Network Analysis - Project Report

Jimi Hytönen Hanna Holtdirk Basil Mashal

28. helmikuuta 2020

1 Introduction

This project is part of network analysis course. The purpose of this project was to analyse real-world network by applying different algorithms to extract some interesting information about the network as well as get hands-on experience with network analysis. We chose to analyse California road networks because TBD - why?

In this report we will cover the technical aspects of gathering the data, tell how the work was divided, explain our network analysis and visualizations, and finally describe our conclusions. The project can be found in github (https://github.com/Jimmeeee/NA-project)

2 Data

We used dataset from https://www.cs.utah.edu/ lifeifei/SpatialDataset.html which was collected, cleaned and formated from multiple different sources into easy-to-use format. Network's nodes were in longitude-latitude coordinate form and network's edges contained information about start node, end node and the distance between them. In addition, the site provided information about California's points of interests such as hospitals, lakes and airports.

We used pandas for data manipulation and processing as we needed to get the data into a certain form for further analysis. We used networkx for building and visualisation of the network. This was really straightforward to do as the data was in an easy to use format. TBD-something else?

3 Methods

For the project we divided the original question into the following subtasks:

1. What is the general structure of the road network
2. Can we find where the (big) cities are from the road network and points of interest
3. Can we learn to make predictions about the placement of the roads and places of interest

For each of these subtasks we planned what analysis was needed to answer the question. After setting these subtasks we created a timeline for the project.

For the first question we analysed the California road network with simple methods to get a better idea what we are dealing with. One of which was calculating the connectivity of the network to determine if there were roads that led nowhere. We also calculated and visualized different centrality measures of the network such as degree centrality, betweenness centrality, eigenvector centrality and katz centrality.

After analysing the general structure of the network, we moved to the second question. For finding the big cities we used Girvan Newman community detection algorithm as it seemed to be the most suitable for the task. However, we didn't use the algorithm on the whole network because it was enormous, instead we used approximation based on previously calculated centrality measures to create a smaller graph and applied the algorithm on that.

For the third question ...

4 Results

4.1 General structure analysis results

4.1.1 Visualisation

We visualised the network using python library called networkx. In the following figure we can see the edge structure of California's road network on the left, and on the right we can see the same graph but with intensity of edge based on number of points of interest.

Fig. 1: California's road network

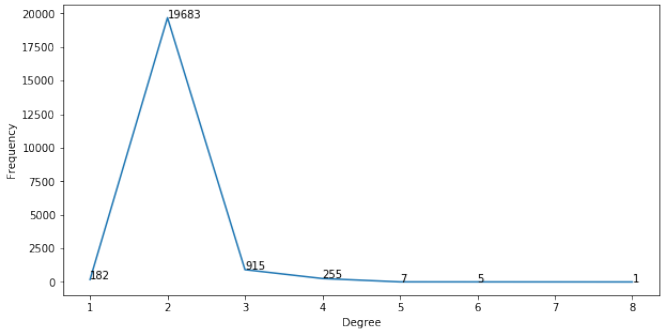
(a) Edges on the network

(b) Points of interest edge intensity

4.1.2 Degree distribution

In order to get more information about the basic structure of the network we checked the connectivity of the network and analysed degree distribution of the nodes. We found out that the California's roads are connected and there are no isolated sections. Also, from the figure below we can see the node degree distributions. It is intuitive that the majority of the nodes have degree of two as they are the middle nodes of the road. However, we can also see that there are lots of nodes that have degree higher than two, they are the intersections. Nodes that have degree of one are dead-ends.

Fig. 2: Degree distribution



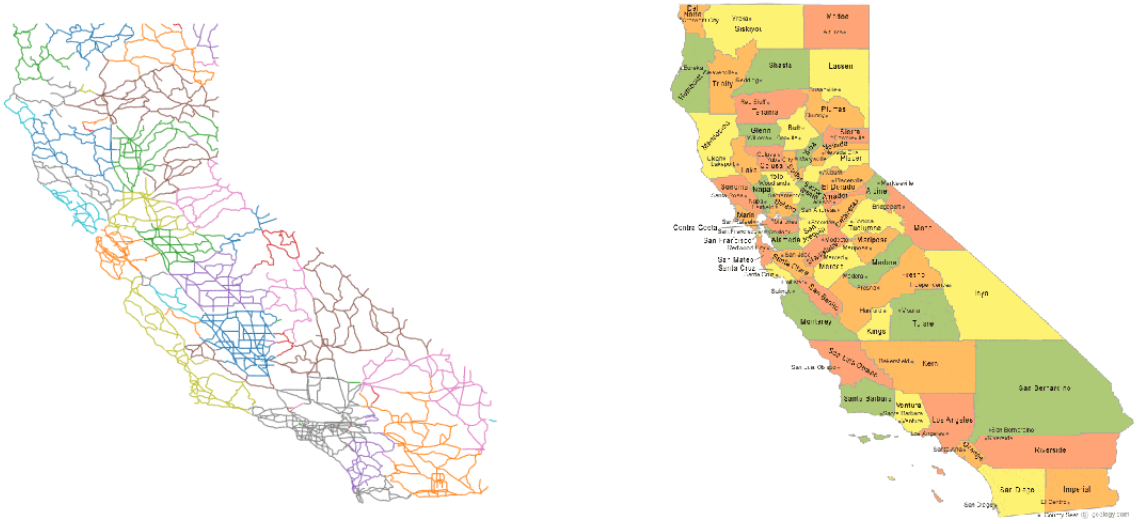
4.1.3 Centrality

We analysed the structure of the network with different centrality measures. These were: degree centrality, betweenness centrality, eigenvector centrality and katz centrality. In the figure below we can see the different centrality measures where higher color intensity and bigger node size means higer centrality value.

4.2 Communitites

For finding out the counties of California we used Girvan Newman community detection algorithm on networkx. Since the size of our network was inormous we had to optimize the algorithm with few methods. We based our edge removal to betweenness centrality as it sped up the algorithm noticeably. Also, we stopped the algorithm after it had found 58 communities as there are only 58 counties in California. In the figure below we can see how well we managed to find different counties based on the road network.

Girvan Newman community detection



4.3 Machine Learning

5 Conclusions

This does not need to be long

6 Contributors

1. Jimi Hytönen - Gathering data, building and visualising the network. Helping out with general structure and communities. Writing project report.
2. Hanna Holtdirk - TBD
3. Basil Mashal - TBD