Railroad Network Mapping and Analysis
Proposal Report
CSE 519 – Data Science Fundamentals

Team Members:
Abhinav Jain(111495982) abjain@cs.stonybrook.edu
Chaitanya Kalantri(111446728) ckalantri@cs.stonybrook.edu
Jitendra Savanur(111491676) jsavanur@cs.stonybrook.edu

Abstract

Railways are one of the most important modes of transportation around the world, with the topological properties of these railway networks attracting huge attention. We study the structural properties of the United States railway network, where railway stations are considered as nodes while edges are represented by trains directly linking two stations. The network displays small world properties and is assortative in nature. Based on betweenness and closeness centralities of the nodes, the shortest distance between two nodes in the network is processed.

Moreover, as recent attacks in trains and train stations show, the protection of such critical infrastructure plays a major role for public decision makers. Thereby, security installations in the railway network are a frequently discussed topic. Especially the need for an open system demands for technologies that do not influence or delay passenger flows. This also leads to the question of optimal placement of security installations such as smart camera systems or sensors. For answering this question we need to observe attacks on railway stations. The observation data was transferred into a quantitative network and analyzed using various measures.

Furthermore, as there is unequal distribution of railway passengers based on various periods of day. We can intimate passengers before hand, that there could be possible heavy crowd at the station or not. This time series data analysis could be effective to save time and take effective actions by the passengers.

To Sum, in this paper, we are trying to come up with the top 1000 places where Sensors could be placed, find the shortest distance between two nodes and analyze the time series data to understand the heaviness of the crowd at different time periods.

Data Acquisition

We have considered dataset from various sources, as per the requirements

1. **Trespassing Casualties Dataset:**
   We have taken dataset from ArcGIS website regarding 'Trespassing causalities' which consists of more than 2700 plus tuples. This web map, utilizing data from the Federal Railroad Administration (FRA), shows the total number of trespasser casualties that occurred on railroad property in the United States from June 2011 to January 2015. More than 500 fatalities a year and nearly as many injuries occur and most are preventable.
   Some of the CSV fields are Date, State, Type of incident, Railroad, etc.

2. **Railway Crossing Dataset:**
   The dataset from Highway-Rail Inventory Datain USA contains the information about the location and the average number of trains scheduled to cross during the day and night. The locations will be mapped to the coordinates from the railroad lines dataset to plot the entire network on map.
   Some of the CSV fields are Longitude, Latitude, CrossingIdSuffix, Railrd_CMPNY, State, CountyFIPS, TRAFICLN, Operating_RRCode, PassCnt, DayThru, NightThru, Gates.

3. **RailRoad Lines:**
   The dataset is taken from Transportation.gov website. The dataset is a comprehensive database of United States' railway system at 1:24000 to 1:100,000 scale. The dataset covers all 50 states plus the district of Colombia. The dataset exists in CSV, Shapefile and KML formats and consists of 235066 lines, each representing a unique railway line, that may have multiple tracks.
   Some of the CSV fields are FROMNODE, TONODE, CountyFIPS, StateFIPS, PASSENGER_type, ROwner1, ROwner2, ROwner3, Track_RGHTS, STATUS, Num_TRACKS, Shape_len, STRAC_net, len_MILES etc.

4. **Population wise 1000 US cities:**
   Dataset we have from GitHub is a JSON file which contains the population of a city and its coordinates as well as the growth of the city from year 2000 to 2013. JSON file consists of data of 1000 US cities. This data can be used to map the population of a given city with the help of its coordinates.
   Fields in JSON file are population, latitude, longitude, growth_from_2000_to_2013, etc.

5. **Census:**
   This Dataset is taken from census.gov. This dataset contains the state wise population of US along with the latitude and longitude coordinates from census 2010. This data can be used to map the cities which have less population density.
   Fields in this dataset are STATEFP, COUNTYFP, COUNAME, STNAME, POPULATION, LONGITUDE, LATITUDE.

Data Pre-Processing

1. Extract only important features/ fields which are relevant and have least correlation between them.
2. Normalize/ remove the null data from the fields to reduce the sparseness.
3. Give weightage to categorical data fields accordingly.
4. Apply page rank algorithm to identify where to place 1000 sensors for better traffic monitoring.
5. Reading .kml file with help of 'beautiful soup' library to extract data line by line and map it accordingly with primary key.
6. Tuples without the latitude and longitude coordinates will be deleted for better prediction.

Data Analysis

1. Create railway network graph
    - We will be using 'networkx' library for plotting railway network graph.
    - We will be plotting the graph by using the FROMNODE and TONODE fields from our dataset and railway lines as edges.
    - We will be analyzing the top 1000 nodes based on the desirability score.

2. Analyze most active railways crossing
    - Will be using the page rank algorithm over the population, and average number of trains passing from that crossing in daytime as well as night.
    - This observation will be useful while placing the Sensors at railway stations.

3. Calculate shortest path
    - We will implement the shortest path algorithm.
    - We will use haversine function, which takes latitude and longitude of source and destination and returns the shortest distance based on the mathematical principle behind it.

4. Place 1000 sensors
    - Based on the active or busiest railways crossing we will determine the best 1000 places to place sensors to monitor the traffic most efficiently.
    - We will be considering the trespassing causalities dataset while determining the best 1000 places.
    - To determine the 1000 best places, we will be applying the page rank algorithm.

5. Time-series Railway data analysis
    - We will be taking into considerations the previous year dataset regarding the passenger count and their travel time

- Using the above data, we will try to predict which route is busiest at what time of the day.
- We will also try to predict which route is busiest at what time of the year or quarter.

6. Classification of trains
    - We will be classifying the trains into clusters as industrial or passenger or freight trains using KNN algorithm.
    - We will also analyze which route taken by industrial or passenger or freight trains the most.

7. Creating a data map for busiest routes
    - We will be color coding the busiest route for better visualization.
    - We will be considering the railways crossing and the average trains passing through that crossing.

References

[1] https://gist.github.com/Miserlou/c5cd8364bf9b2420bb29
[2] http://osav-usdot.opendata.arcgis.com/datasets/2553aa5e457349efb600502050bf9c3c_0
[3] http://osav-usdot.opendata.arcgis.com/datasets/4ad2a211393b4bf5b9af664967f4a57a_0
[4] http://safetydata.fra.dot.gov/OfficeofSafety/publicsite/downloaddbf.aspx
[5] https://www.arcgis.com/home/item.html?id=93c8bc810064449d9376d68f996ead10
[6] http://fedmaps.maps.arcgis.com/home/item.html?id=722aca4326b94108b84296c7fd76a023
[7] https://www2.census.gov/geo/docs/reference/cenpop2010/county/CenPop2010_Mean_CO.txt
[8] https://www2.census.gov/geo/docs/reference/cenpop2010/county/CenPop2010_Mean_CO.txt
[9] http://ieeexplore.ieee.org/abstract/document/7492815/