Group Members:
Shabnam Zareshahraki, shabnam.zareshahraki@studenti.unipd.it, 2091106
Mattia Toffanin, mattia.toffanin.1@studenti.unipd.it, 2096045
Taha Shieenavaz, taha.shieenavaz@studenti.unipd.it, 2092671
Link to download the code: https://github.com/tahashieenavaz/learning-from-networks

# LFN Project: Final Report

# Comparison and Analysis of Subway Networks in London, Paris, and São Paulo

## Abstract

This research compares the subway networks of Paris, London, and São Paulo, aiming to assess efficiency and guide strategic investments. Advanced network analysis techniques are employed to evaluate centrality measures, connectivity patterns, and network robustness. The findings identify key areas for targeted investments to enhance overall transit network performance, offering valuable insights for urban planners, policymakers, and investors.

## Introduction

In today's ever-evolving urban landscape, optimizing metropolitan transportation networks has become a crucial priority to ensure efficient and sustainable mobility. In this context, our research project focuses on three of the world's major cities: Paris, London, and São Paulo. The primary objective is to examine and compare the characteristics of the transportation networks in these metropolises, aiming to identify key points for strategic investments. The research questions we sought to address include:

- Efficiency and Connectivity: How do measures of efficiency and connectivity in the transportation networks compare across the cities of Paris, London, and São Paulo?
- Node Centrality: What are the central nodes in the transportation networks of each city? Which geographic areas could benefit most from targeted investments?
- Inter-City Comparison: To what extent do differences in the structures and performances of transportation networks influence the overall quality of urban life in these metropolises?

Through the application of network analysis techniques, we aim to provide answers to these questions, thus contributing to a deeper understanding of metropolitan transportation network dynamics and offering crucial insights for well-informed decision-making in infrastructure investments and urban planning.

## Methodology

### Dataset

To explore data, a common format that represents the relevant information, should be established. The following depicts the data repositories:

- 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. This structure enables precise analysis of the relationships within the dataset.
- 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**.
- 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**.

### 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
    - The clustering coefficient, which measures how connected the neighbors of a station are

- Graph level
    - Diameter, representing the longest shortest path between any two nodes
    - Average degree, which indicates the average number of edges per node
    - The average clustering coefficient, which indicates if stations are well-connected
    - Average shortest path length, which indicates the average number of stops from one random station to another random one

## Algorithms

We utilized exact algorithms to compute node and graph features, as the networks under consideration were relatively small.

## Implementation

### Dataset Import

Initially, all datasets were transformed from their original format into JSON format. Within the JSON files, there is an attribute “directed” indicating whether the graph is directed or not. There is also a “nodes” attribute containing a list of all graph nodes composed of “id”, “label” and “color”. Additionally, there is a “links” attribute containing the list of all edges composed of “source”, “target” and “color”. Subsequently, the Model class creates a NetworkX graph for each JSON file.

### Features Computation

We utilized 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 the graph feature, we used the following functions provided by NetworkX:

- `diameter(G[, nodes, weight, ...])` to compute the diameter of the graph;
- `average_clustering(G[, nodes, weight, ...])` to compute the average clustering coefficient for the graph;
- `average_shortest_path_length(G[, weight, method])` to compute the average shortest path length of the graph.

For node-feature, we used the following functions provided by NetworkX:

- `degree_centrality(G)` to compute the degree centrality for nodes;
- `closeness_centrality(G[, u, distance, ...])` to compute closeness centrality for nodes;
- `betweenness_centrality(G[, k, normalized, ...])` to compute the shortest-path betweenness centrality for nodes;
- `clustering(G[, nodes, weight])` to compute the clustering coefficient for nodes.

Once the node features are calculated using networkX functions, the first 10 nodes with the highest node feature values are printed.

### Final Score Computation

To identify the stations for investment, a score was calculated for each node by combining the values of various node features previously computed.

Initially, the process involved normalizing the diverse node features to ensure their values fell within a consistent range, preventing any single feature from disproportionately overshadowing others. Following this normalization step, the standardized features were amalgamated using the linear model $s(n) = a * degree(n) + b * closeness(n) + c * betweenness(n) + d * clustering(n)$, where $a, b, c, d$ are the coefficients chosen concerning the feature node importance.

We selected network efficiency as the primary focus for your investment strategy, thus, both betweenness centrality and closeness centrality are particularly relevant for optimizing network efficiency.

- Betweenness Centrality: Given its role in identifying crucial connectors and facilitating efficient traffic flow, we assign a relatively high weight to betweenness centrality. Therefore, $c$ = 2.
- Closeness Centrality: As closeness centrality measures the proximity of a station to all other stations, enhancing overall network accessibility, we also assign a significant weight to closeness centrality, so, $b$ = 1.5
- Degree Centrality and Clustering Coefficient: While these factors are still important, we assign lower weights to them in this context, as your primary focus is on network efficiency. As a result, $a$ = 1 and $d$ = 1.

### Machines

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

## Experimental Evaluation and Results

### Investment Factors

The factors of degree centrality, betweenness centrality, closeness centrality, and clustering coefficients play crucial roles in guiding investments in subway systems.

1. Degree Centrality:
   - Significance: Degree centrality measures the number of edges connected to a node, indicating how well-connected and central a station is within the entire subway network.
   - Investment Implications: Stations with a high degree of centrality are pivotal for investment as they are well-connected to multiple other stations. Investing in neighborhoods around these high-degree stations enhances accessibility to a larger portion of the subway network. These stations often attract more passenger traffic, making them strategic for businesses and urban development.
2. Betweenness Centrality:
   - Significance: Betweenness centrality quantifies the extent to which a node lies on the shortest paths between other nodes in the graph. It identifies stations that play a crucial role in facilitating traffic flow and communication between different parts of the network.
   - Investment Implications: Stations with high betweenness centrality are strategic connectors. Investing in these stations ensures efficient traffic flow, reduces travel times for commuters, and contributes to the resilience and redundancy of the transit system. Economic activities around these stations may flourish due to increased footfall.
3. Closeness Centrality:
   - Significance: Closeness centrality measures the proximity of a station to all other stations in the subway network, indicating how quickly passengers can reach various parts of the network from that station.
   - Investment Implications: Stations with high closeness centrality are critical for overall network accessibility and efficiency. Investing in these stations enhances the ease of commuting, making them attractive locations for businesses and residential development. Efficient connections provided by these stations contribute to the convenience of the transit system.
4. Clustering Coefficient:
   - Significance: The clustering coefficient measures how interconnected the neighbors of a particular node are, providing insights into the local cohesion and connectivity of that station within its immediate vicinity.
   - Investment Implications: Stations with high clustering coefficients are important for local connectivity and community development. Investing in these areas can lead to the establishment of community-centric facilities and services. High clustering coefficients may also indicate areas with vibrant economic activity, making them attractive for various types of investments.

## Conclusions

### São Paulo

#### Graph Analysis

The Sao Paulo network is strongly connected, indicating that there is a path between every pair of nodes in the graph.

Average Degree: 2.0253

#### Graph Features

- Diameter: 33
- Average Clustering Coefficient: 0.0000, indicating a low level of local clustering.
- Average Shortest Path Length: 12.4216.

#### Node Features

Each node has associated centrality and clustering coefficient values, providing insights into their importance and local clustering.

1. Top 20 Nodes with Highest Degree Centrality:
   Nodes with the highest degree of centrality include Republica, Santa Cruz, Se, Ana Rosa, Chacara Klabin, and Luz.
2. Top 20 Nodes with Highest Closeness Centrality:
   Nodes with the highest closeness centrality include Paraiso, Vergueiro, Sao Joaquim, Japao Liberdade, Ana Rosa, and Se.
3. Top 20 Nodes with Highest Betweenness Centrality:
   Nodes with the highest betweenness centrality include Se, Paraiso, Vergueiro, Sao Joaquim, Japao Liberdade, and Ana Rosa.
4. Top 20 Nodes with Highest Clustering Coefficient:
   All nodes exhibit a clustering coefficient of 0.0000, suggesting a lack of local clustering.

### Overall

The Sao Paulo network displays strong overall connectivity, with certain key nodes having high centrality and betweenness values. The network's diameter indicates the presence of long-distance connections, and the low average clustering coefficient suggests a limited level of local clustering. These findings may have implications for network robustness, efficiency, and potential areas for improvement or targeted

## Paris

### Graph Analysis

The Paris network is strongly connected, indicating that there is a path between every pair of nodes in the graph.

This connected component exhibits the following characteristics:

Average Degree: 2.3069

### Graph Features

- Diameter: 42
- Average Clustering Coefficient: 0.0164,  indicating a low level of local clustering.
- Average Shortest Path Length: 14.7068.

### Node Features

Each node has associated centrality and clustering coefficient values, providing insights into their importance and local clustering.

1. Top 20 Nodes with Highest Degree Centrality:
   The top nodes based on degree centrality (e.g., Chatelet, Gare du Nord) are crucial hubs in the Paris transportation network. Nodes with high closeness centrality (e.g., Chatelet Les Halles, St Michel Notre Dame) are geographically central, offering efficient connections. Nodes with high betweenness centrality (e.g., Chatelet Les Halles, Gare du Nord) are critical for maintaining connectivity.
2. Top 20 Nodes with Highest Closeness Centrality:
   The top nodes based on closeness centrality play pivotal roles as central hubs for efficient connectivity. Nodes such as Chatelet Les Halles, St Michel Notre Dame, Gare du Nord, and Gare de Lyon demonstrate short average distances to other nodes, indicating their significance in facilitating quick and accessible connections throughout the network.
3. Top 20 Nodes with Highest Betweenness Centrality:
   Nodes with the highest betweenness centrality hold crucial positions in facilitating communication and traffic flow between various parts of the network. Chatelet Les Halles, Gare du Nord, and Gare de Lyon emerge as key transit points with the highest betweenness centrality, indicating their pivotal role in connecting different routes and enabling efficient travel. St Michel Notre Dame and St Denis also exhibit significant betweenness centrality, emphasizing their importance as transit links. These nodes likely serve as major transfer points, contributing significantly to the overall network's resilience and accessibility.
4. Top 20 Nodes with Highest Clustering Coefficient:
   Paris has nodes with a clustering coefficient of 1.0000, indicating that certain areas (e.g., Falguiere, Reullly Diderot) form highly connected clusters. This suggests localized efficiency and strong interconnectivity within specific regions.

### Overall

The network analysis of the Paris transportation system reveals a complex and interconnected structure. With a diameter of 42 and an average shortest path length of 14.7068, the network exhibits a considerable degree of connectivity. The top 20 nodes with the highest degree of centrality, such as Chatelet, Gare du Nord, and Montparnasse Bienvenue, represent major transportation hubs. Similarly, nodes like Chatelet Les Halles, St Michel Notre Dame, and Gare de Lyon, with high closeness centrality, underscore their efficiency in minimizing travel times and enhancing overall accessibility. Nodes with the highest betweenness centrality, including Chatelet Les Halles, Gare du Nord, and Gare de Lyon, play crucial roles in maintaining efficient traffic flow and connectivity throughout the network. The high clustering coefficient in nodes like Falguiere and Reullly Diderot suggests localized connectivity within these areas. Overall, the Paris network demonstrates a well-organized and strategically positioned transportation system, with key nodes facilitating effective mobility and transit within the city.

## London

### Graph Analysis

This connected component exhibits the following characteristics:

Average Degree: 2.3113

### Graph Features

- Diameter: 38
- Average Clustering Coefficient: 0.0328,  indicating a low level of local clustering.
- Average Shortest Path Length: 14.0985.

### Node Features

Each node has associated centrality and clustering coefficient values, providing insights into their importance and local clustering.

1. Top 20 Nodes with Highest Degree Centrality:

   These nodes represent crucial and highly connected locations within the system. Notable among them are Bank, Monument, and Barking, indicating these stations have a large number of direct connections with other stations. These nodes are likely major interchange points or central hubs in the network, facilitating efficient transfers and providing accessibility to various destinations. The prominence of these stations in degree centrality underscores their pivotal role in shaping the overall structure and connectivity of London's transit system, highlighting key areas of concentration and importance for commuters and transportation planning.
2. Top 20 Nodes with Highest Closeness Centrality:

   These nodes are critical in terms of overall network accessibility and efficiency. Holborn, Epping, and Bounds Green stand out, indicating that these stations are not only well-connected to nearby nodes but also efficiently linked to the broader network. High closeness centrality suggests that these nodes serve as effective central points for minimizing travel time, playing a pivotal role in enhancing the ease of commuting across the network. Commuters at these stations can generally reach other parts of the network more swiftly, underscoring their strategic importance in ensuring efficient and well-connected transportation within the London transit system.
3. Top 20 Nodes with Highest Betweenness Centrality:

   These nodes play crucial roles in facilitating the flow of traffic and information across the entire system. Holborn, Ruislip Gardens, and Barking emerge as key transit hubs with high betweenness centrality, signifying their significance in connecting different routes and serving as essential transfer points. These stations are likely to witness a substantial volume of transit traffic passing through them, making them vital in maintaining the network's overall connectivity. Nodes with high betweenness centrality, such as Bank and Epping, act as critical bridges between various parts of the network, ensuring efficient communication and transit options for commuters traveling through the London transit system.
4. Top 20 Nodes with Highest Clustering Coefficient:

   These nodes exhibit a notable tendency to form tightly-knit clusters or communities within the overall network structure. Nodes like Cyprus, Kennington, and Kensal Green have a clustering coefficient of 1.0000, indicating that these stations, along with their connected neighbors, form cohesive groups where many nodes are interconnected. These clusters contribute to the network's resilience and local efficiency, as they represent areas of increased connectivity and collaboration among nearby stations. This clustering pattern suggests that certain stations in the

London transit system have a higher likelihood of passengers transferring between nearby stops within these closely interconnected groups, possibly due to geographical proximity, shared transportation routes, or specific urban characteristics.

### Overall

The analysis of the London transportation network reveals a complex and well-connected system with diverse centrality patterns across its nodes. The network's relatively large diameter of 38 and an average shortest path length of 14.0985 indicate that, on average, stations are within reasonable proximity of each other. The average clustering coefficient of 0.0328 suggests a moderate level of local clustering, emphasizing the presence of cohesive groups of stations. Examining the centrality measures, Holborn stands out with the highest values in both closeness and betweenness centrality, signifying its critical role in connecting different parts of the network and facilitating efficient passenger movement. Bank, Ruislip Gardens, and Barking also exhibit significant centrality, playing key roles in maintaining network cohesion. The high clustering coefficient of certain nodes, such as Cyprus and Kennington, reflects the existence of tightly-knit station clusters. Overall, the London transit system displays a robust and interconnected structure, with specific nodes playing pivotal roles in facilitating efficient and resilient passenger flows.

## Comparison

The analysis of the transportation networks in Sao Paulo, Paris, and London highlights both commonalities and distinctions in their structural and centrality characteristics. In terms of graph features, Sao Paulo has the largest diameter and average shortest path length, indicating a more dispersed network, while Paris and London have smaller values, suggesting more compact structures. All three networks exhibit low average clustering coefficients, indicating a limited level of local connectivity.

Regarding centrality measures, the top nodes in degree centrality for all cities are major transportation hubs, such as Chatelet and Gare du Nord in Paris, Chatelet Les Halles in Sao Paulo, and Holborn in London. These stations play pivotal roles in facilitating connectivity within their respective networks. In terms of closeness centrality, commonalities include key stations like Chatelet Les Halles in Paris, Republique in São Paulo, and Holborn in London, emphasizing their significance in ensuring efficient passenger travel. Nodes with high betweenness centrality, representing critical connectors between different parts of the network, are observed in all cities, such as Chatelet Les Halles in Paris, Republique in São Paulo, and Holborn in London.

Despite these similarities, differences emerge in the specific nodes highlighted by each city. São Paulo emphasizes Republique and Santa Cecilia, Paris prioritizes Gare du Nord and St Michel Notre Dame, and London showcases Holborn and Ruislip Gardens. These distinctions reflect the unique characteristics and operational dynamics of each city's transportation system.

In summary, while São Paulo, Paris, and London share certain commonalities in the centrality of major transportation hubs, the specific nodes and structural features of their networks differ, reflecting the distinct urban layouts, population distributions, and historical development patterns of each city.

# 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

# Members contribution

Shabnam Zareshahraki

- Writing the final report
- Writing the project proposal (initial and mid-term)
- Refactoring the code
- Implementing the analytics functions
- Final analysis of the networks and conclusions

Mattia Toffanin

- Project ideation
- Writing project proposal (initial and mid-term)
- Computation of graph and node features through NetworkX functions
- Creation of a combination score to find the best places to invest
- Normalization of node features
- Fixed not strongly connected graph in the import module
- Writing final report

Taha Shieenavaz

- Project ideation
- Writing the project proposal (initial and mid-term)
- Importing the datasets
- Implementing models and helper functions related to the databases
- Creating and managing the code repository and project structure