGM) to account for this dependence while statistically examining the placement of links. But ERGMs are temperamental, subject to random factors in the calculations and model specification, and prone to failure when miss-specified. Current techniques for addressing this specification require the analyst to search through many different specifications seeking a possible best-outcome. Though in line with a sensitivity analysis, this approach is still inefficient at best, subject to getting stuck in a local optimization rather than a global optimization even when performed systematically, and often performed haphazardly. New machine learning techniques offer promise in improving on this method.

### Data

In 2014 and 2015, researcher with the Vermont Experimental Program to Stimulate Competitive Research (Vermont EPSCoR),[[1]](#footnote-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). Table 1 lists the four organizational groups into which we sorted our respondents and reports the rate at organizations responded, by group. An online survey was developed and deployed, with respondents recruited through personal outreach. Each respondent was presented with a list of all the organizations in the survey and asked 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. Several characteristics, or node attributes, for each organization are also recorded, including a measure of the organization’s budget and staff size, the 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). Each network and the attribute data are recorded on separate. Since these data were gathered with federal research dollars, a de-identified version is available for free distribution and will be posted to Github, along with a full codebook.[[2]](#footnote-2)

Table 1: Survey Response Rates

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Organizational Group | Number of Contacts | | Completed Responses | | Response Rate (%) | | Observation Rate (%)[[3]](#footnote-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 |

### Impact

A range of different interests can be addressed through this project. 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.

Many organizations and firms will find an interest in forecasting unobserved portions of a network from observed portions. The United States Department of Defense is interested in this technique for forecasting the structure of illicit networks, such as terrorist networks, where network targeting techniques are useful but the context prevents accessing information on the full structure of the network. Marketing firms may find this technique interesting for trying to use social media to market products. This project will build this visualization through organizational networks, but it will seek to build a demonstrate a tool that could be used in any situation where the data could be organized as a network.

### Deliverables

This project’s primary deliverable will be an application built in R’s *shiny* package. This package is specifically designed for making interactive online graphs and figures. R also offers a wealth of network analytic packages, including the *igraph* package and *statnet* suite of packages, as well as *ergm*, a package for doing ERGMs, included as part of *statnet*. This makes R an ideal venue to create a tool for dynamically manipulating network maps that can communicate both global and local network structures to non-specialists. Few examples of this approach currently exist, with one possible example being a network published by Resilient Infrastructure as Seas Rise (RISER) project, located at the University of California-Berkeley and the Center for Environmental Policy and Behavior at UC-Davis.[[4]](#footnote-4) This project will build an application with greater abilities to focus on small pieces of the network than that application allows.

Additional deliverables will include a code for a machine learning script to study link placement in the network as well as documentation of the results and, possibly, a comparison with current ERGM techniques.

1. I personally led this effort, leading the drafting of the survey instrument, outreach to survey respondents, and data preparation, cleaning, and analysis. [↑](#footnote-ref-1)
2. Data are found here: <https://github.com/wmirecon/Water_Quality_Governance_Networks> [↑](#footnote-ref-2)
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. [↑](#footnote-ref-3)
4. The map can be seen here: http://riser.berkeley.edu/sea-level-rise-management-network/ [↑](#footnote-ref-4)