UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

A.Y. 2023/2024

Master's Degree

Computer Science

# Metro Transport Network

Data Analytics Project - February 2025

## Group Partecipants:

Piacente Cristian - 866020

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

# Introduction

# Context

Île-de-France Mobilités manages a complex public transport network, including:

- **Metro**: 16 lines, serving Paris
- RER**:** 5 train lines
- Trams and Buses

Our analysis will focus on the Metro.

# Goals

We will represent the Paris Metro Network as an **undirected weighted graph**, with the nodes being the stations and the edges will be weighted by the Travel Time (in minutes).

It will let us achieve the following goals:

- Construct **Interactive Graph Visualizations**.
- Compute **Network Metrics.**
- Simulate **Failure Scenarios**.
- Perform a **Load Analysis**.

# Dataset

The data is in the **GTFS** (General Transit Feed Specification) standardized format, from the Mobility Database. In this project we use the version released on **February 6th, 2025**.

Preprocessing steps include:

- **Filter Metro Data**: Extract the Metro data.
- **Remove Duplicate Stations**: Merge stations with the same name (in the raw data they're duplicated because of multiple lines).

After Preprocessing, we have
**321** Metro stops and **16** Metro lines.

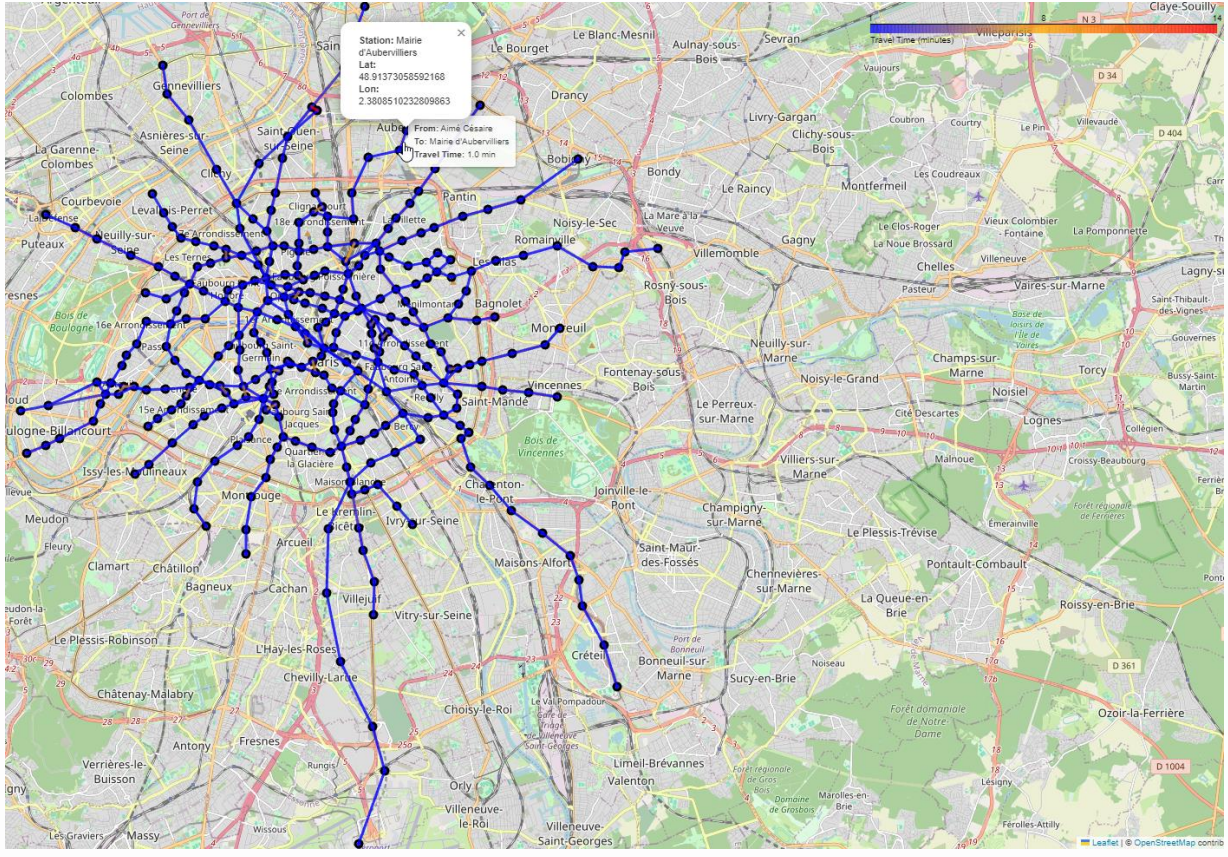
| Dataset | Number of Entries |
|---|---|
| Metro Routes | 16 |
| Metro Trips | 47,258 |
| Metro Stop Times | 1,176,628 |
| Metro Stops (after deduplication) | 321 |
| Metro Transfers | 498 |

**Route**: Metro Line.
**Trip**: Metro journey from start to end.
**Stop Times**: Scheduled Arrival & Departure for each Metro stop.
**Transfer**: Connection between different Metro Lines.

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

# Network Visualization & Properties

# Interactive Map Visualization




Before computing Network Properties, we construct and visualize an **interactive** graph on OpenStreetMap:

- A **node** is a Metro stop with the **name** and **coordinates**.
- An **edge** is weighted by the Travel Time in minutes.

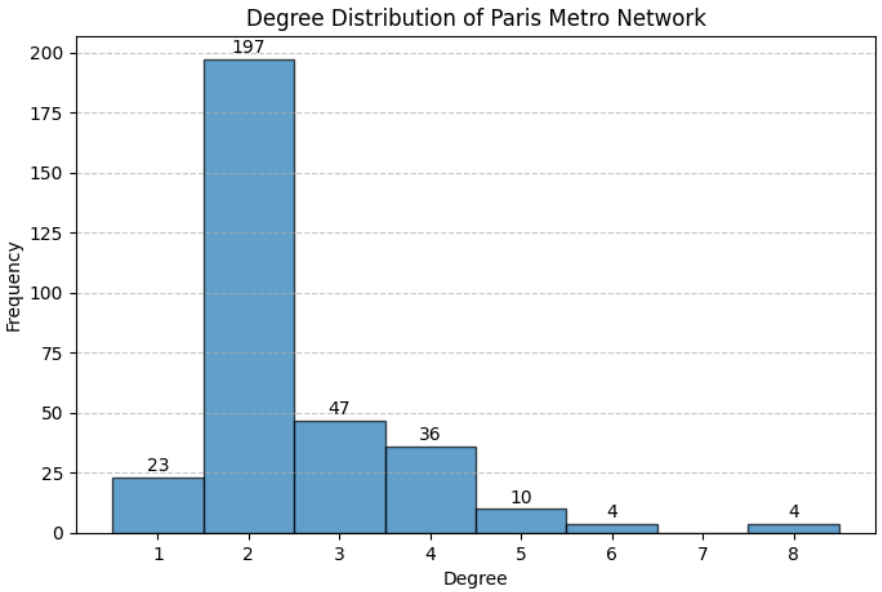
# Overview

We will now compute the following:

- **Degree Distribution**.
- **Diameter**.
- **Density**.
- **Efficiency**.
- **Average Clustering Coefficient**.
- **Average Shortest Path Length**.
- **Assortativity**.

# Degree Distribution

The Degree indicates how many connections each station has within the network.




The Average Degree is defined as

$$\langle k \rangle = \frac{2|E|}{|V|}$$

For the Paris Metro, we have $\langle k \rangle = 2.52$.

Most stops have only a few connections, but we also observe the presence of **high-degree hubs**.

# Diameter

The Diameter is defined as the longest shortest path between any two stations.

In our case, we have

$$D = 36$$

«the maximum number of stops required to travel between any two stations is **36**».

This means the network is geographically extensive and we will see high-centrality stations are important because of this.

# Density

The Density is the ratio of existing edges to the total possible edges in a fully connected graph, which is $|V|(|V|-1)\ /\ 2$ for an undirected graph.

$$\rho = \frac{2|E|}{|V|(|V|-1)}$$

The Paris Metro network presents

$$\rho = 0.0079$$

which is what we expect, since in a Metro network we don't have a fully connected graph.

# Efficiency

The Efficiency (Latora and Marchiori) is used to evaluate the performance after the removal of nodes/links, defined as

$$E = \frac{1}{n(n-1)} \sum_{i,j \in V, i \neq j} \frac{1}{d_{ij}}$$

where $d_{ij}$ is the distance between nodes i and j.

We have

$$E = 0.1124$$

This suggests a vulnerability of the network: geographic coverage is prioritized rather than optimal connectivity.

# Average Clustering Coefficient

The Average Clustering Coefficient is the probability that two neighbors of a randomly selected node link to each other, defined as

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i$$

with $C_i$ being the local clustering coefficient for a node i, which is $C_i = \frac{2E_i}{k_i(k_i - 1)}$

We have

$$\langle C \rangle = 0.0088$$

which confirms the network is not redundant.

# Average Shortest Path Length

The Average Shortest Path Length determines the typical journey time in minutes.

We have

$$\langle L \rangle = 12.25 \text{ minutes}$$

«a typical passenger must travel approximately **12 minutes** to reach their destination».

# Assortativity

Assortativity indicates whether hubs (high-degree stations) tend to connect to other hubs.

We have

$$r = 0.0588$$

which is close to zero: the network is almost **neutral** and hubs do not necessarily interconnect.

# Summary

| Metric | Value |
|---|---|
| Average Degree | 2.52 |
| Network Diameter | 36 |
| Network Density | 0.0079 |
| Network Efficiency | 0.1124 |
| Clustering Coefficient | 0.0088 |
| Average Shortest Path Length | 12.25 mins |
| Assortativity | 0.0588 |

Results suggest that hubs play a critical role and their disruptions could have a significant impact on connectivity, which we will explore.

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

# Vulnerability Analysis

# Overview

To assess the vulnerability, we will analyze **centrality** measures, specifically:

- **Degree Centrality**.
- **Betweenness Centrality**.
- **Closeness Centrality**.
- **Eigenvector Centrality**.
- **PageRank Centrality**.
- **Spectral Gap & Algebraic Connectivity**.

After doing that, we will:

- Measure the **impact** of removing the most central stations.
- **Simulate Failure Scenarios** (Random, Based on Degree/Betweenness, Cascading).
- Perform a Percolation Threshold Analysis.

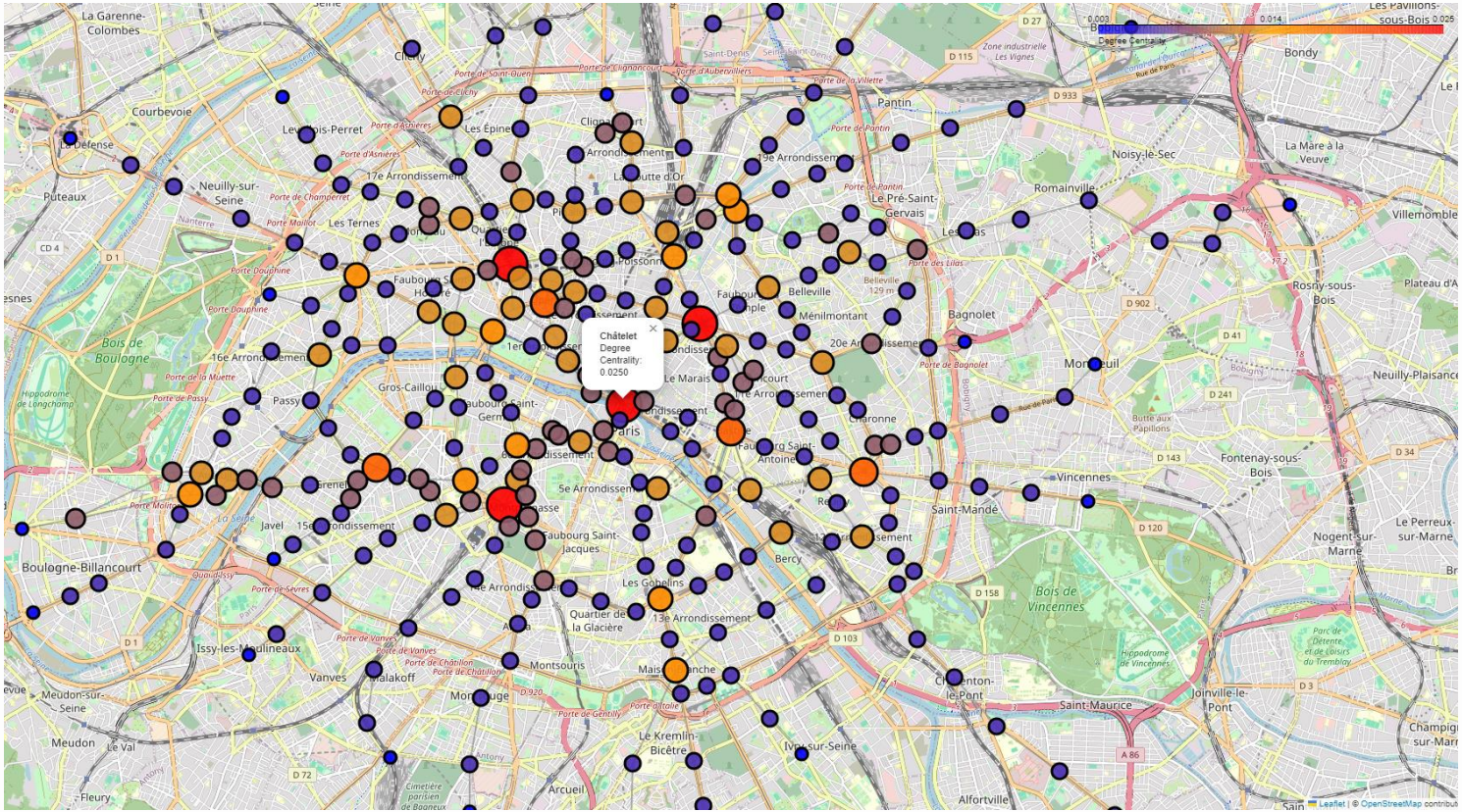
# Degree Centrality

Stations like **Châtelet**, **Saint-Lazare** and **Montparnasse Bienvenue** are key interchange hubs.

| Station | Degree Centrality |
|---|---|
| Châtelet | 0.025000 |
| Saint-Lazare | 0.025000 |
| Montparnasse Bienvenue | 0.025000 |
| République | 0.025000 |
| Opéra | 0.018750 |
| Bastille | 0.018750 |
| La Motte-Picquet - Grenelle | 0.018750 |
| Nation | 0.018750 |
| Duroc | 0.015625 |
| Michel-Ange - Molitor | 0.015625 |

Table 3.1: Top 10 Stations by Degree Centrality

# Degree Centrality



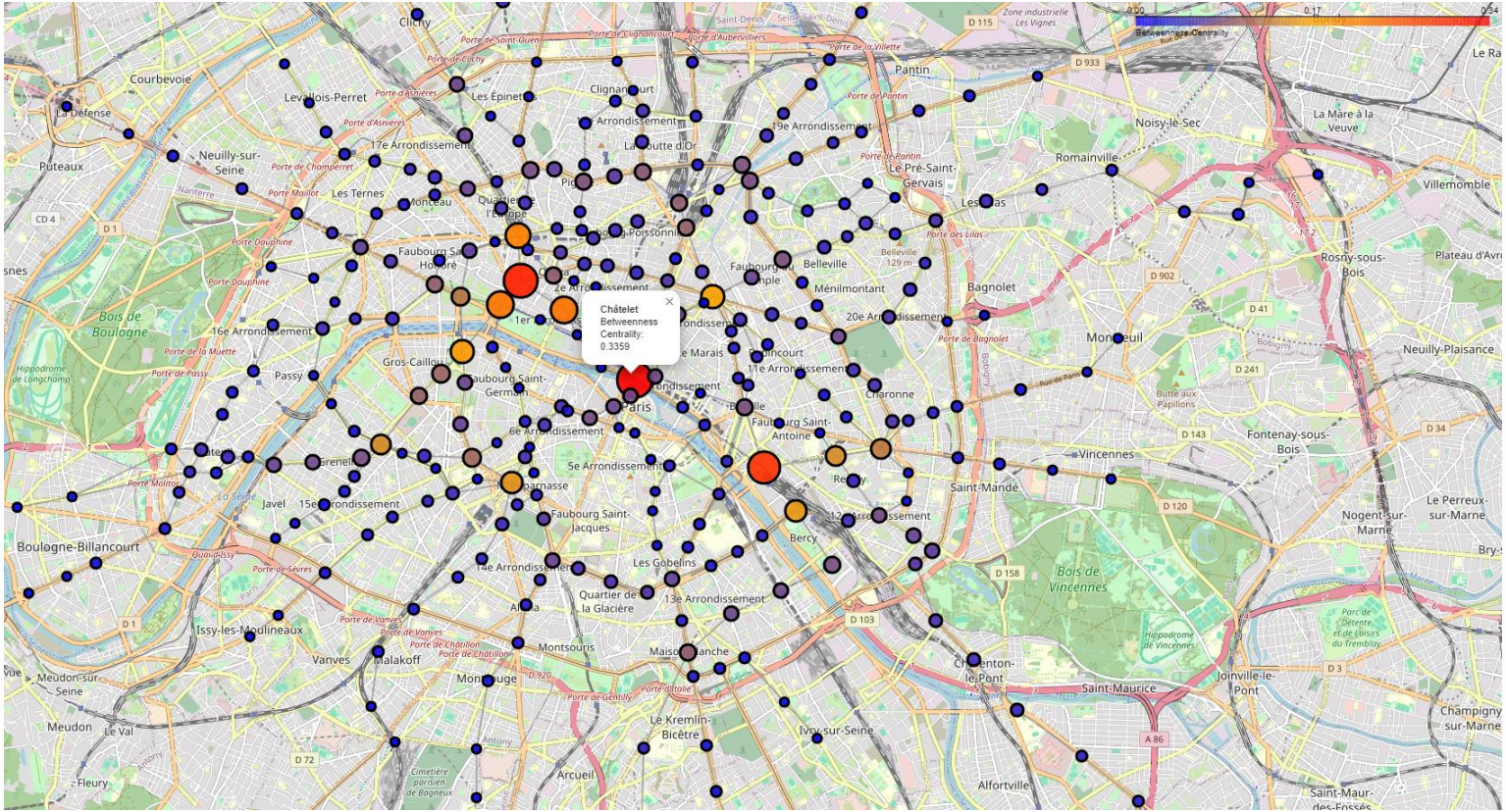
# Betweenness Centrality

**Châtelet**, **Madeleine** and **Gare de Lyon** function as network bridges.

| Station | Betweenness Centrality |
|---|---|
| Châtelet | 0.335880 |
| Madeleine | 0.304445 |
| Gare de Lyon | 0.285838 |
| Pyramides | 0.220902 |
| Concorde | 0.216997 |
| Saint-Lazare | 0.206688 |
| Invalides | 0.180556 |
| République | 0.168462 |
| Bercy | 0.154855 |
| Montparnasse Bienvenue | 0.150316 |

Table 3.2: Top 10 Stations by Betweenness Centrality

# Betweenness Centrality



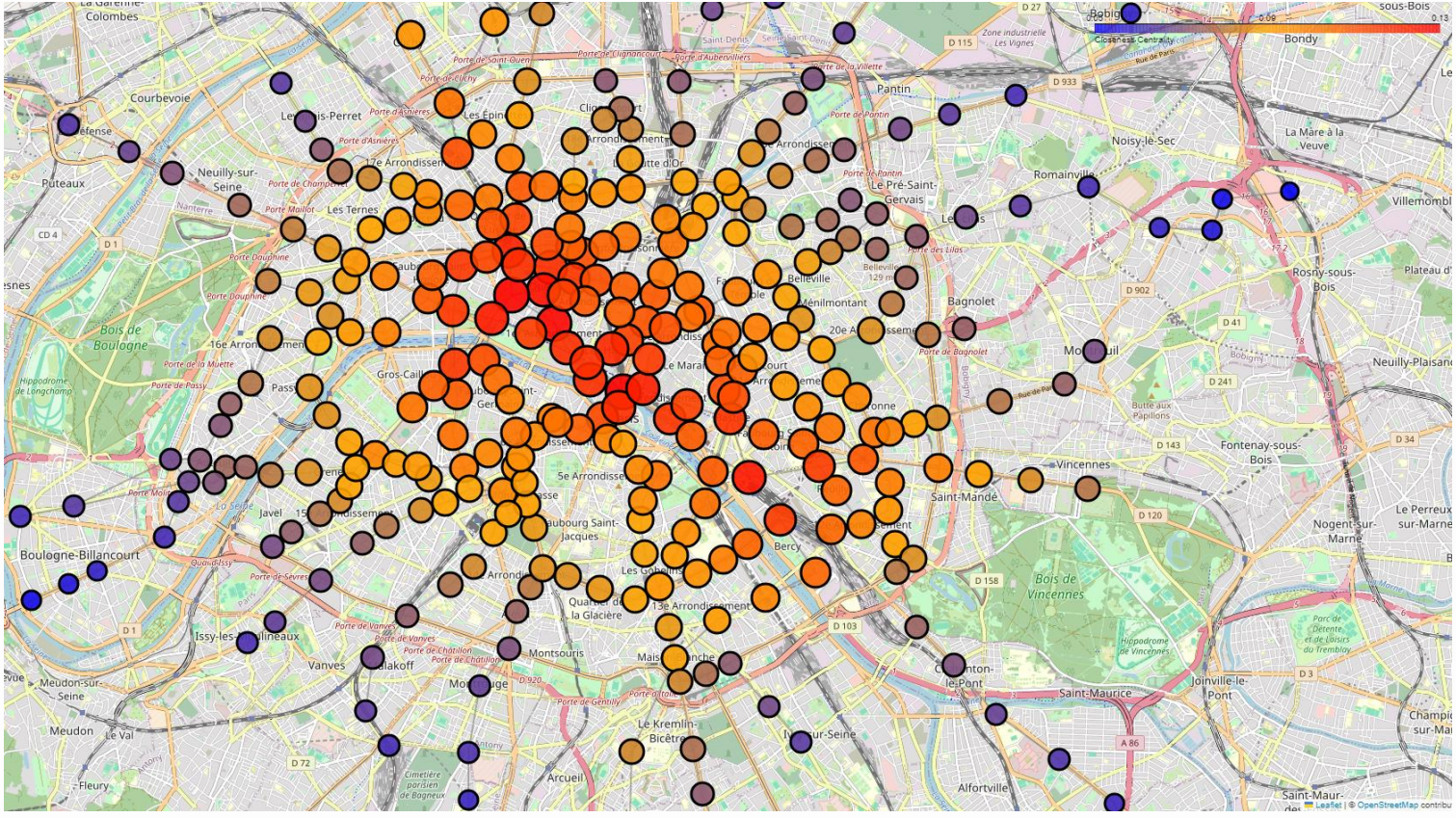
# Closeness Centrality

**Châtelet**, **Madeleine** and **Pyramides** allow to reach most destinations with minimal travel time.

| Station | Closeness Centrality |
|---|---|
| Châtelet | 0.129450 |
| Madeleine | 0.128411 |
| Pyramides | 0.127847 |
| Opéra | 0.125049 |
| Gare de Lyon | 0.123504 |
| Concorde | 0.122277 |
| Hôtel de Ville | 0.121581 |
| Saint-Lazare | 0.120937 |
| Les Halles | 0.119225 |
| Cité | 0.118212 |

Table 3.3: Top 10 Stations by Closeness Centrality

# Closeness Centrality



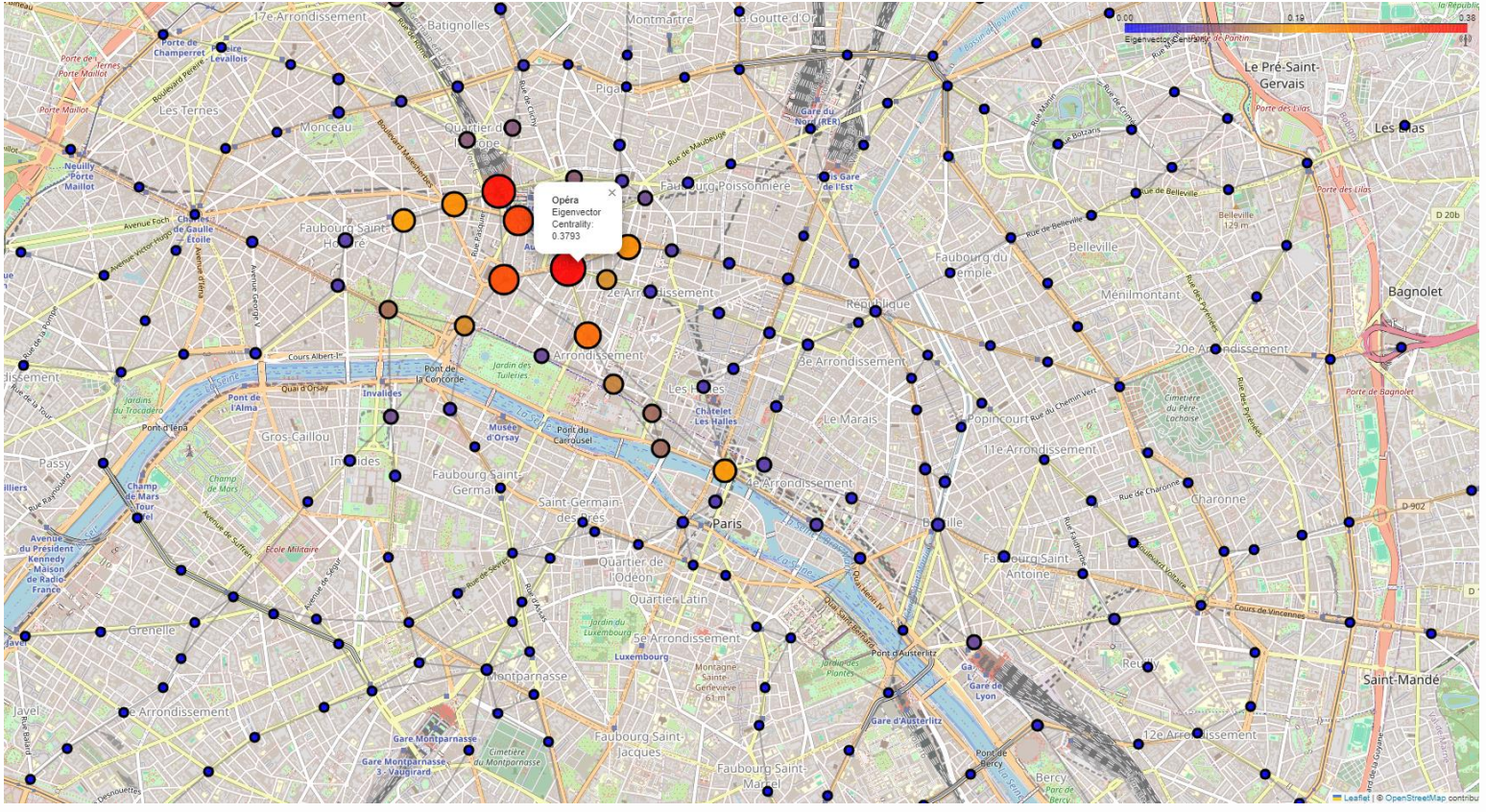
# Eigenvector Centrality

**Opéra**, **Saint-Lazare** and **Havre-Caumartin** connect to other important stations.

| Station | Eigenvector Centrality |
|---|---|
| Opéra | 0.379321 |
| Saint-Lazare | 0.362677 |
| Havre-Caumartin | 0.307907 |
| Madeleine | 0.295104 |
| Pyramides | 0.261041 |
| Chaussée d'Antin - La Fayette | 0.250421 |
| Saint-Augustin | 0.220115 |
| Richelieu - Drouot | 0.216467 |
| Châtelet | 0.206400 |
| Miromesnil | 0.196227 |

Table 3.4: Top 10 Stations by Eigenvector Centrality

# Eigenvector Centrality



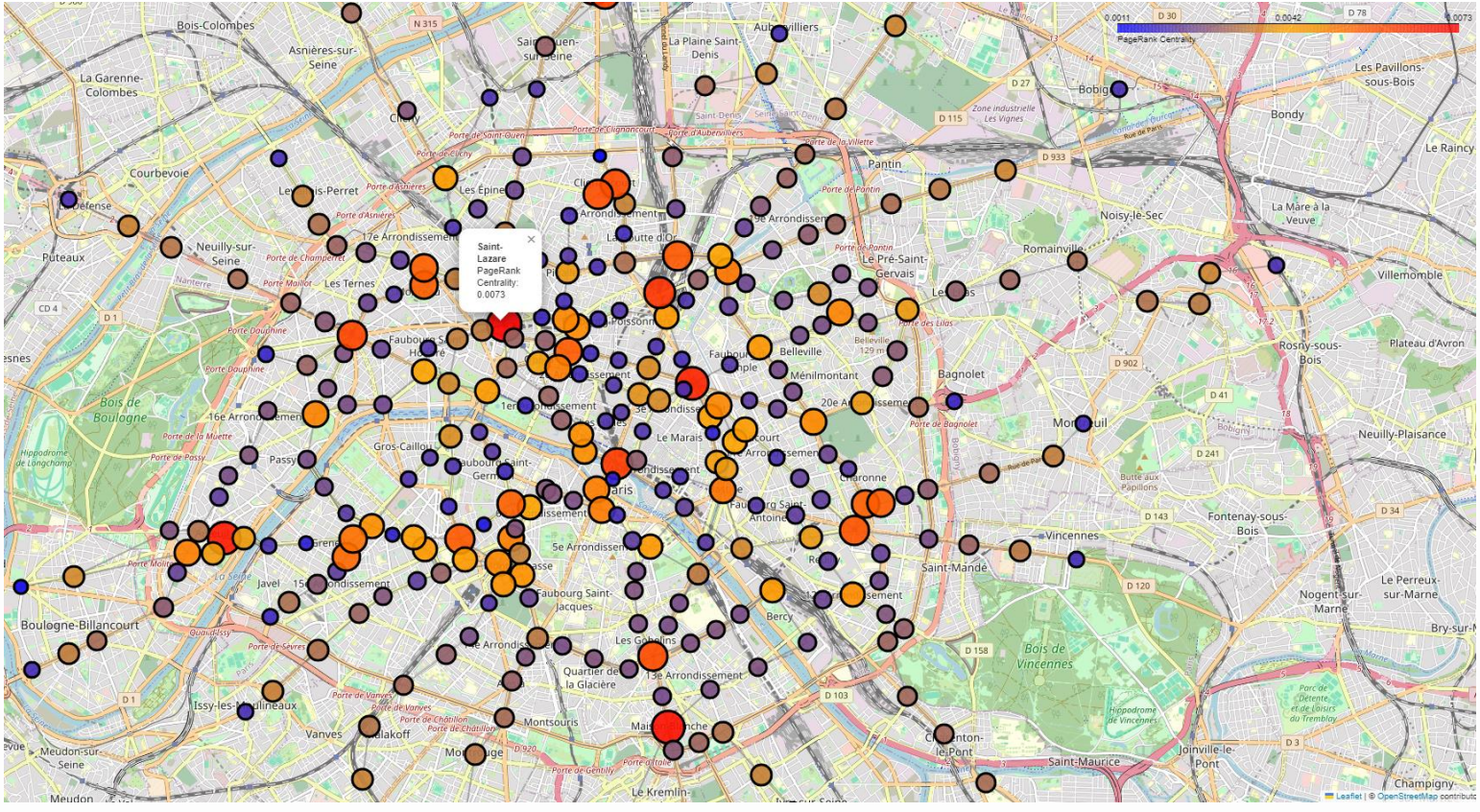
# PageRank Centrality

**Saint-Lazare** ranks highest, reinforcing its role as a strategic interchange station.

| Station | PageRank Centrality |
|---|---|
| Saint-Lazare | 0.007317 |
| Carrefour Pleyel | 0.007145 |
| Maison Blanche | 0.007033 |
| Église d'Auteuil | 0.006848 |
| République | 0.006642 |
| Gare du Nord | 0.006339 |
| Châtelet | 0.006194 |
| Saint-Denis - Pleyel | 0.006085 |
| Charles de Gaulle - Étoile | 0.006034 |
| Place d'Italie | 0.005906 |

Table 3.5: Top 10 Stations by PageRank Centrality

# PageRank Centrality



# Spectral Gap & Algebraic Connectivity

Given the Laplacian matrix $L(G) = D(G) - A(G)$, where $D(G) = diag(k_i)$ with $k_i$ the degree of node i and $A(G)$ adjacency matrix, the second-smallest eigenvalue is called **algebraic connectivity** and it indicates how well the graph remains connected if stations are removed.

We have

$$\text{Algebraic Connectivity} = 0.011314$$

The **spectral gap** is defined as the difference between the second-smallest and the smallest eigenvalue, but in the Laplacian matrix the smallest is always 0, so

$$\text{Spectral Gap} = 0.011314$$

This means removing key stations could significantly disrupt the structure.

# Impact of Removing Key Stations

We can measure the disruptions by computing the **global efficiency loss** after removing the stops with the highest betweenness. We can see **Châtelet** is the most influential hub.

| Station | Efficiency Drop (%) |
|---|---|
| Châtelet | 7.09 |
| Montparnasse Bienvenue | 4.87 |
| Gare de Lyon | 4.48 |
| Saint-Lazare | 4.43 |
| Pyramides | 2.79 |
| République | 2.54 |
| Bercy | 2.27 |
| Invalides | 2.16 |
| Concorde | 2.06 |
| Madeleine | 1.91 |

Table 3.7: Network Efficiency Drop After Removing Key Stations

# Failure Scenarios

The following types of failures are simulated:

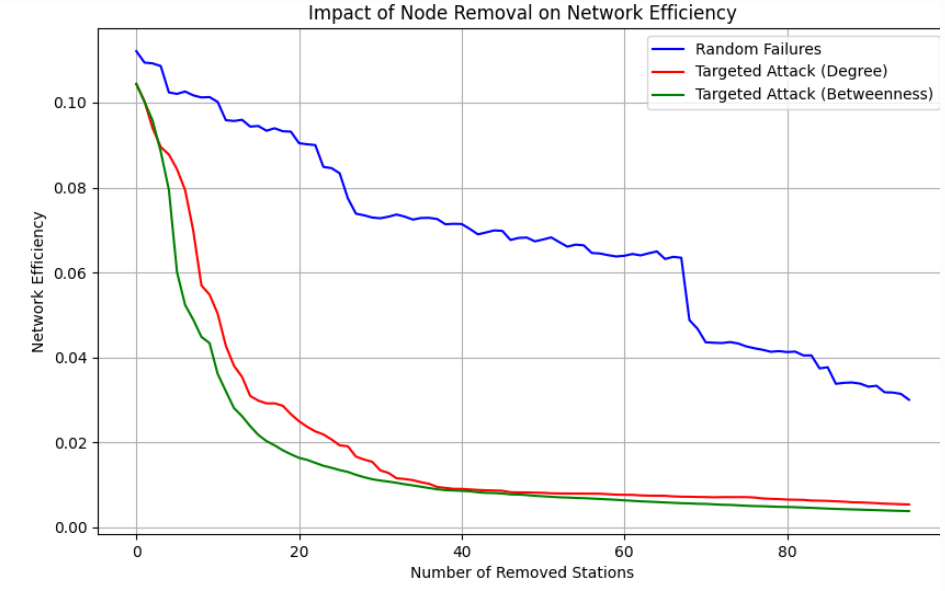
- **Random Failures**: Stations are removed randomly**.**
- **Targeted Attacks**: Stations are removed based on Degree/Betweenness.
- **Cascading Failures**: Dynamic scenario with a progressive breakdown.

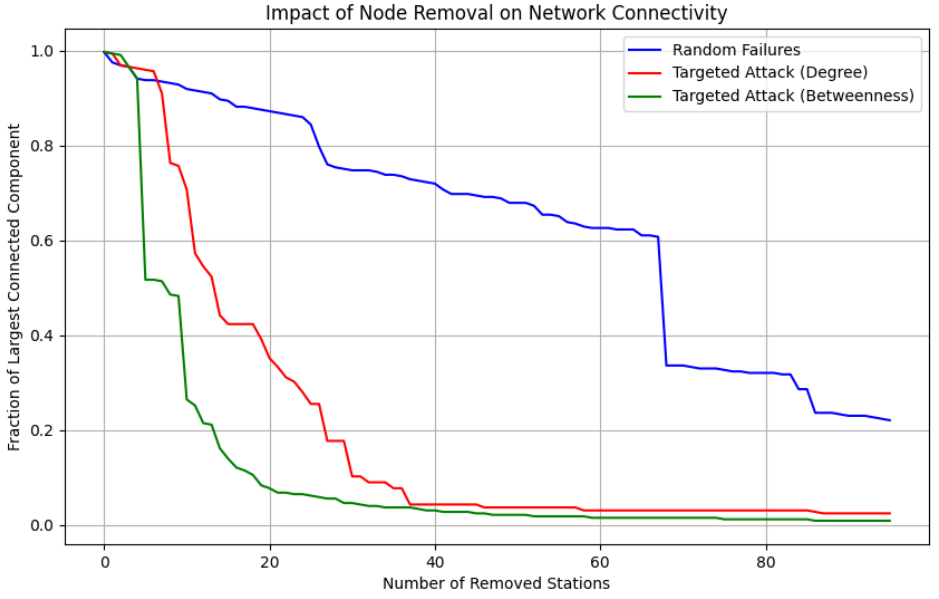
To analyze the effects, we measure the degradation of **Efficiency** and **Largest Connected Component** (i.e., fraction of stations that remain interconnected).

Cascading Failures are analyzed separately, since they are a dynamic scenario.

# Impact of Random & Targeted Attacks

**Random Failures** (blue) show a slow decline.
**Targeted Attacks**, based on **Degree** (red) and **Betweenness** (green, more damaging) result in a fast collapse of both Efficiency and Largest Connected Component (critical hubs are removed).



# Impact of Cascade Failures

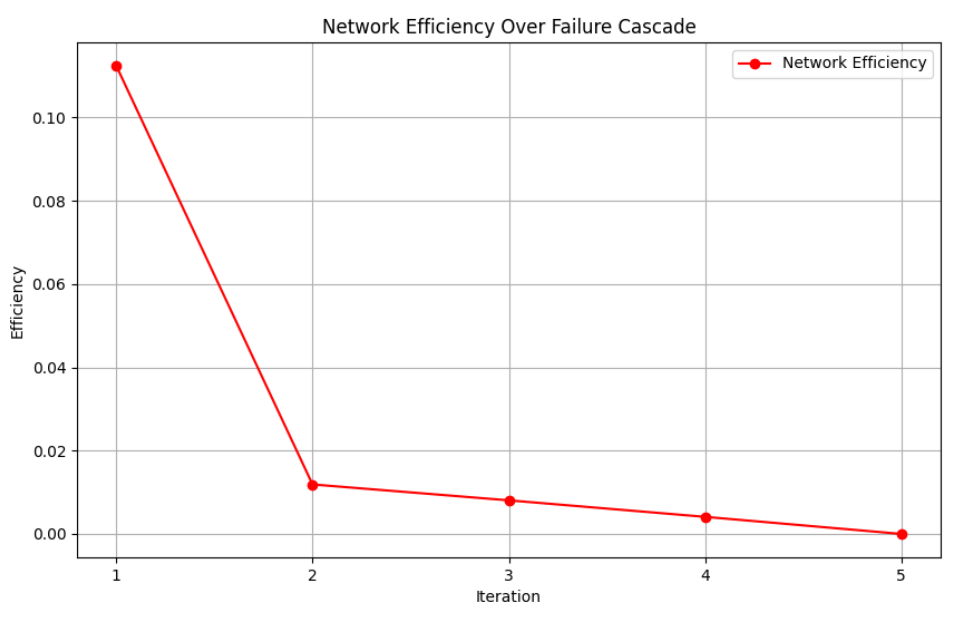
This failure process is an iterative breakdown simulation, where stations fail dynamically based on **Betweenness Centrality Overload**. Given a threshold of **10%**:

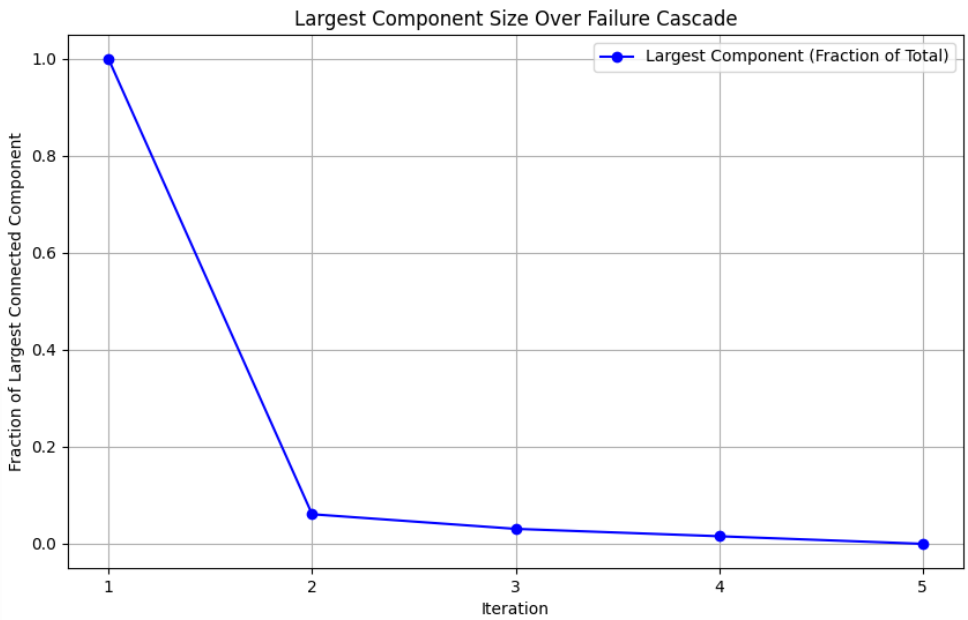
- Stations with **high betweenness** (at least 10% of the maximum betweenness) fail first.
- After each step, **betweenness is recalculated**, reflecting the new network structure.
- The process **repeats iteratively** until no additional stations exceed the threshold and fail.

We can observe that all it takes is 4 iterations for the whole network to fail.

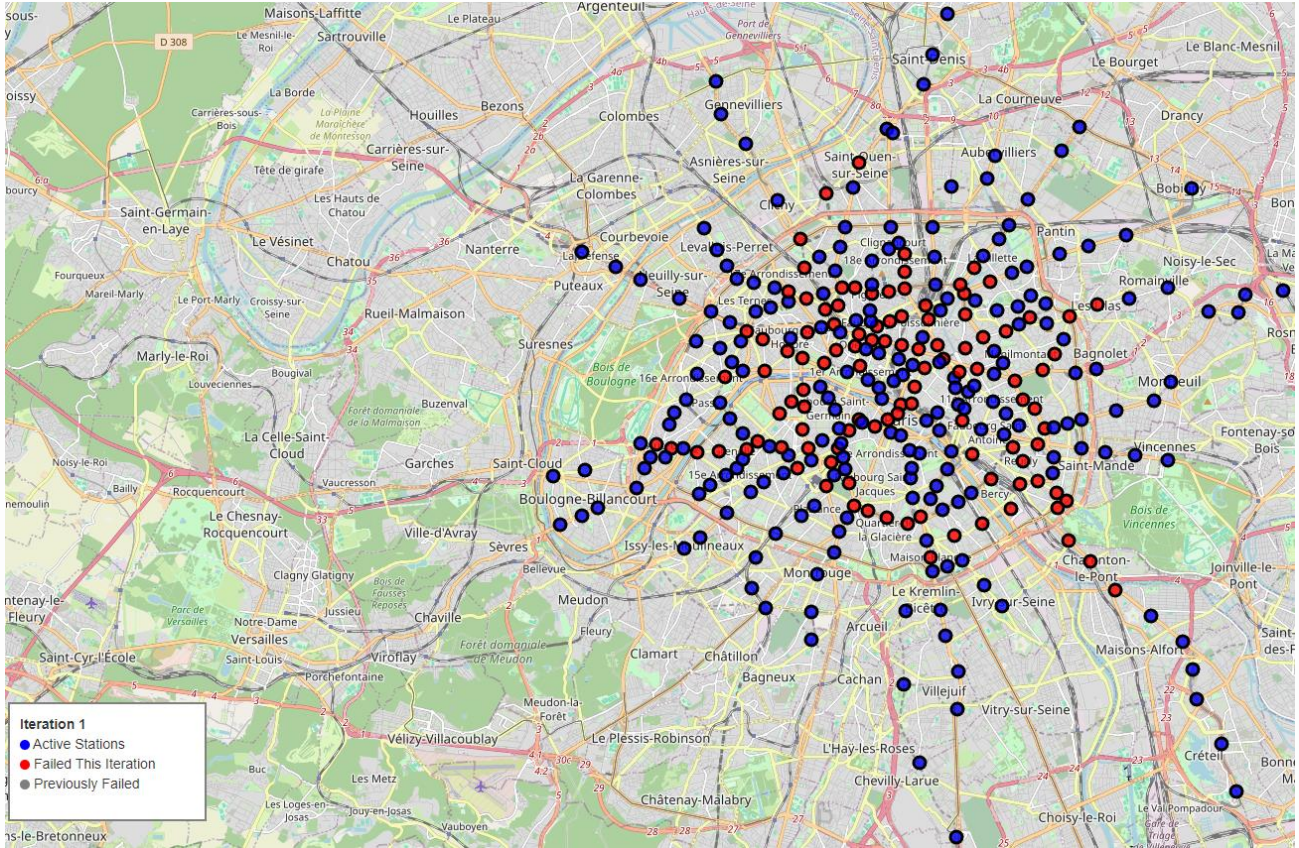
| Iteration | Failed Stations |
|---|---|
| 1 | 108 |
| 2 | 51 |
| 3 | 34 |
| 4 | 128 |

# Impact of Cascade Failures

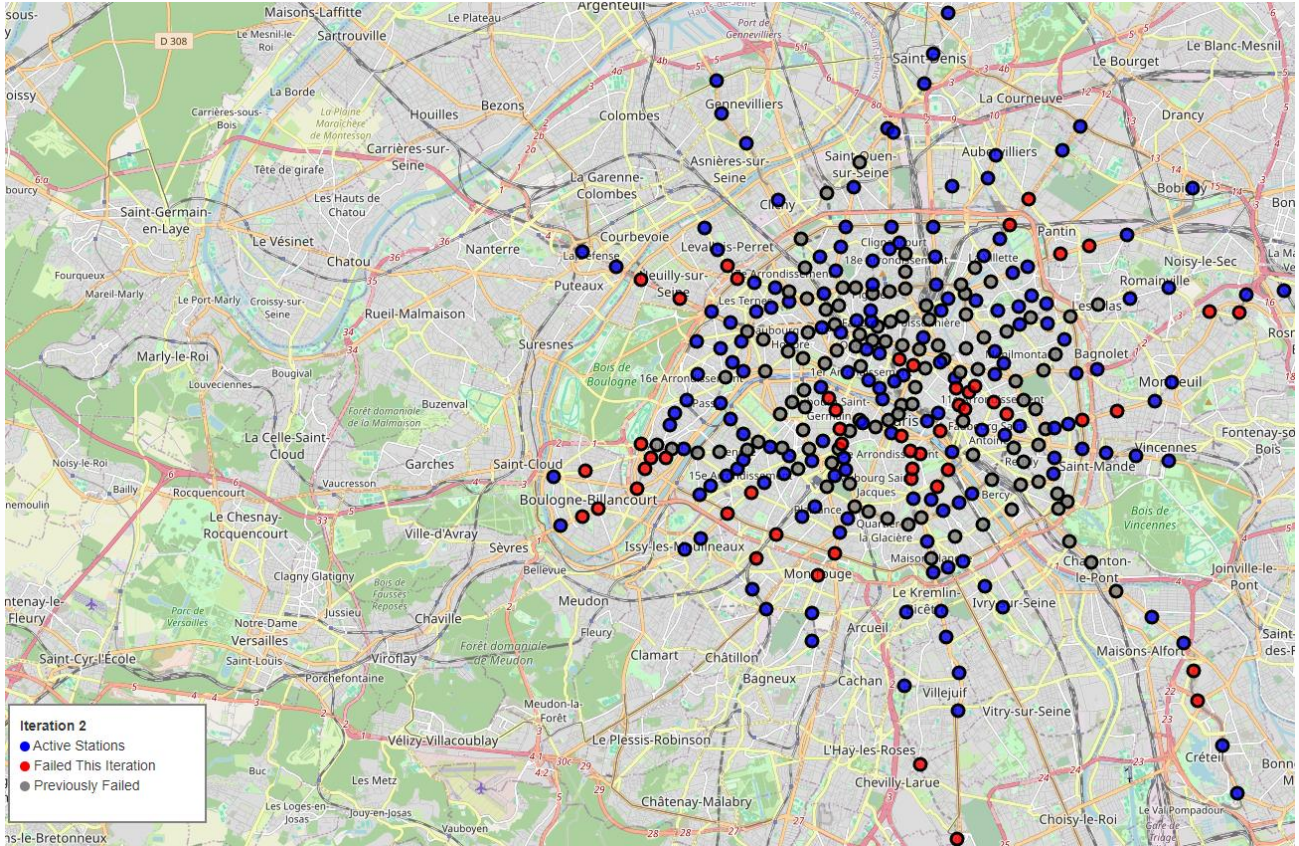


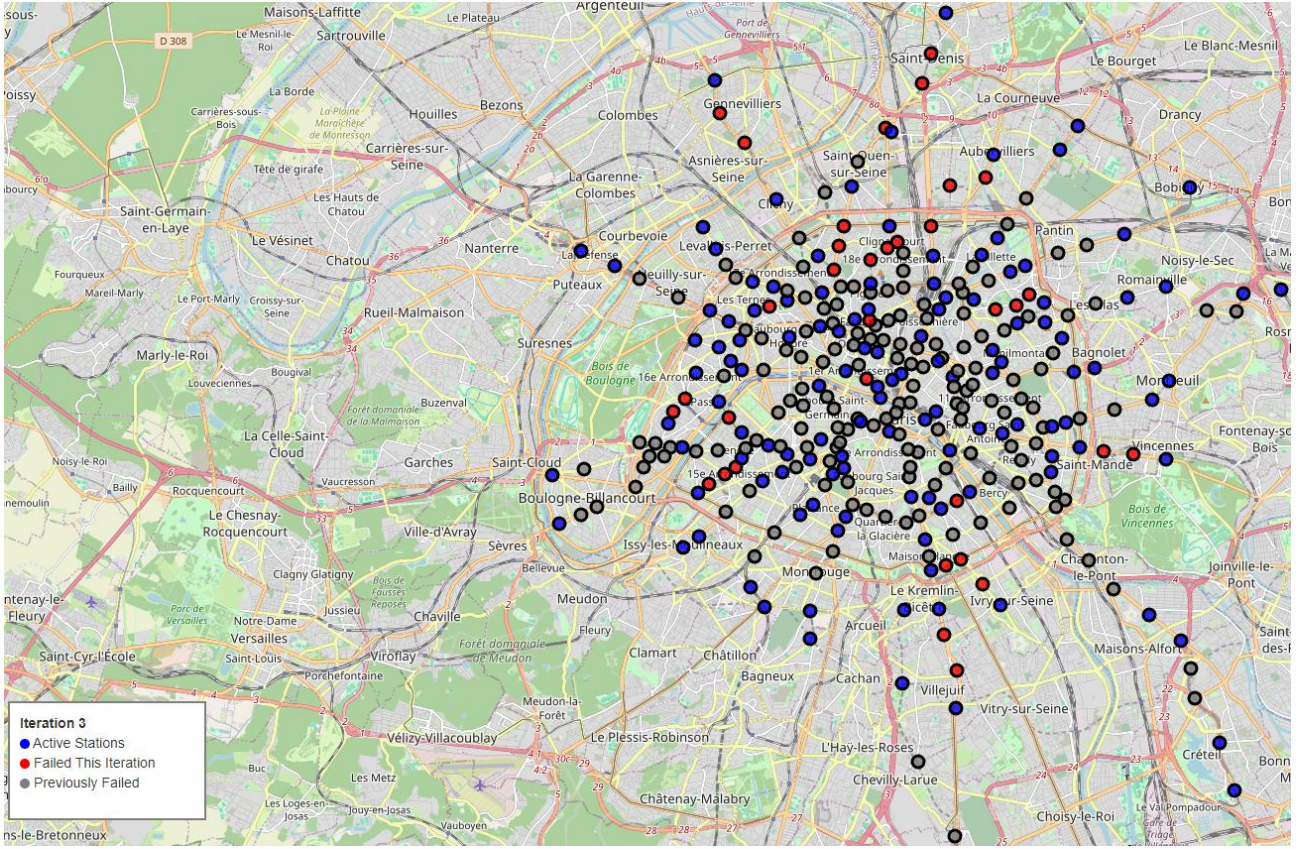
# Impact of Cascade Failures – Iteration 1



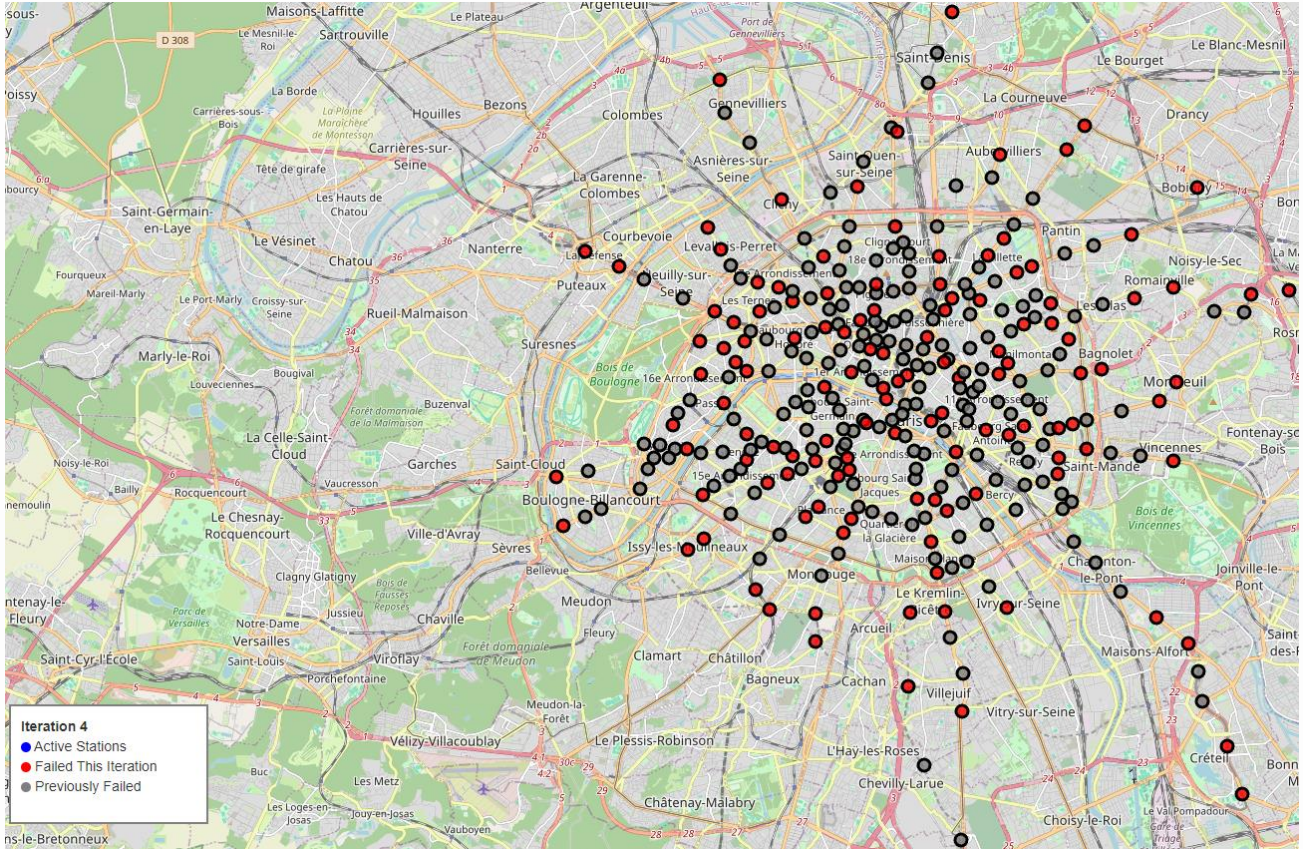
# Impact of Cascade Failures – Iteration 2



# Impact of Cascade Failures – Iteration 3



# Impact of Cascade Failures – Iteration 4

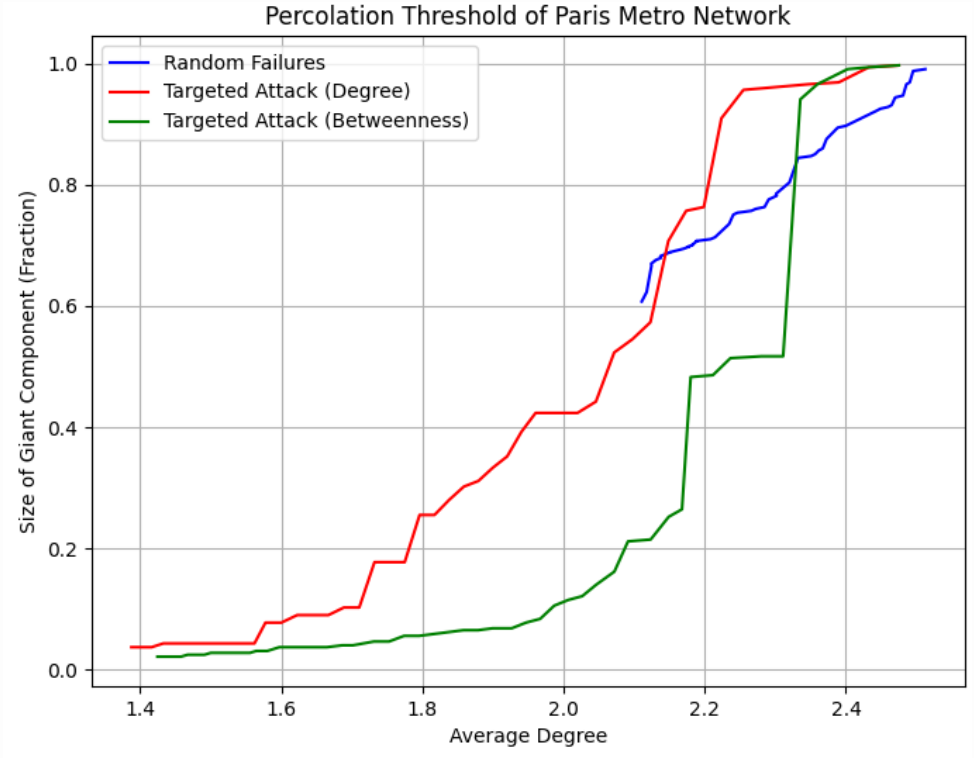


# Percolation Threshold Analysis

This analysis shows how the Largest Connected Component size decreases as the average degree decreases due to station removals.

We can confirm that **Random Failures** (blue) lead to a slow degradation, and **Betweenness-Based Attacks** (green) are even more severe than **Degree-Based Attacks** (red).

Only 30% of the stations were removed: even a **partial attack** can cause huge disruptions.



UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

# Station Load Analysis

# Overview

We define **Station Load** as the number of train arrivals at each station throughout the day.
We will examine the following time windows:

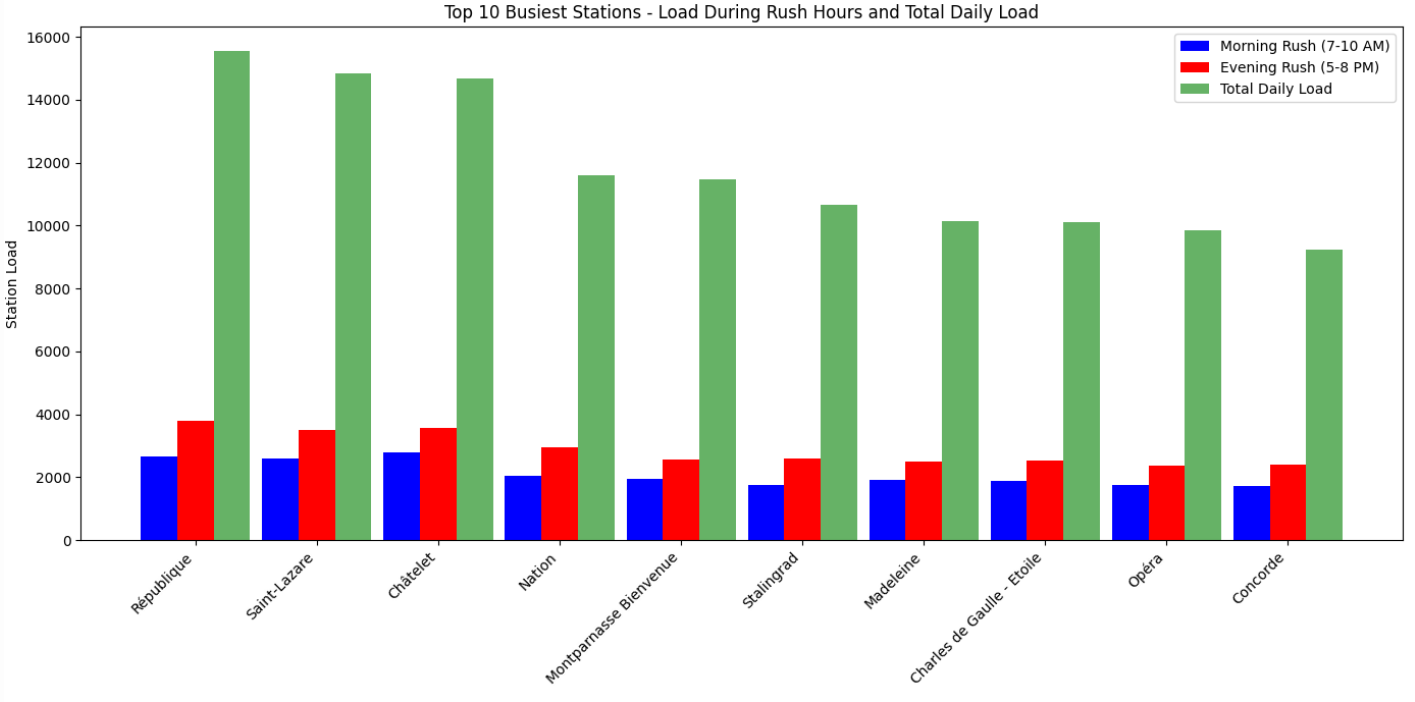
- **Morning Rush (7-10 AM)**: Commuting pattern of workers and students.
- **Evening Rush (5-8 PM)**: Captures return trips.
- **Total Daily Load**: Usage throughout the whole day.

After doing that, we will:

- Analyze the **Hourly Load** during the day.
- Show the **correlation** between Station Load and Degree/Betweenness **Centrality**.
- **Visualize the Hourly Load** Evolution.

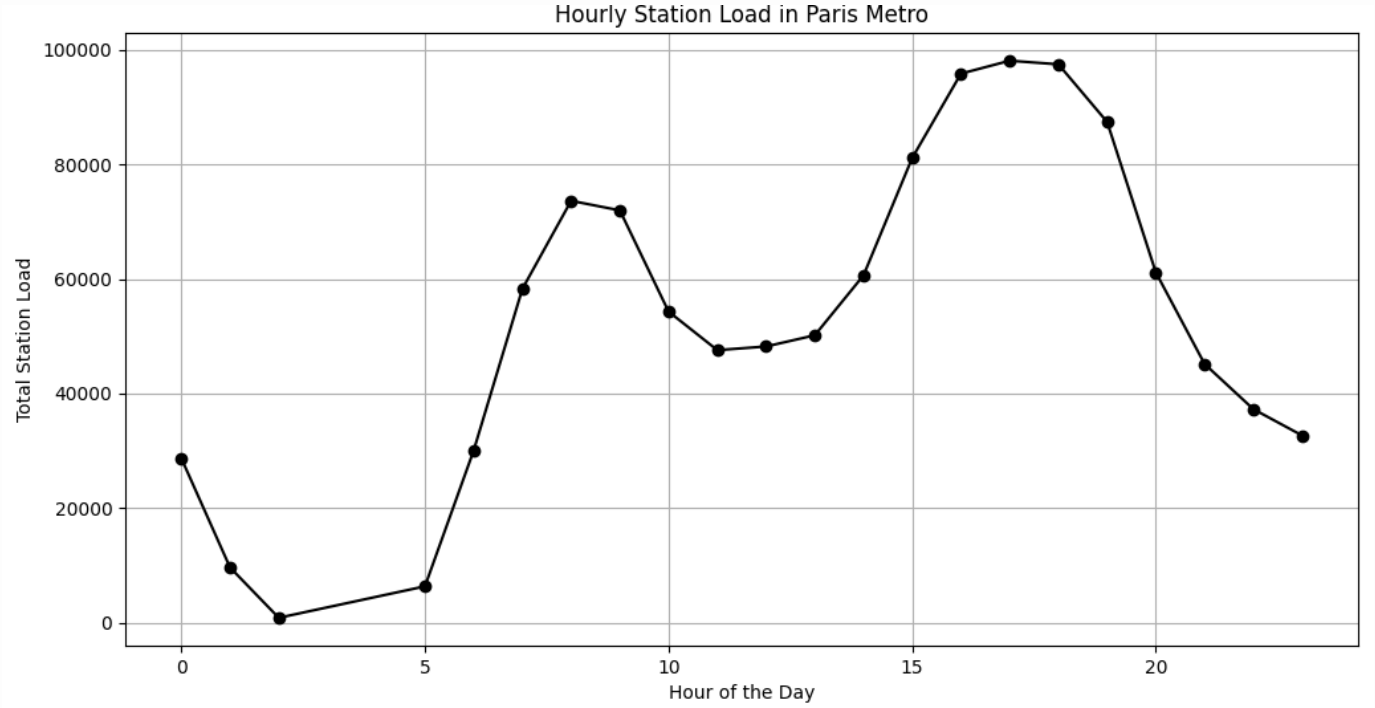
# Daily Load Distribution

**République**, **Saint-Lazare** and **Châtelet** are the most active stations, registering high arrivals in both peak periods and across the entire day.



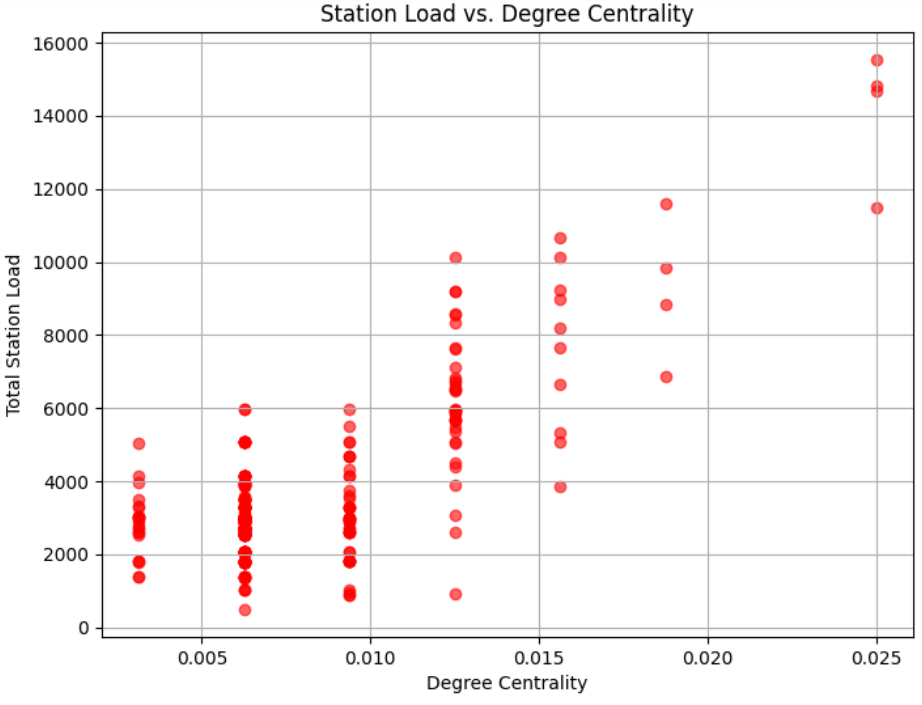
# Hourly Load Distribution

Analyzing the Hourly Load, we observe a **sharp increase** during the morning, then a second **peak** during the evening and a **late-night decline** after 10 PM.



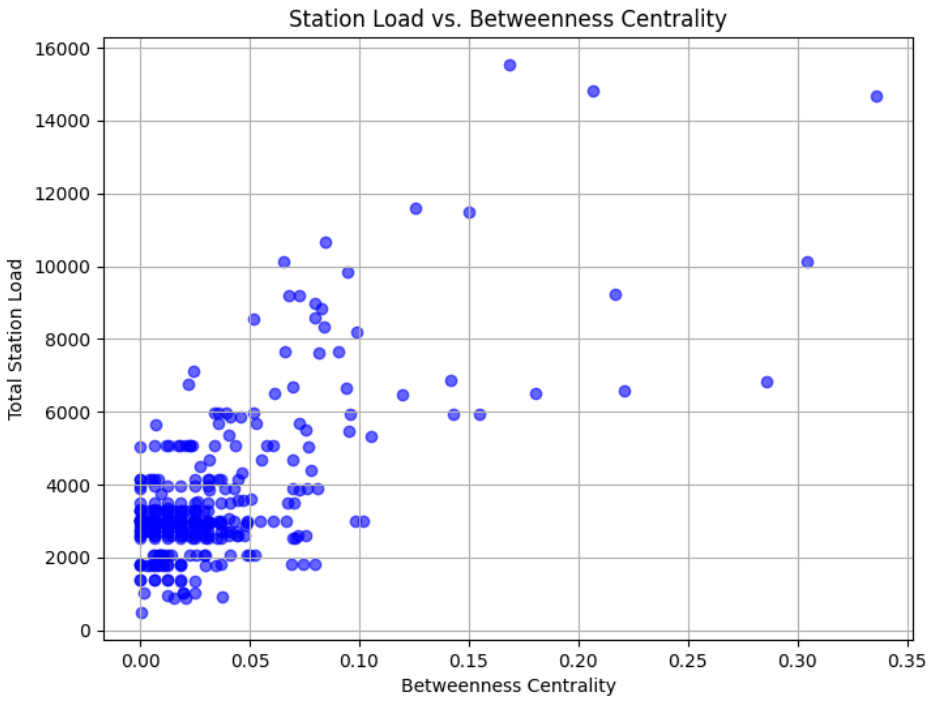
# Station Load vs. Degree Centrality

There is a **positive correlation** between **Station Load** and **Degree Centrality**, meaning that stations with many direct connections experience higher Station Load.



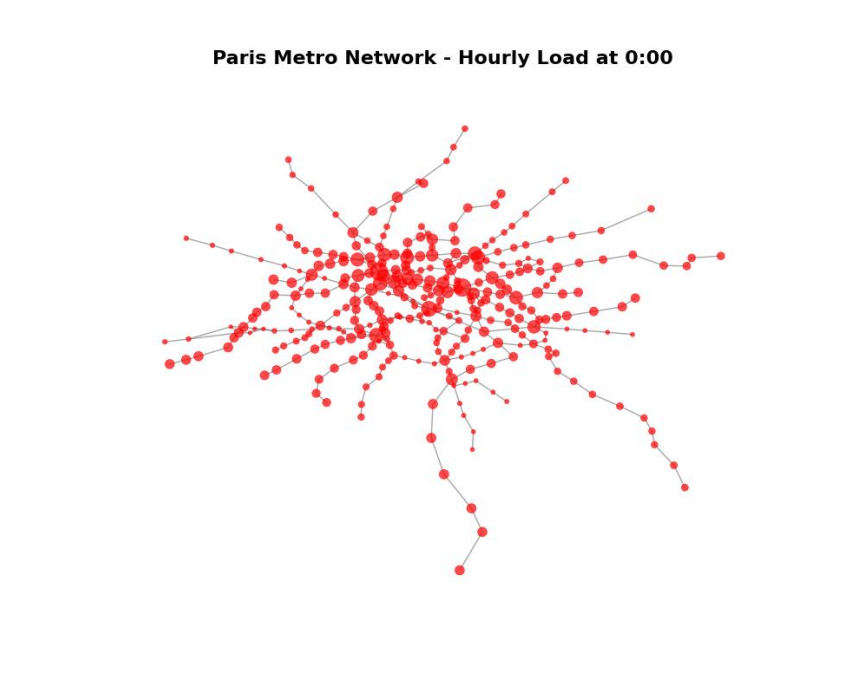
# Station Load vs. Betweenness Centrality

The correlation is even stronger with **Betweenness Centrality**, as High-Betweenness stations are critical transfer points.



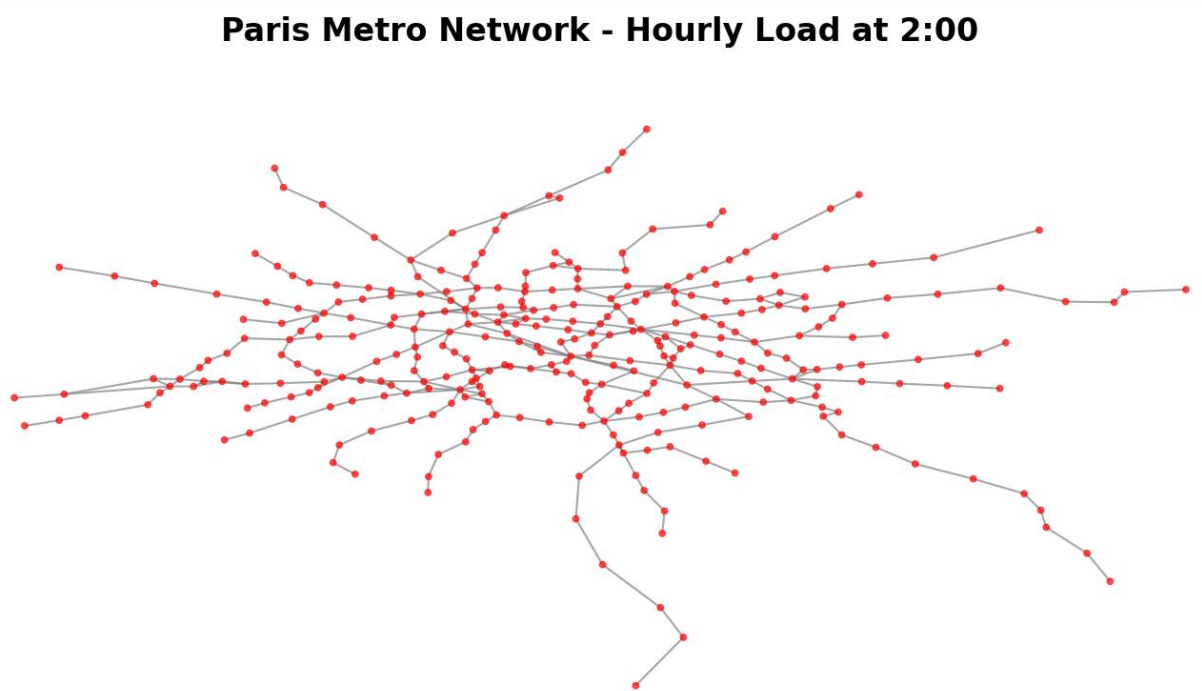
# Visualization of Hourly Load Evolution

At the end, we **animated** the Hourly Load in the graph.
We will analyze snapshots at the following times: **2 AM**, **5 AM**, **8 AM**, **11 AM** and **6 PM**.



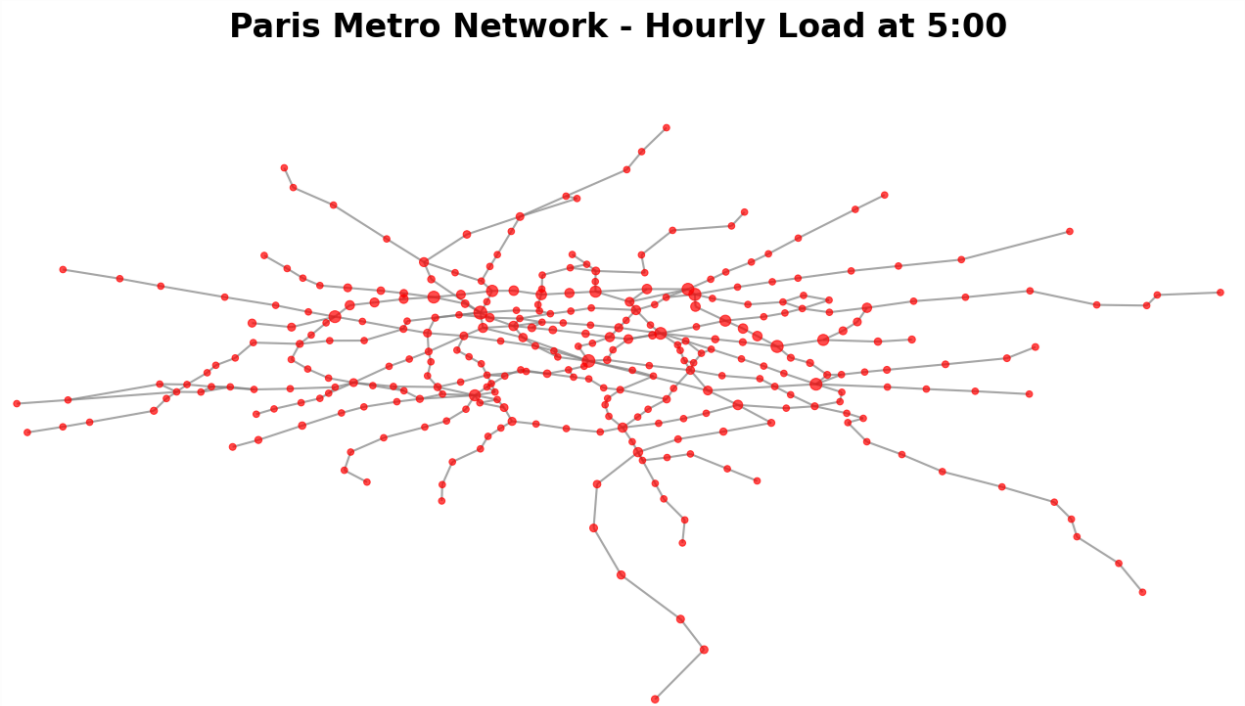
# Visualization of Hourly Load Evolution

Minimal activity is observed at 2 AM.



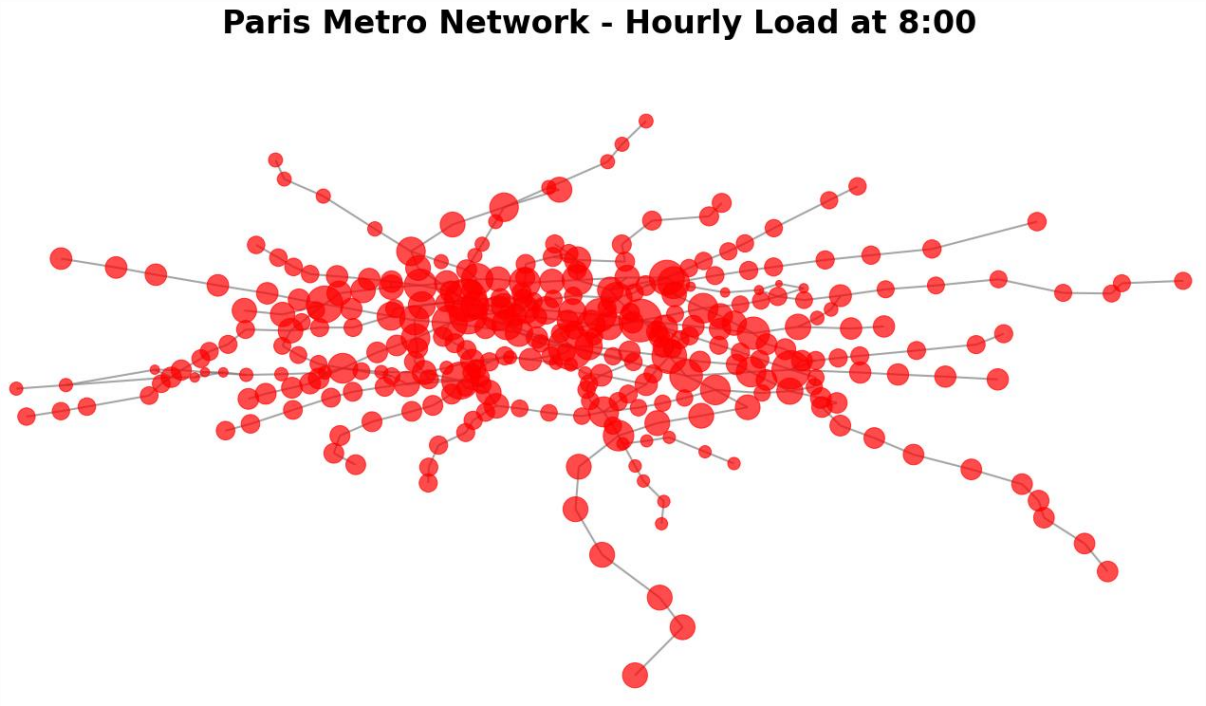
# Visualization of Hourly Load Evolution

Activity rises slightly at 5 AM, as services begin.



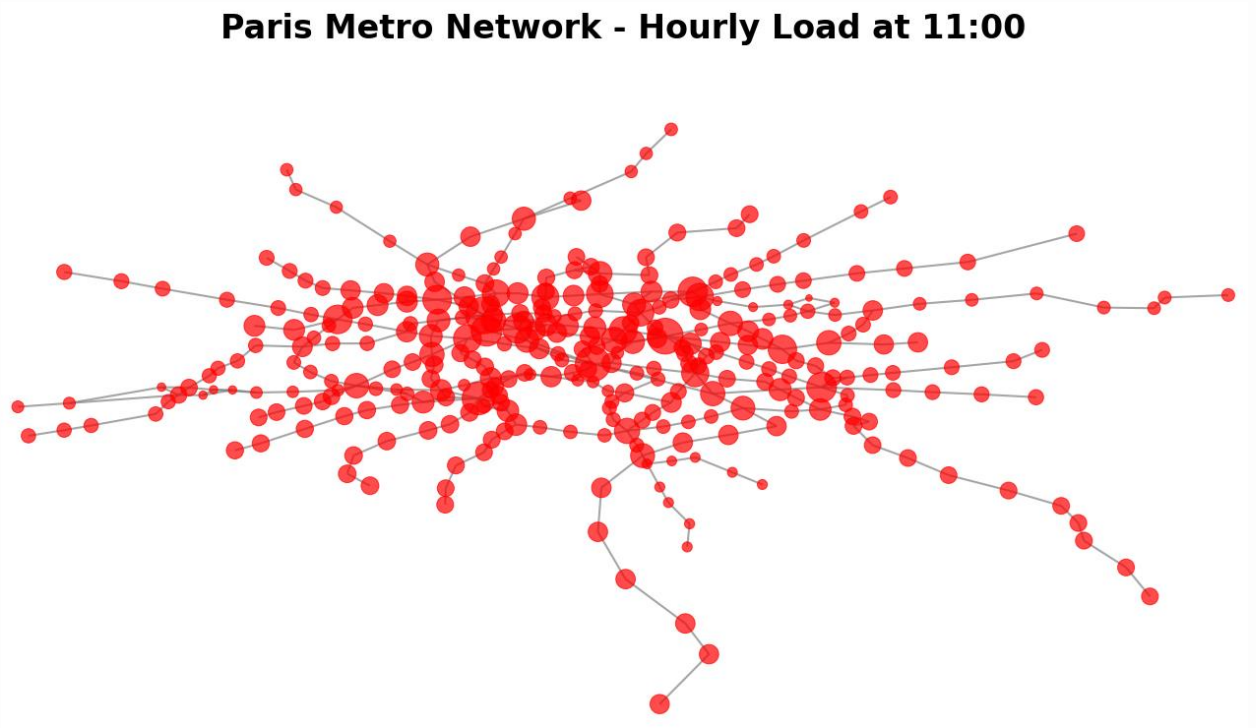
# Visualization of Hourly Load Evolution

At 8 AM, a **strong increase** in the Station Load is observed, particularly in interchange hubs.



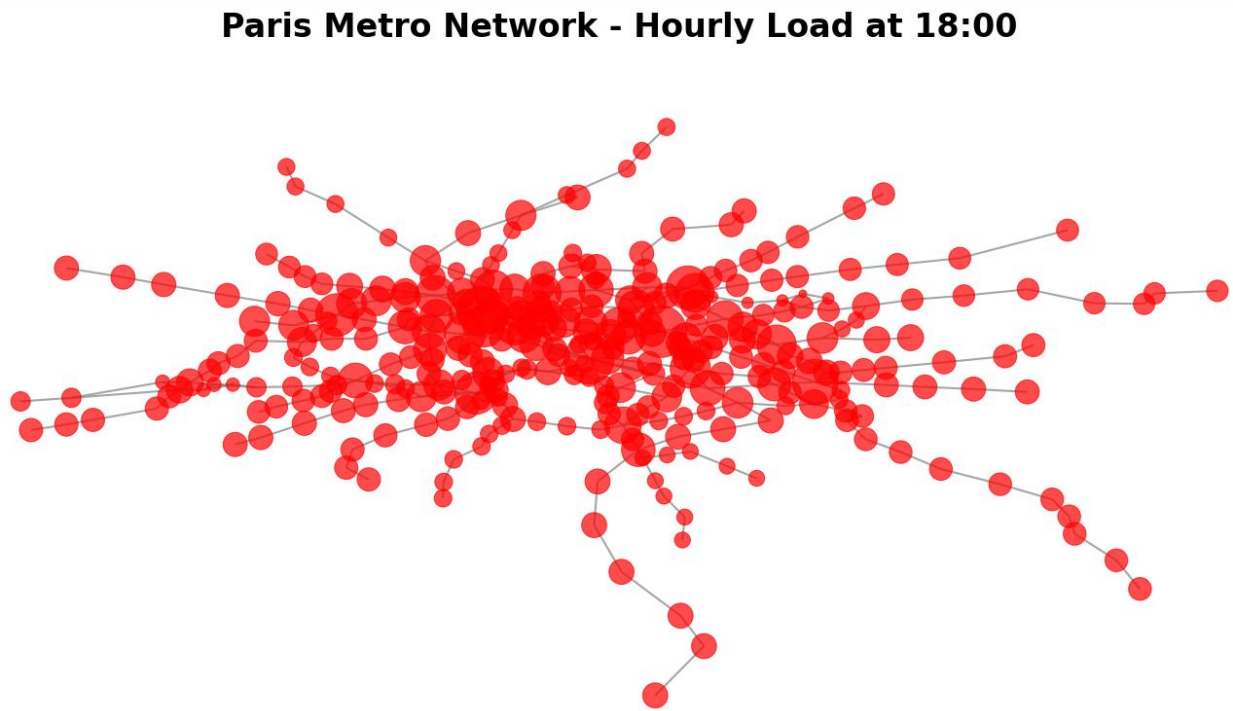
# Visualization of Hourly Load Evolution

At 11 AM, the Load distribution becomes more uniform.



# Visualization of Hourly Load Evolution

At 6 PM, we observe a **new increase** in Load, mirroring the morning peak.



# Conclusions & Future Work

# Conclusions & Future Work

We can summarize the results as follows:

- The network has a **sparse structure** with an **average degree** of **2.52** and a **low clustering coefficient** (0.0088).
- Interchange hubs such as **Châtelet** and **Saint-Lazare** play a **central role** in maintaining the network connectivity.
- The network is **highly sensitive** to **Targeted Attacks**.
- A breakdown of **High-Betweenness stations** showed that the network collapses after just **four iterations**, losing **321** stations.
- High-load stations coincide with high centrality.

In the future, the following aspects could be explored:

- Include ridership data to model passenger flow.
- Expand the network, including RER, trams and buses to analyze the interconnections.

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

# Thanks for your attention!