EPFL, EE-558

A Network Tour of Data Science, Fall 2019

Project Proposal: Analysis of delays on the New Jersey railway network

Group 24: Rami Azouz, Linah Charif, Jasso Espadaler Clapés, Lynn Fayed

1 Motivation

Coming from a transportation background, evaluating delays and identifying their causes is a very crucial topic given the detrimental and cumulative cost it has on the economy. In analogy with the epidemiological spreading model, existence of hubs in every railway network results in a faster propagation of delays, mainly because the majority of network lines converges or diverges from the central stations. As a consequence, the first objective of this project is to assess whether value of delays at stations radially vary as we move further away from the center. The basic assumption behind this proposition is that trains can compensate delays on peripheral stations by increasing their speed. This is nevertheless impossible to achieve at central stations where high frequencies imply a higher level of restrictions. The previous analysis can be performed during morning and evening peaks through inversion of the graph directionality. In this context, and once the delay distribution is grasped, the eventual objective is to be able to predict its value given the time dependency of the dataset.

2 Choice of Dataset

An overwhelming amount of datasets in the field of transportation is available on open platforms. After a thorough research of the multiple existing datasets, we chose to base our project on the NJ Transit + Amtrak (NEC) Rail Performance dataset available on Kaggle. The New Jersey rail commuter rail network is the second largest of the United States in terms of passengers, and allows the commuting between the New Jersey state and New York. The dataset allows to assess the performance of the network, which is essential given its importance in the New York area. To do so, the dataset contains detailed information on the trains operating in the rail network. Indeed, more than 287,000 train trips for the NJ transit network are available at stop-level and with minute resolution for the schedules and delays. The data covers train trips from March 1, 2018 to April 30, 2019.

3 Methodology and Tools

The unique characteristic of this chosen railway network (Figure 1) is that its features (delays) are time dependent. Nevertheless, at the first stage, we will proceed with the assessment of the total average delays which originall vary as function of the time of the day, day of the week, and station location. Next, the dataset will be filtered based on the inward and outward trips because we expect the directional delays to be significantly different. Clustering is then performed to evaluate spatial variation of delays. The main incentive behind this approach is that the network under consideration follows a power law distribution (Figure 2). Hence, we expect the nodes closer to the center to have the highest level of delay. An initial computation of the average delay per station shows a significant smooth variation which we will attempt to correlate to the network configuration. Finally, one additional task that we would like to apply on our dataset is to predict the final delay of a train, given the delays at the previous stops. Regarding the tools we would like to use, our intuition would be to consider Recurrent Neural Network given the temporal aspect of our data. In fact, the accumulation of delays from the beginning to the end of a line will depend on the punctual delay at each station. However, the overall delay is not the sum of all the delays at each node: the delay between two stations can be compensated at the next stop by a change of speed for instance. Since the patterns are progressively learned in time in RNN, we thought that this tool could be adapted to the given task.

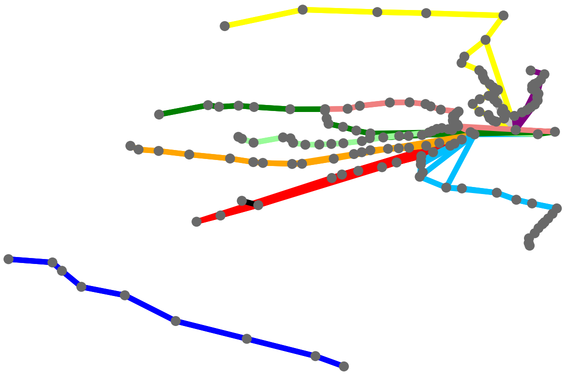

Figure 1: Graph of the NJ transit network

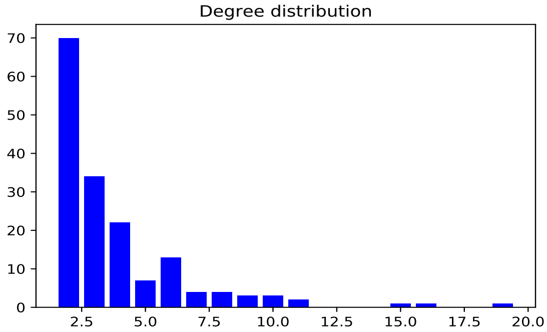

Figure 2: Degree distribution of the network

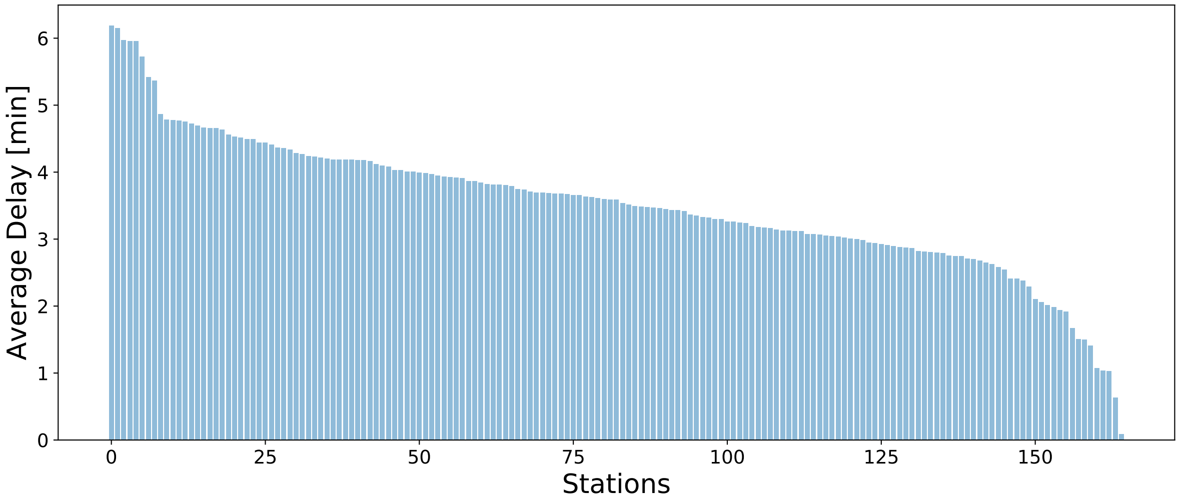

Figure 3: Average sorted delays on the different stations