Members: Shabnam Zareshahraki - Mattia Toffanin - Taha Shieenavaz

LFN Project: Compare the Subway Systems in Big Cities

Data Repository

We're getting ready to explore data from three cities, each presenting in its unique way. Our task is to compare the graphs, looking specifically at how stations are connected. However, before we start digging into the data, we first need to establish a common format and pick out the relevant information from the repositories.

1. Paris: The dataset, a text-formatted file, consists of **933 edges** and **376 vertices**. Each vertex is associated with a unique ID, and similarly, each edge is assigned an ID. The file provides detailed information about how edges connect to each other. This structure enables precise analysis of the relationships within the dataset.
2. São Paulo: The dataset is in CSV format, detailing stations and their neighbors. Extracting information about the number of edges requires some additional effort. Nonetheless, the graph comprises **79 vertices.**
3. London: The dataset is organized into separate files: one dedicated to nodes and another comprehensive file that encapsulates edges. This intricate network consists of **503 edges** interconnecting **369 nodes**.

Motivation

1. Do subway networks of major cities showcase similar patterns in the distribution of node scores?
2. We could focus on one city and pinpoint the best neighborhood for investment by considering several key graph measures.

Method

Problem:

Compute graph analytics and network features

- Node level
  - Centrality measures, to build a ranking of the best places to invest
    - Closeness centrality, which measures the proximity of a station to all other stations in the subway network
    - Betweenness centrality, which quantifies the extent to which a node lies on the shortest paths between other nodes in the graph
  - Clustering coefficient, which measures how connected the neighbors of a station are
- Graph level

- Diameter, which indicates the size of the subway system
- Average degree, which indicates the average number of edges per node
- Average clustering coefficient, which indicates if stations are well connected to each other
- Average shortest path length, which indicates the average number of stops from one random station to another

Algorithms:

We'll start with precise algorithms since the networks aren't huge. If things take too long, we'll switch to quicker approximation algorithms.

Implementation:

We'll be utilizing Python along with handy libraries like NetworkX (accessible at https://networkx.org) to handle graph data structures and implement algorithms for computing graph analytics and network features. For instance, the `average_clustering(G[, nodes, weight, ...])` function will be employed to calculate the average clustering coefficient of the network. It's worth noting that the NetworkX package offers both exact and approximate methods.

Our initial steps involve importing the datasets to create a manageable graph with NetworkX for each city. Following that, we'll dive into the computation of graph analytics and network features using the functionalities provided by NetworkX.

Machines:

- Mac Mini (M2, 8GB RAM)
- Macbook Pro (M1, 8GB RAM)

Experiments:

We will evaluate graph analytics and network features of subway systems of different cities. We will visualize and analyze the results to compare the distribution of node scores (node-level features) to understand the best places to invest in that city. We will analyze different graph-level feature scores to understand the differences between different subway systems.

References

NetworkX, https://networkx.org
Paris Subway Dataset, https://github.com/BTajini/Paris-Metro-Project
São Paulo Subway Dataset, https://www.kaggle.com/datasets/thiagodsd/sao-paulo-metro
London Subway Dataset, https://manliodedomenico.com/data.php