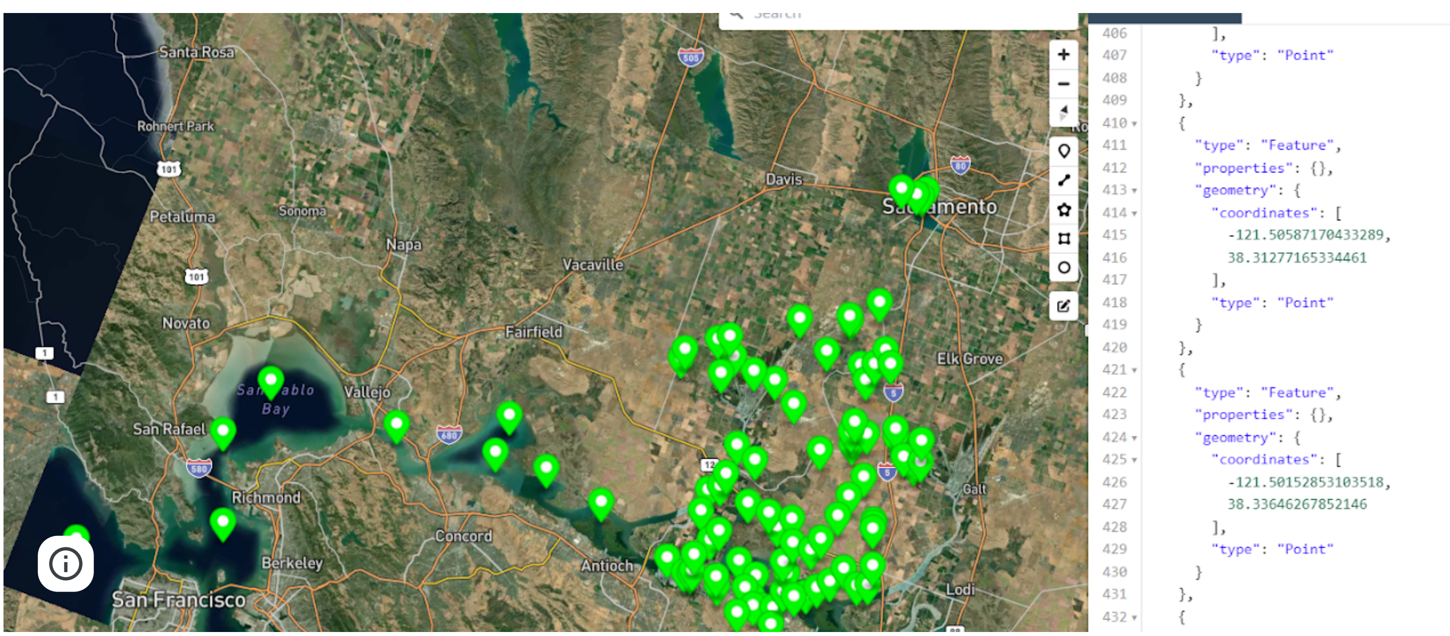
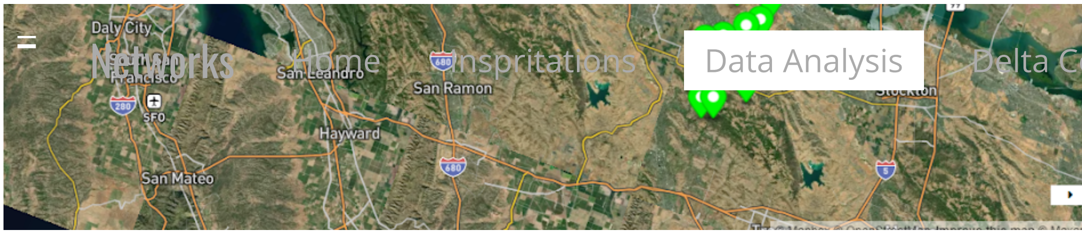
# DATA ANALYSIS

### Data Collection:

Data for the Sacramento-San Joaquin river delta was scares, at least for formulating it as a network. For this reason, we took to GeoJson to plot nodes, each representing a part of or the whole river, aqueduct, channel, slough among other water features. Then we manually took notes of the other nodes that they were connected to(those that water fed from one to the other). This allowed us to formulate a network of 102 nodes and 146 links. Simple statistics of our data include having an average degree of 2.863 and a max degree of 13, the San Joaquin River. Our system was one fully connected component with an average clustering coefficient of .220. As previously stated, we categorized each node into one of 15 types of water feature. This then allowed for us to progress into modeling, clustering, and visualization tasks to understand the complex network that we were working with. Of course we would love more data, so in the future, we can create and extract data from online sources to increase knowledge on each node. Ideally we would have a metadata table that tells which pieces of data we are in possession of, and then have sub tables that will contain either time series data, or more continuous/ever changing data.

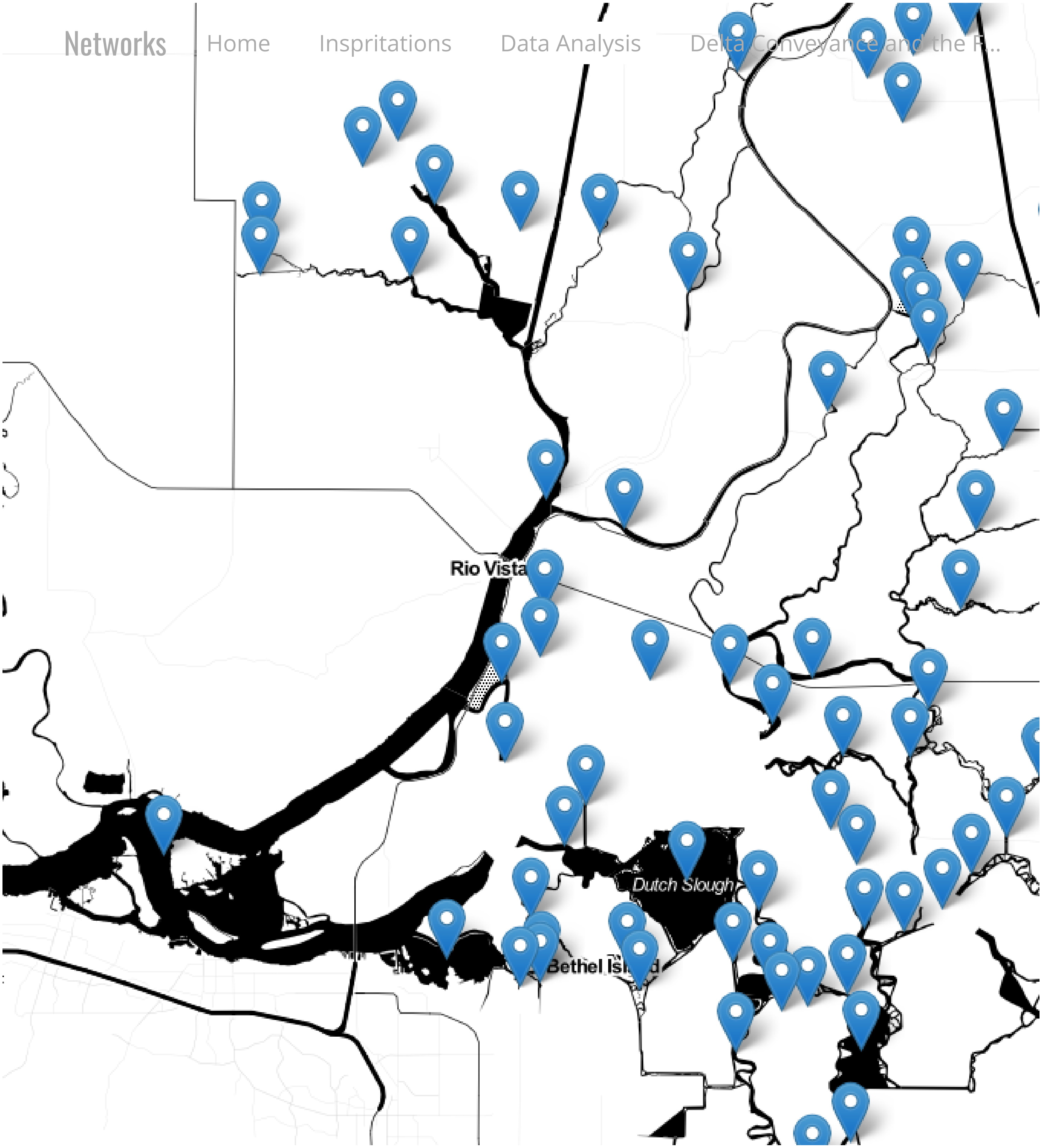


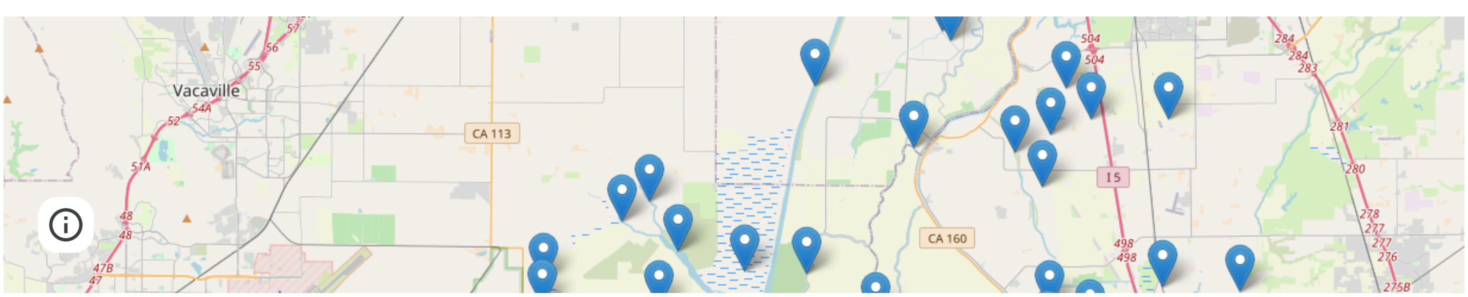


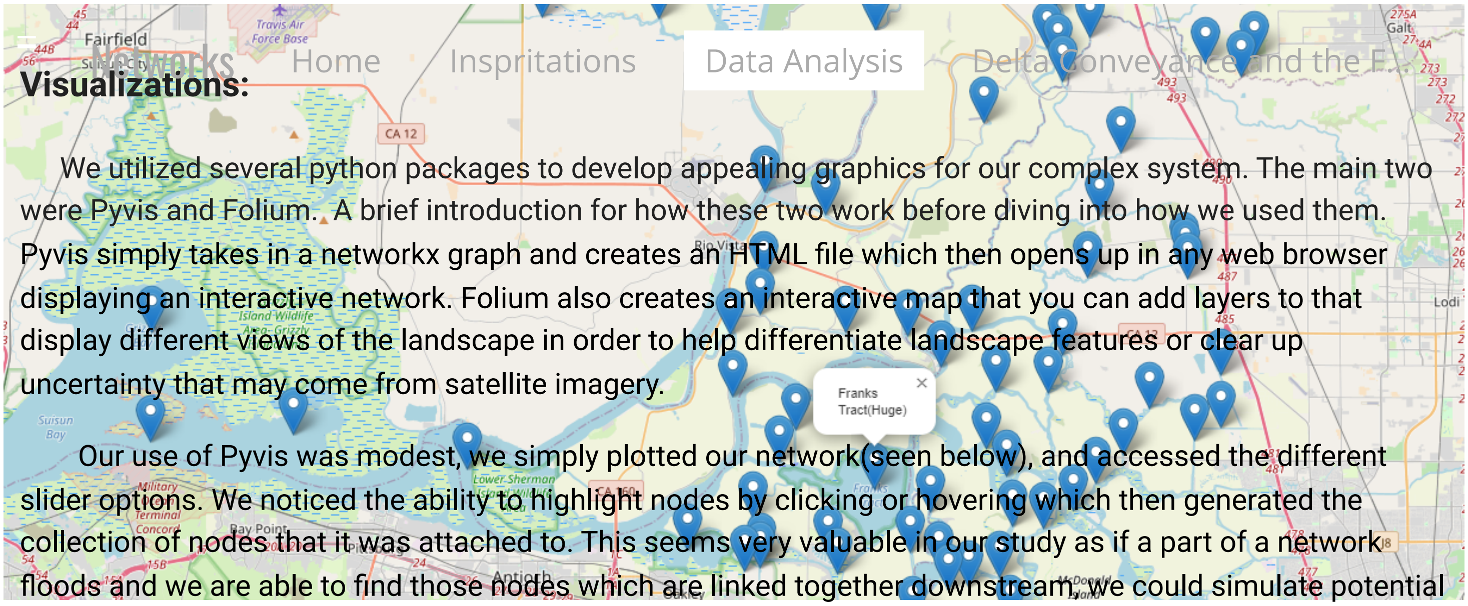


```
433       "type": "Feature",
434       "properties": {},
435       "geometry": {
436         "coordinates": [
437           -121.49223322062895,
438           38.357169587067574
439         ],
440         "type": "Point"
441       }
442     },
```

Networks | Home | Inspritations | Data Analysis | Delta Conveyance and the F...





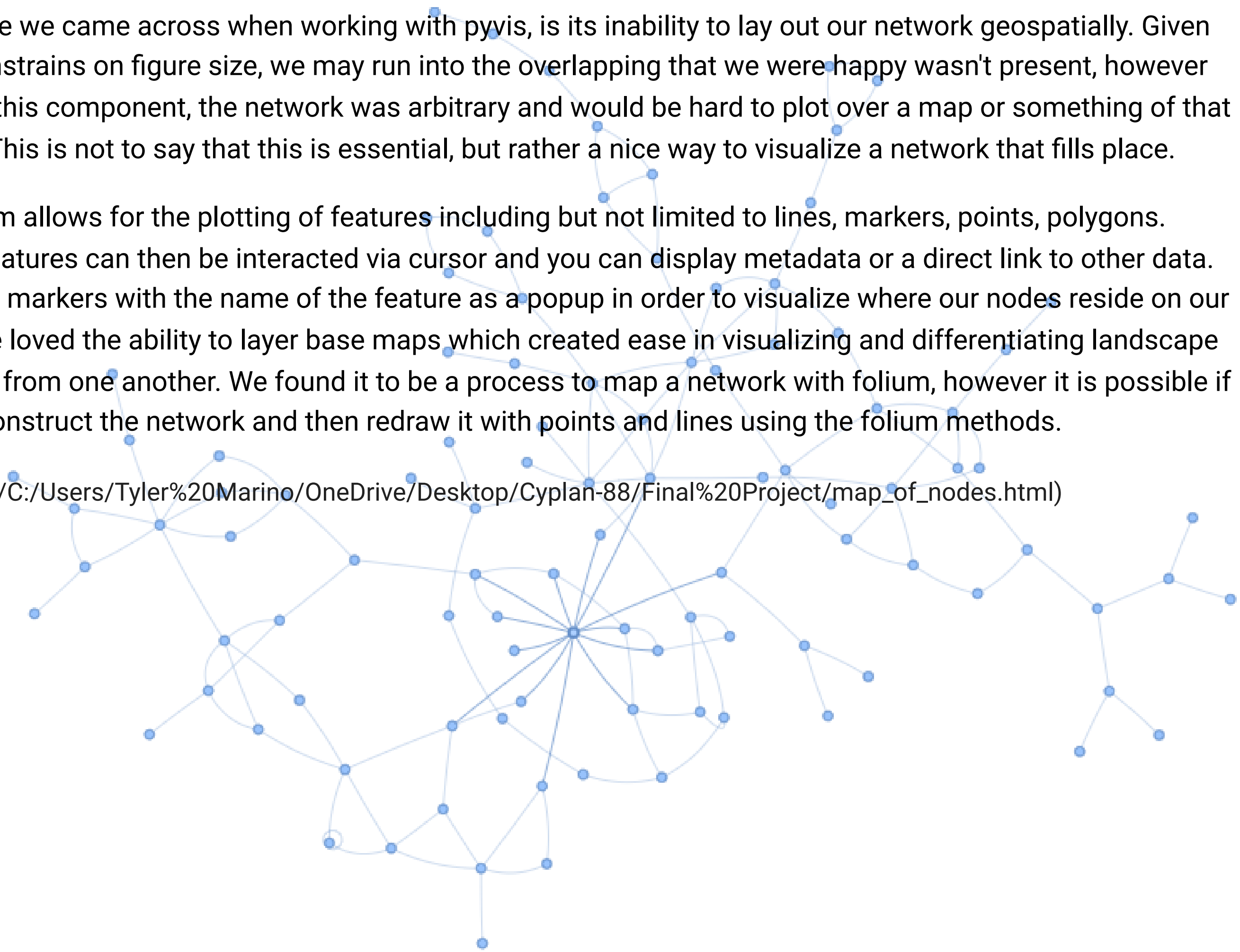

## Visualizations:

We utilized several python packages to develop appealing graphics for our complex system. The main two were Pyvis and Folium. A brief introduction for how these two work before diving into how we used them. Pyvis simply takes in a networkx graph and creates an HTML file which then opens up in any web browser displaying an interactive network. Folium also creates an interactive map that you can add layers to that display different views of the landscape in order to help differentiate landscape features or clear up uncertainty that may come from satellite imagery.

Our use of Pyvis was modest, we simply plotted our network(seen below), and accessed the different slider options. We noticed the ability to highlight nodes by clicking or hovering which then generated the collection of nodes that it was attached to. This seems very valuable in our study as if a part of a network floods and we are able to find those nodes which are linked together downstream, we could simulate potential flood patterns. We were happy with pyvis' ability to distribute the network nicely, minimizing overlaying of nodes and edges. This left a clean visualization that minimizes the chance for human perception error. The one issue we came across when working with pyvis, is its inability to lay out our network geospatially. Given tight constrains on figure size, we may run into the overlapping that we were happy wasn't present, however without this component, the network was arbitrary and would be hard to plot over a map or something of that nature. This is not to say that this is essential, but rather a nice way to visualize a network that fills place.

Folium allows for the plotting of features including but not limited to lines, markers, points, polygons. These features can then be interacted via cursor and you can display metadata or a direct link to other data. We used markers with the name of the feature as a popup in order to visualize where our nodes reside on our map. We loved the ability to layer base maps which created ease in visualizing and differentiating landscape features from one another. We found it to be a process to map a network with folium, however it is possible if you deconstruct the network and then redraw it with points and lines using the folium methods.

(file:///C:/Users/Tyler%20Marino/OneDrive/Desktop/Cyplan-88/Final%20Project/map_of_nodes.html)

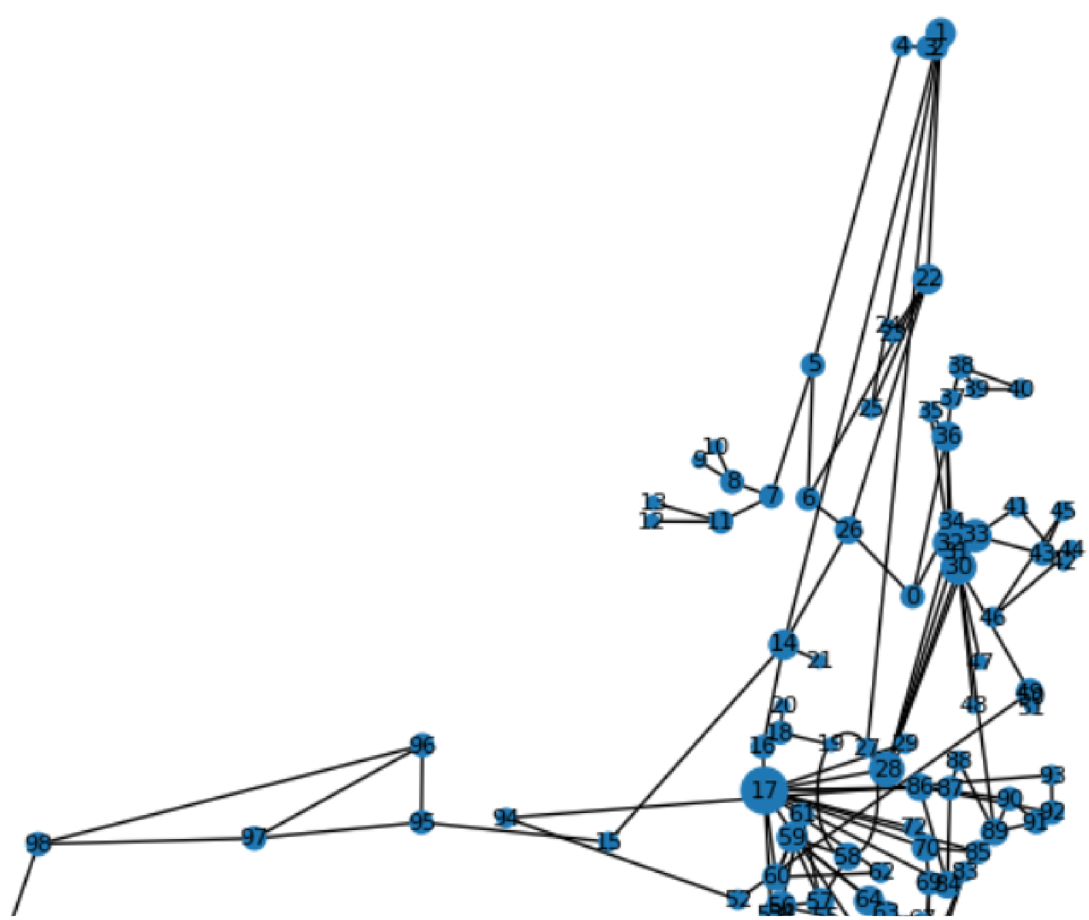


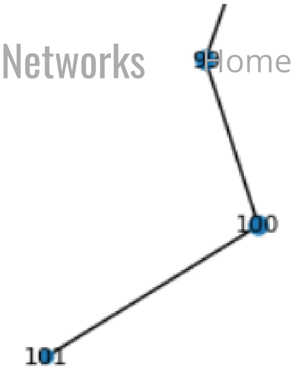
**Modeling:**

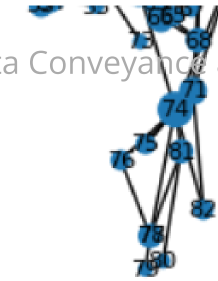
With a better understanding of how our network is set up and oriented, we began attempting to classify our system. Upon first glance, it seemed as if we had a Barabassi Albert model. We took this to be true by looking at the degree distribution(visualized below). There are clearly hubs in our network, those being nodes containing 6 or more edges attached. Upon further analysis and modeling, our network does not compare to a random graph, small world network, or Barabasi Albert model. There is an exceptionally high average shortest path when you compare these models, roughly 6.15. Of the models we have looked at in class, We conclude that this network fails to fall into one of these categories. This also has to do with the ever changing nature of theses complex systems. There is no steady state as tide, weather, and humans are constantly changing the environment in non-deterministic ways. Through analysis of connectivity and location modeling, the central nodes are the hubs. In a river system this makes sense. However trying to model this was difficult. Also, in future work we would love to sample arbitrary nodes at following the same spatial distribution as our network and then create a network and see how we can recreate the most similar. This is where we believe size of river, flow rate, and other factors will come in to make classification and clustering clearer.

---

River System Drawn with degree centrality for size





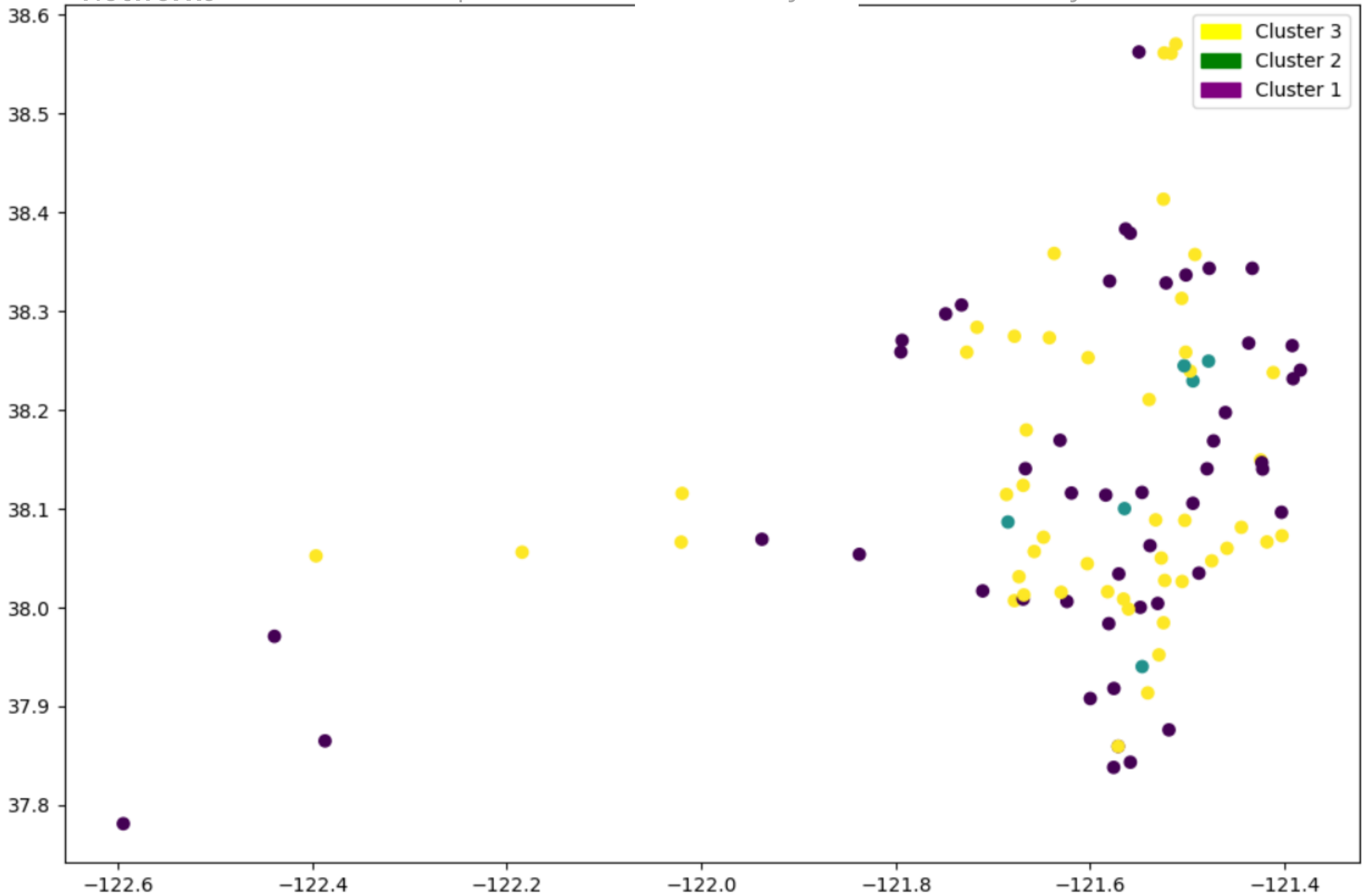



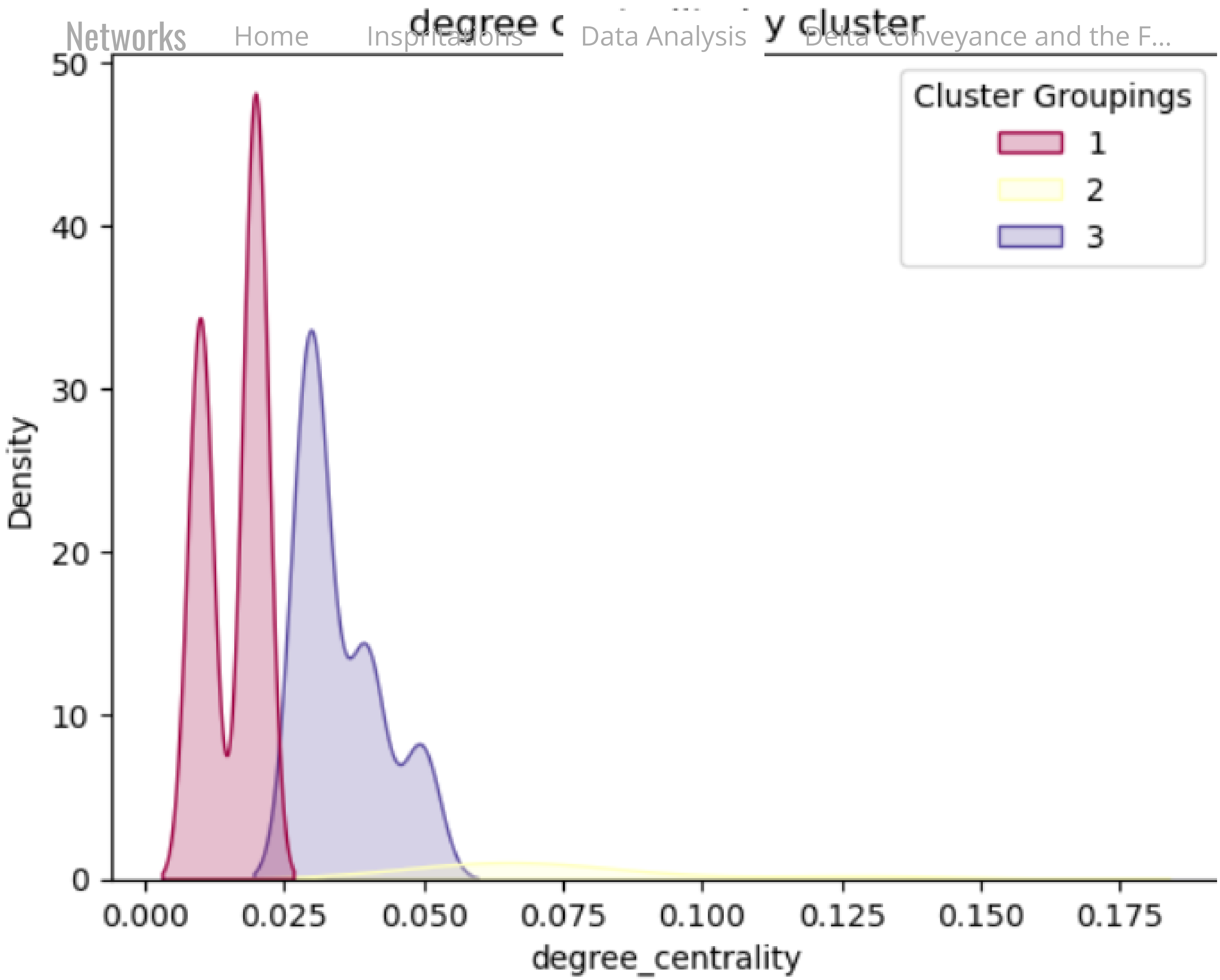

**Clustering:**

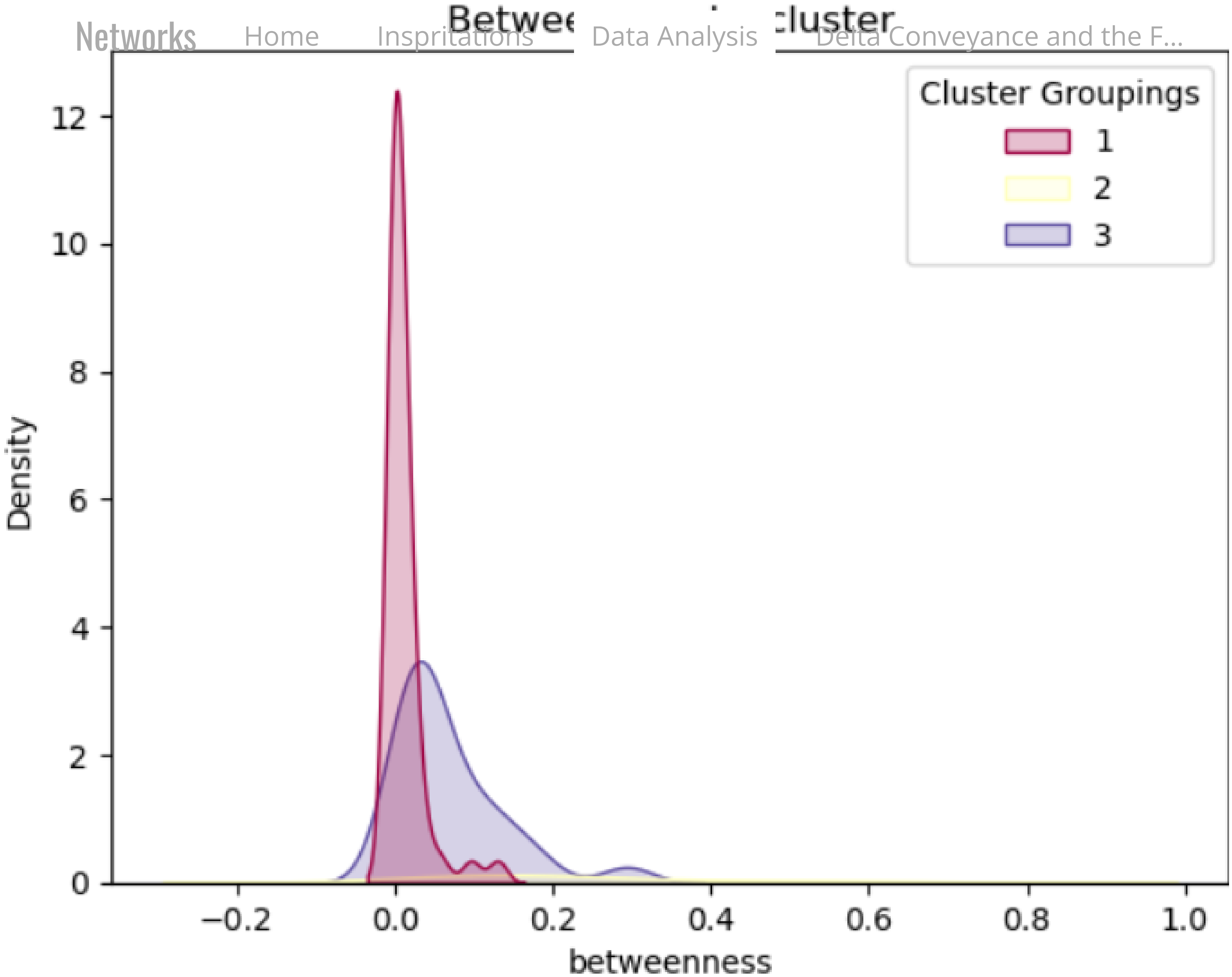
To further understand the type of network we are working with, we decided to perform clustering using the K Means technique. We calculated network statistics, betweenness, degree centrality, betweenness centrality, and degree for clustering. Then, after running through the clustering process and using the elbow method to determine the best number of clusters we believed it to be 4. However, we noticed that with 4 clusters the San Joaquin River was in a cluster to itself. This led us to believe that 3 clusters was truly the optimal choice.

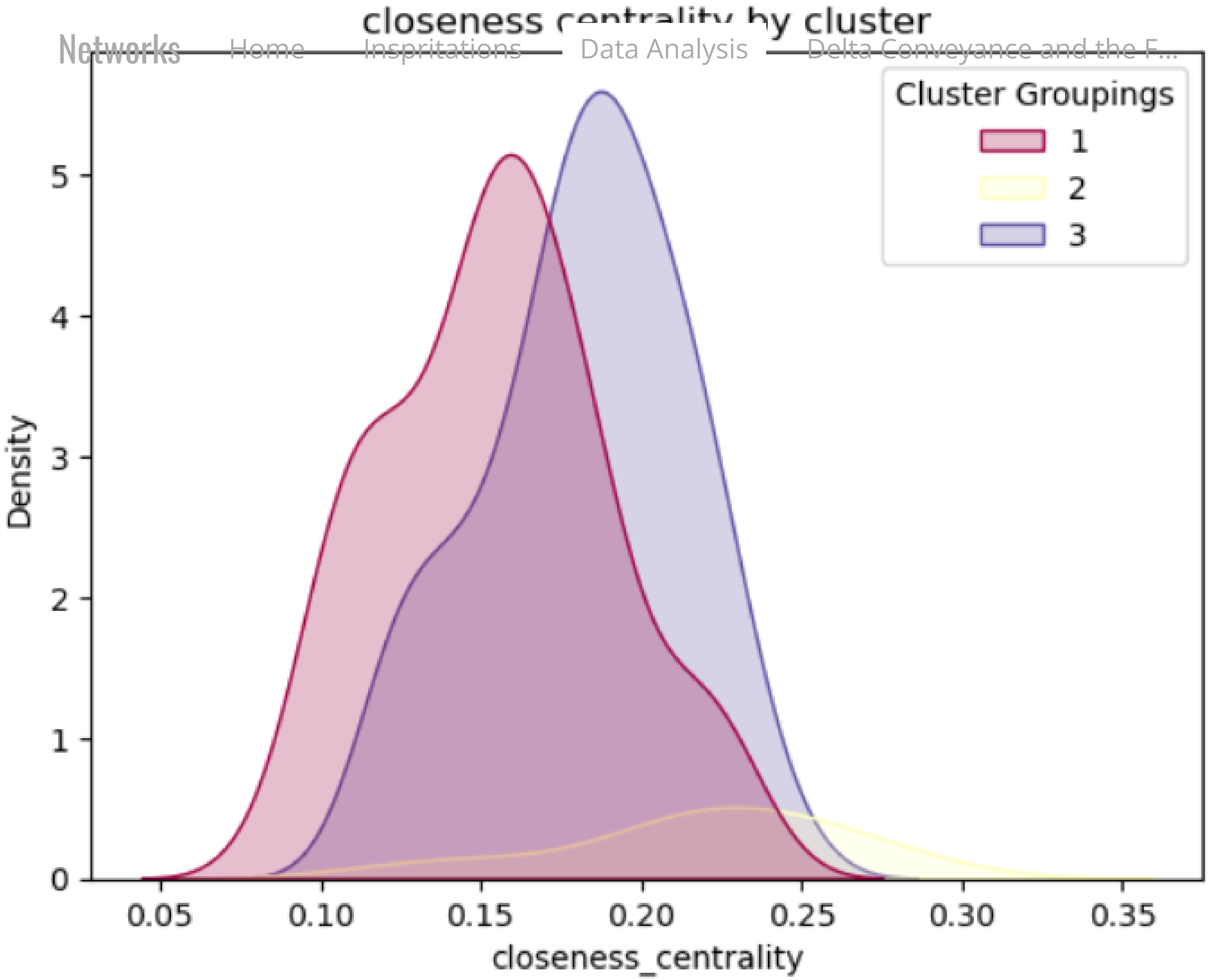
Our network is centrally focused and builds out. If there were weights on nodes, weights on influx of flows and other important factors to analyze we may see different results. However with the limited data, and evenly weighted components, we see that the network is controlled by those of the several key components/hubs.

Networks Home Inspritations Scatter Data Analysis s Delta Conveyance and the F...




degree c
y cluster
Cluster Groupings
1
2
3
Density
50
40
30
20
10
0
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
degree_centrality


Betwee
cluster
Cluster Groupings
1
2
3
Density
betweenness
12
10
8
6
4
2
0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0


closeness centrality by cluster
Cluster Groupings
1
2
3
Density
0
1
2
3
4
5
0.05
0.10
0.15
0.20
0.25
0.30
0.35
closeness_centrality