Claude Bernard Lyon University Lyon1

Option: M2 Data Science

Module: Data Mining

TCL Networking

Prepared by

Mr.BEN TOUHAMI Mohamed
Rida
Mr.Motassim Hamza

Supervised by

Pr.Cazabet Rémy

Academic Year 2025/2026

Table of Contents

1. Project Overview

This project models the TCL (Transports en Commun Lyonnais) public transportation network as a graph structure to enable efficient route planning and network analysis across Lyon's metros and trams systems (Zone 1).

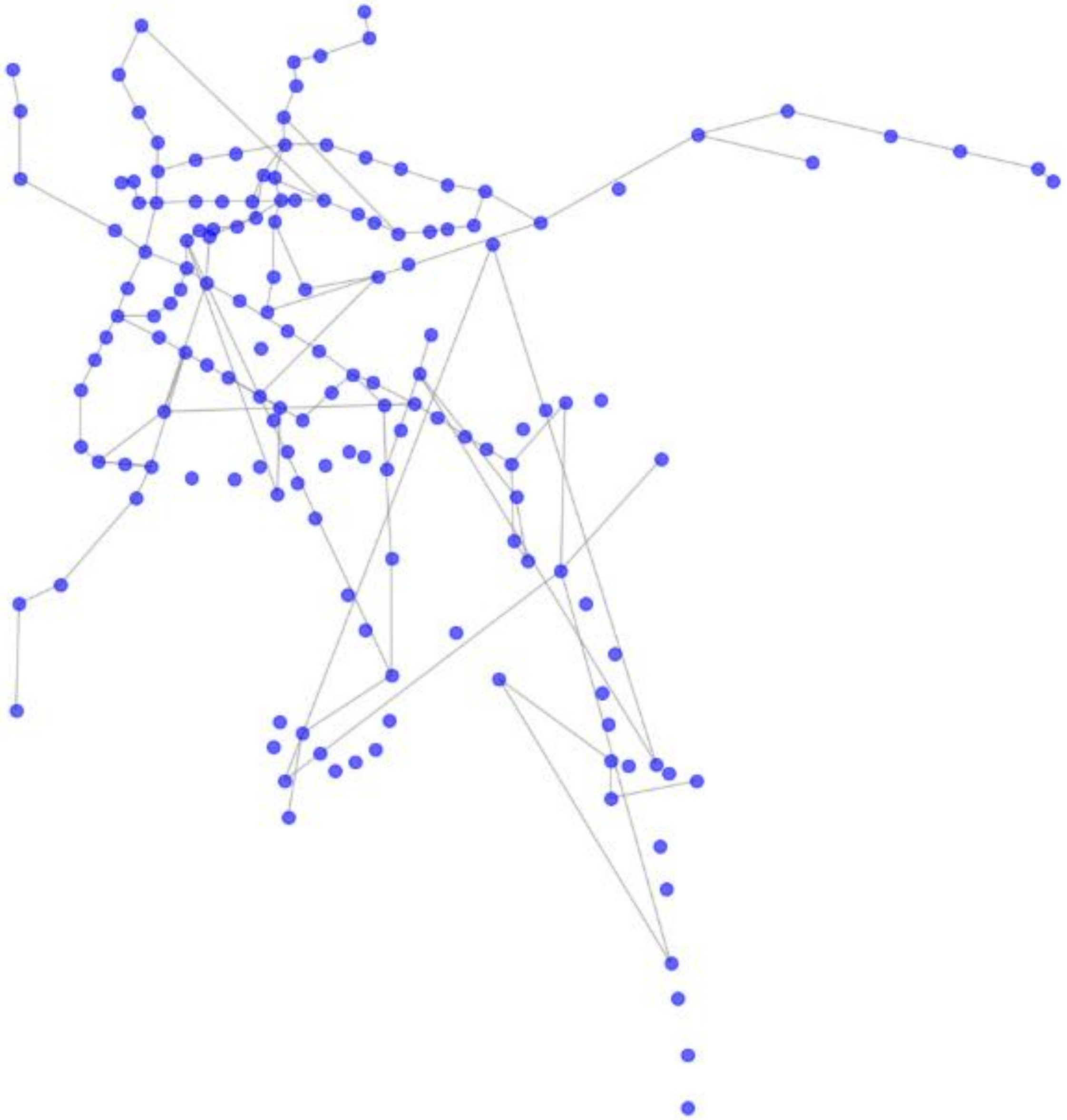


Figure 1: Graph overview

a. Core Concept: Unified Stop Representation

The Node Merging Approach

A key design principle of this project is that all stops sharing the same name or address are treated as a single nod**e** in the graph, regardless of which transport line serves them.

Why This Matters

In real-world transportation networks, the same physical location is often served by multiple lines:

Metro stations where multiple lines intersect (e.g., "Charpennes Charles Hernu" serves Metro A, B, Tram T1, T4, and bus C2)

Tram stops that share platforms or are within walking distance

Implementation Benefits

1. **Natural Transfer Points**: When multiple lines serve the same stop, passengers can transfer between them at that location. By merging these into a single node, transfers are implicit in the graph structure.
2. **Simplified Route Planning**: From any point in the network, the graph automatically considers all available transport options, not just a single line.
3. **Real-World Navigation**: Mirrors how passengers actually think: "I'm at Bellecour - which lines can I take to reach my destination?" rather than "I'm at Bellecour on Metro A specifically."
4. **Reduced Graph Complexity**: Instead of having separate nodes for "Bellecour (Metro A)", "Bellecour (Metro D)", "Bellecour (Bus C10)", we have one "Bellecour" node connected to all relevant lines.

Example Scenario

Question: "I'm at Saxe - Gambetta. How can I reach Vaulx-en-Velin La Soie using the TCL network?"

Without Node Merging:

- You'd need to specify: "I'm at Saxe - Gambetta on Metro B"
- The system would only consider paths starting from that specific node
- Transfer points would need explicit modeling

With Node Merging (Our Approach):

- You simply say: "I'm at Saxe - Gambetta"
- The system knows this location is served by:
  - Metro B
  - Metro D
  - Tram T1
- All possible routes from any of these lines are automatically considered
- Optimal path selection happens naturally

b. Graph Structure

Graph Properties

- **UnDirected Graph**: Travel is bidirectional on all lines
- **UnWeighted Edges**: Can represent time, distance, or stop count but ignored for basic analysis

Nodes

- **Node ID**: Unique stop name (e.g., "Bellecour", "Perrache", "Charpennes Charles Hernu")
- **Node Attributes**:
  - Stop name
  - List of lines serving this stop
  - Coordinates

Edges

- **Connection Type**: Sequential stops on the same line

Multi-Modal Integration

The graph seamlessly integrates:

- **4 Metro lines** (A, B, C, D)
- **7 Tram lines** (T1-T7)
- **TramBus** (TB11)

c. Use Cases

1. **Route Finding**: Find the shortest path between any two stops in the network
2. **Transfer Optimization**: Identify optimal transfer points for multi-line journeys
3. **Network Analysis**: Analyze connectivity, central hubs, and network robustness
4. **Accessibility Mapping**: Determine reachability from any point in the network
5. **Service Planning**: Identify underserved areas or potential new connections

d. Key Advantages

1. **Intuitive**: Matches mental model of "being at a location" rather than "being on a specific line"
2. **Comprehensive**: Can consider all transport options
3. **Flexible**: Works for any journey type (single line, transfers, multi-modal)

d. Future Enhancements

1. **Walking Connections**: Add edges for nearby stops that aren't officially connected
2. **Real-Time Data**: Integrate live departure times and service disruptions
3. **Cost Optimization**: Consider fare zones and ticket prices
4. **Accessibility Features**: Include wheelchair access and elevator availability
5. **Time-Dependent Routing**: Account for service schedules and frequencies

2.Data Description

a.Overview

This dataset contains information about the **public transport network of Lyon (TCL)**, including stops, geographic coordinates, and the lines that serve each stop. It was collected from data.grandlyon.com and generated using LLMS knowledge (Claude & ChatGPT) and processed to produce a **clean, structured version** suitable for graph analysis, geospatial visualization, and network modeling.

b. Raw Data

The first **raw dataset** was originally obtained in JSON format from data.grandlyon.com. Each record contained nested structures, duplicated entries, and inconsistent formatting.

You can find the raw data file here:
`/data/raw/points-arret-reseau-transports-commun-lyonnais.json`

Common Issues in Raw Data

- Many transport types included (bus, tram, metro) → focus needed on **base network only**.
- Duplicates across stops (same name or address appearing multiple times).
- Mixed formatting in the `"desserte"` column (e.g. `"A : Perrache"`, `" T1"`, `"t3"`).
- Inconsistent case and trailing spaces.
- Redundant or irrelevant columns for network analysis (`ascenseur`, `pmr`, etc.).

Raw Data Schema

Column	Type	Description
`id`	`int`	Unique stop identifier
`nom`	`string`	Stop name
`adresse`	`string`	Stop address
`commune`	`string`	Municipality of the stop
`insee`	`string`	INSEE code of the commune
`lat`	`float`	Latitude coordinate
`lon`	`float`	Longitude coordinate
`desserte`	`string`	List of lines serving the stop:A/R as Aller Retour for each csv (e.g. `A:R,T1:A,C26:A`)
`pmr`	`bool`	Accessibility flag (PMR access)
`ascenseur`	`bool`	Presence of elevator
`escalator`	`bool`	Presence of escalator
`last_update`	`datetime`	Last update date of the record
`zone`	`string`	Zone classification (`1, 2, 3, 4, zone exterieur`)

Raw Data Example Record

```json
{
  "id": 2435,
  "nom": "PD - Vivier Merle",
  "adresse": "Boulevard Vivier Merle, Lyon 3e",
  "commune": "Lyon",
  "insee": "69383",
  "lat": 45.76056,
  "lon": 4.85932,
  "desserte": "A:R,T1:A,C26:A",
  "pmr": true,
  "ascenseur": true,
  "escalator": true,
  "last_update": "2022-10-20T14:32:00Z"
}
```

Basic Cleaning Steps Applied

1. **Zone Filtering**
   - Retained only stops within Zone 1:
2. **Column Pruning**
   - Removed unused or redundant metadata:
3. **Desserte Cleaning and Filtering**
   - Split desserte strings by commas.
   - Retained only **metros (A–D)** and **trams (T1, T2, …)**.
   - Removed empty entries.
   - Grouped rows having the same **stop name** or **address** to merge duplicates.
   - Merged all unique transport lines into a comma-separated string.

c. Stops Data

The second Json file was mainly generated by LLMs with high varacity considerations and same names formatting as in the original data to obtain the sequences of stops for each transport (only trams, tramBus, and Metros for simplicity and missing data after several attempts to scrap original data from official TCL website) then the data was verified mannualy to ensure its veracity

You can find the raw data file here:
/data/raw/trans-lines.json

Common Issues in Stops Data

- Some missing stops.
- Some stops names are not the same as in the first data file (e.g Vivier Merle vs V. Merle)

Stops Data Schema

Column	Type	Description
desserte	string	Transport
stops	List[string]	Stops names of the transport

Stops Data Example Record

```
{
  "B": [
    "Charpennes Charles Hernu",
    "Brotteaux",
    "Gare Part-Dieu V.Merle",
    "Place Guichard",
    "Saxe - Gambetta",
    "Place Jean Jaurès",
    "Jean Macé",
    "Debourg",
    "Stade de Gerland Le LOU",
    "Gare d'Oullins",
    "Oullins Centre",
    "St-Genis-Laval Hôp. Sud"
  ],
}
```

d. Cleaned Data (Final Version)

The **cleaned dataset** (nodes_data.csv & edges_datav2.csv) represents the final, structured version after full preprocessing that is concerned to be used as nodes and edges for network construction. It contains base data (Trams T1, .., T7, TB11 & Metros A, B, C, D) and links between each

3. Network *Analysis*

a. Degree

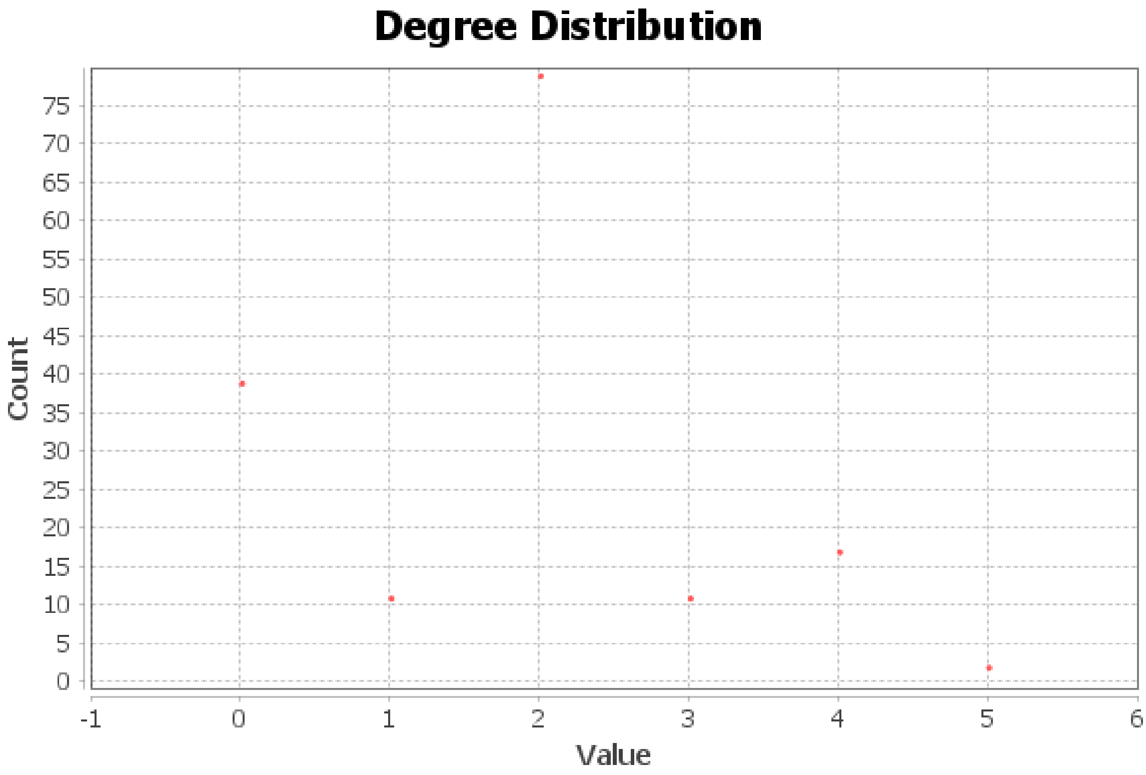
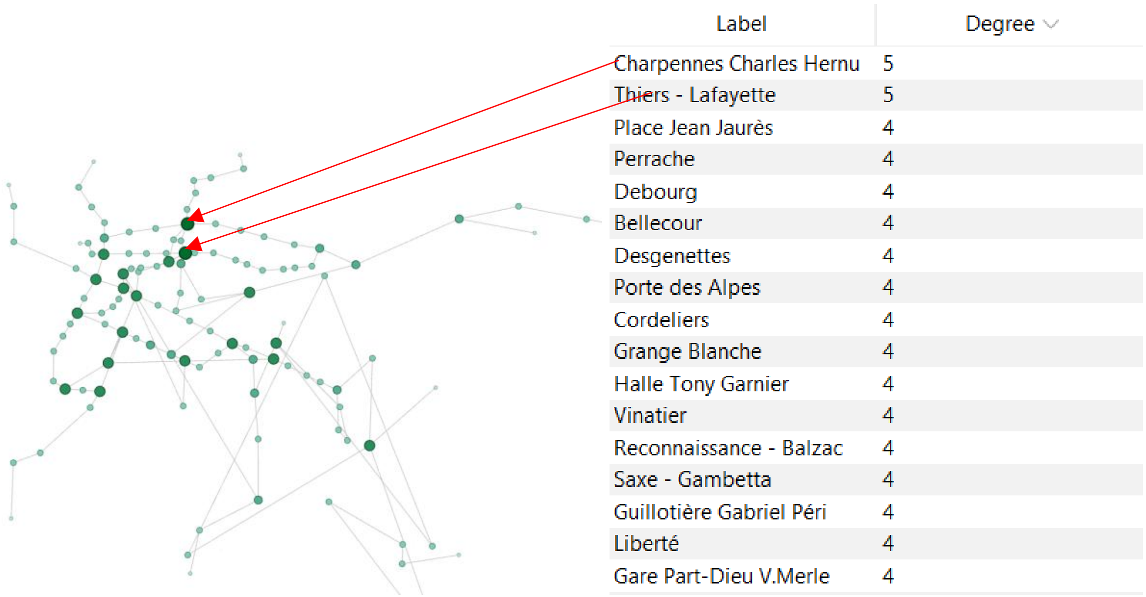


Figure 2: Degree Distribution

We notice in the graph that we have about 38 vertices without edges. This represents 38 isolated vertices (nodes with a degree of zero), as indicated in the degree distribution analysis. This discontinuity is a direct consequence of the incomplete data scope, which was restricted exclusively to metro and tram lines. We were unable to secure data for other significant transport modalities, such as bus routes, resulting in thirty-eight distinct stations lacking any connecting edges within the generated graph. Addressing this data gap is essential for creating a truly comprehensive and representative public transport network analysis.

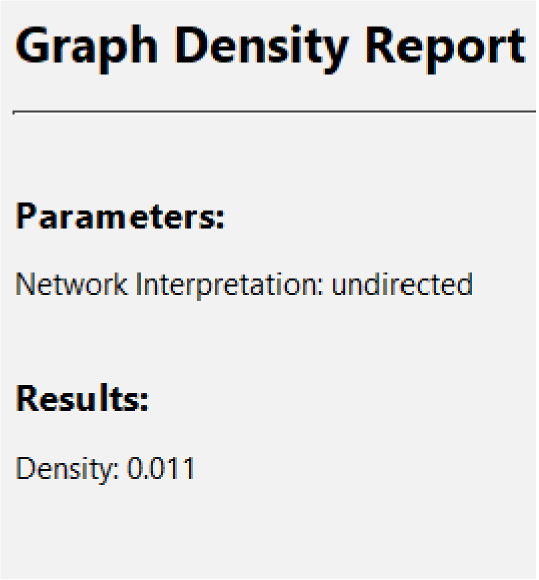


Label	Degree
Charpennes Charles Hernu	5
Thiers - Lafayette	5
Place Jean Jaurès	4
Perrache	4
Debourg	4
Bellecour	4
Desgenettes	4
Porte des Alpes	4
Cordeliers	4
Grange Blanche	4
Halle Tony Garnier	4
Vinatier	4
Reconnaissance - Balzac	4
Saxe - Gambetta	4
Guillotière Gabriel Péri	4
Liberté	4
Gare Part-Dieu V.Merle	4

Figure 3: Graphical Visualization and Node Connectivity Degree

Conversely, the stations exhibiting the highest connectivity in our current model are **Charpennes Charles Hernu** and **Thiers - Lafayette**, both displaying a **degree of 5 ,** they are an important **transfer hub** where passengers can switch modes or lines easily.

b. graph density



Graph Density Report

Parameters:

Network Interpretation: undirected

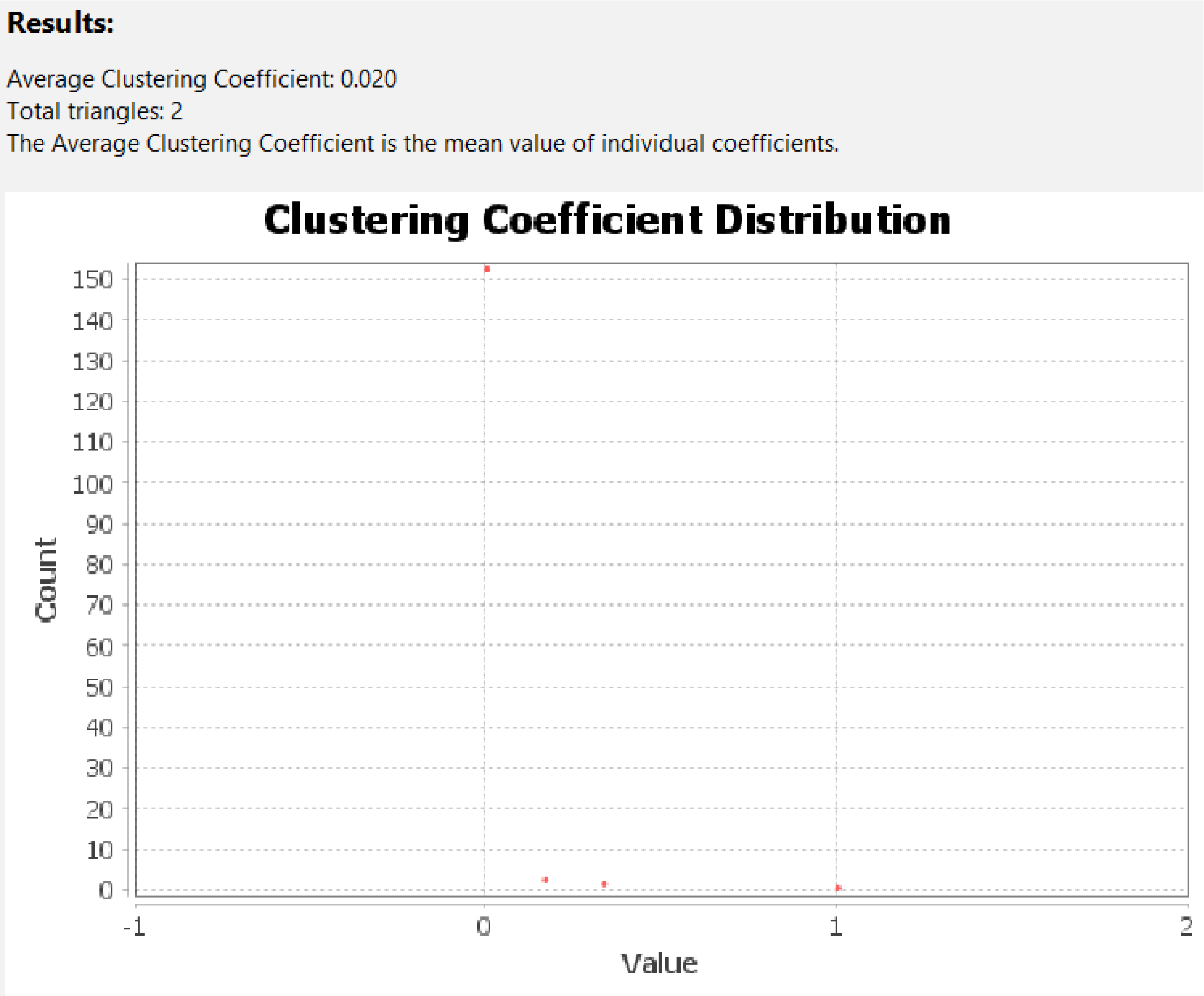
Results:

Density: 0.011

Figure 4 :Graph Density

The extremely low-density value highlights the structural simplicity of the current metro/tram sub-network. Based on the analysis of these data points, the low density confirms a minimal level of direct redundancy within the current sub-network. This suggests that service disruptions would have an exaggerated impact on passenger journeys if no other transport modes were available, confirming that the scope of the current analysis is insufficient for comprehensive resilience assessments across the entire TCL system.

c. coefficient degree



Results:

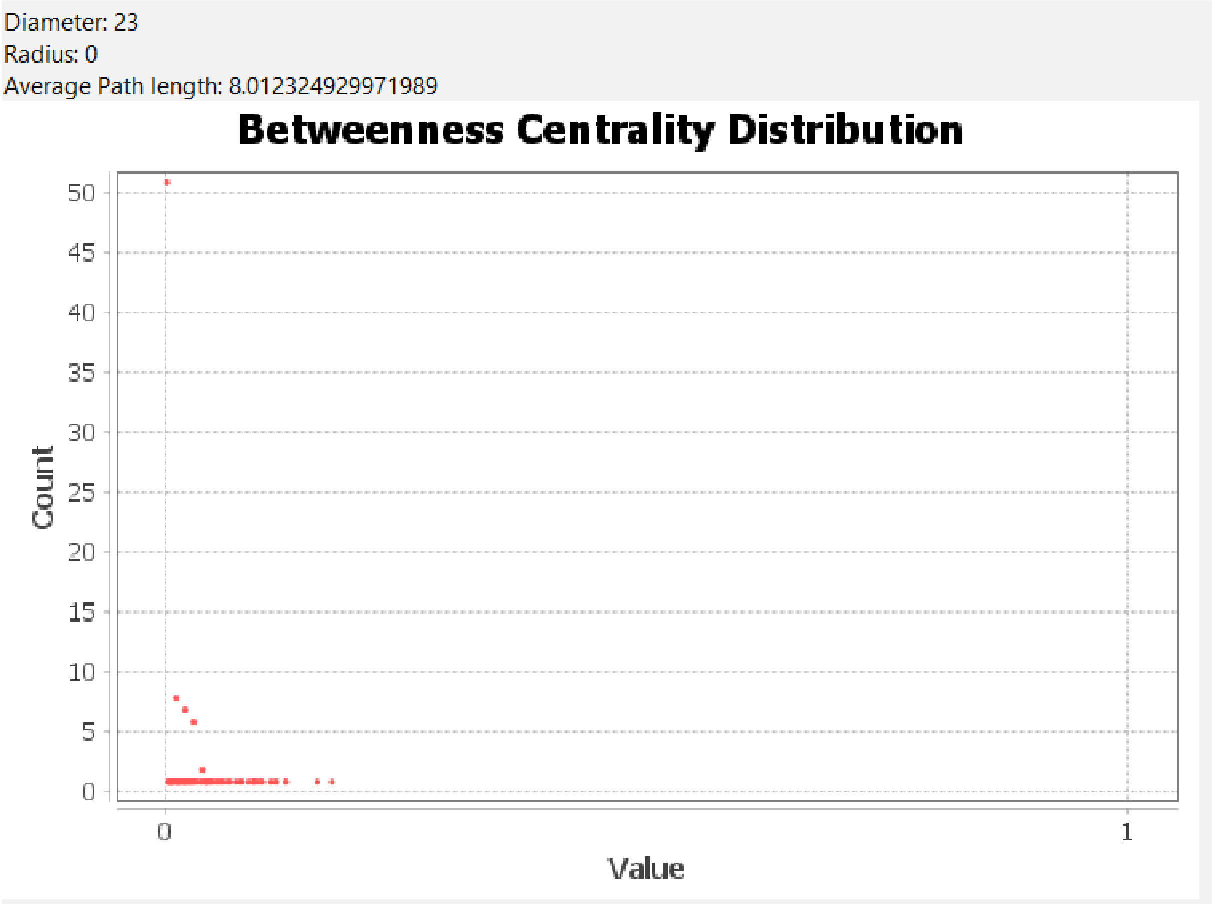
Average Clustering Coefficient: 0.020
Total triangles: 2
The Average Clustering Coefficient is the mean value of individual coefficients.

Figure 5: Clustering Coefficient Distribution

The **Average Clustering Coefficient is extremely low at 0.020**, with only 2 total triangles identified across the entire graph. This result is **structurally expected** for a metro and tram network, as these systems are fundamentally designed as linear or sequential routes (chains) rather than densely interconnected web-like structures (triangles).

When relying strictly on the metro and tram sub-network for passenger journeys, the low clustering coefficient confirms there is **minimal local route redundancy**. This means passengers must follow the strict linear sequence of the line or transfer via a designated hub. If a single segment fails, passengers cannot typically reroute using a nearby alternative segment, forcing them onto the missing bus network or relying heavily on the high-degree hubs (Charpennes, Thiers-Lafayette) to bypass the issue.

d. Betweenness Centrality



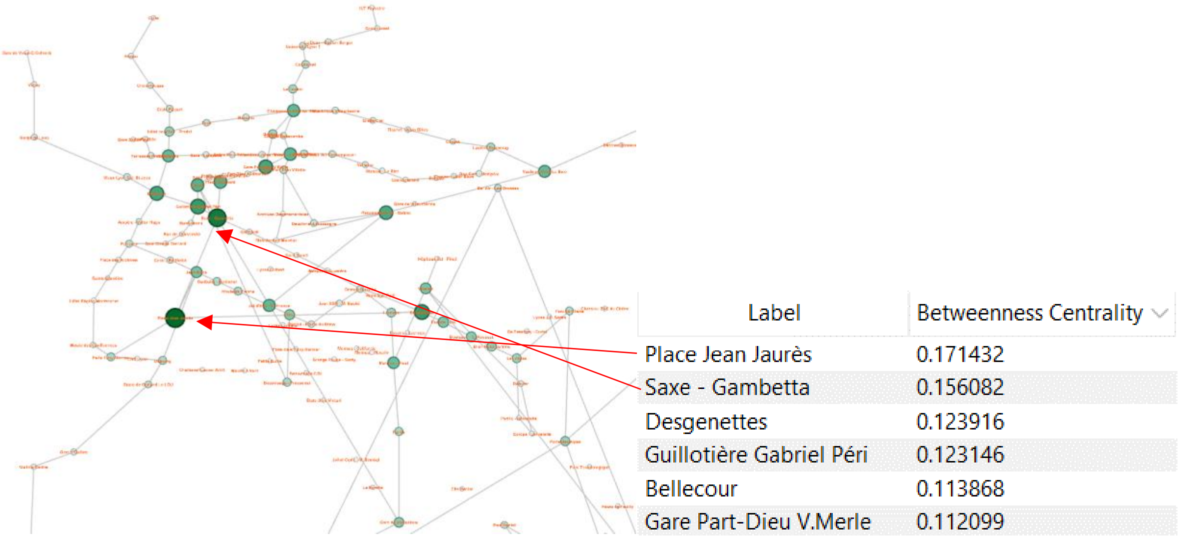
```
Diameter: 23
Radius: 0
Average Path length: 8.012324929971989
```

Figure 6: Betweenness Centrality Distribution

Dominance at Zero: The vast majority of stations (more than 50) have a betweenness value close to 0 (as shown by the heavy concentration of points at the far left of the graph). This is expected in a linear network (chains) where most stations only serve a single line segment and are not transit points for long trips.

Rare Critical Points: The distribution extends to the right, but with very few points. This means that only a few stations hold significant betweenness value. These rare high-betweenness stations are the essential stations: removing or disrupting these nodes would isolate large sections of the network.

Rares Points Critiques : La distribution s'étend vers la droite, mais avec très peu de points. Cela signifie que seules quelques stations détiennent une valeur d'intermédiarité significative. Ces rares stations à haute intermédiarité sont les **stations indispensables** : la suppression ou la perturbation de ces nœuds isolerait de vastes sections du réseau.



Label	Betweenness Centrality
Place Jean Jaurès	0.171432
Saxe - Gambetta	0.156082
Desgenettes	0.123916
Guillotière Gabriel Péri	0.123146
Bellecour	0.113868
Gare Part-Dieu V.Merle	0.112099

Figure 7: Betweenness Centrality

The stations Place Jean Jaurès (0.171432) and Saxe - Gambetta (0.156082) are the most critical connecting hubs in the network. A significant number of shortest paths rely on passing through these two points.

The high centrality of these nodes suggests they are points of high traffic concentration and represent potential single points of failure. Disruption at these stations would cause the most significant delays and rerouting across the entire network.

For network planning, these hubs are strategically vital for transfer and access. Any investment in capacity, maintenance, or security should prioritize these high-centrality locations.

e. Clossness Centrality and Harmonic Closeness Centrality

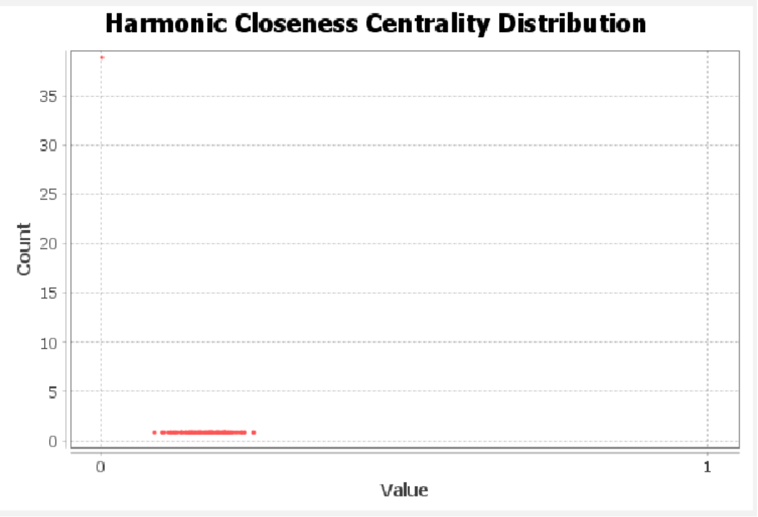


Figure 8: Harmonic Closeness Centrality Distribution

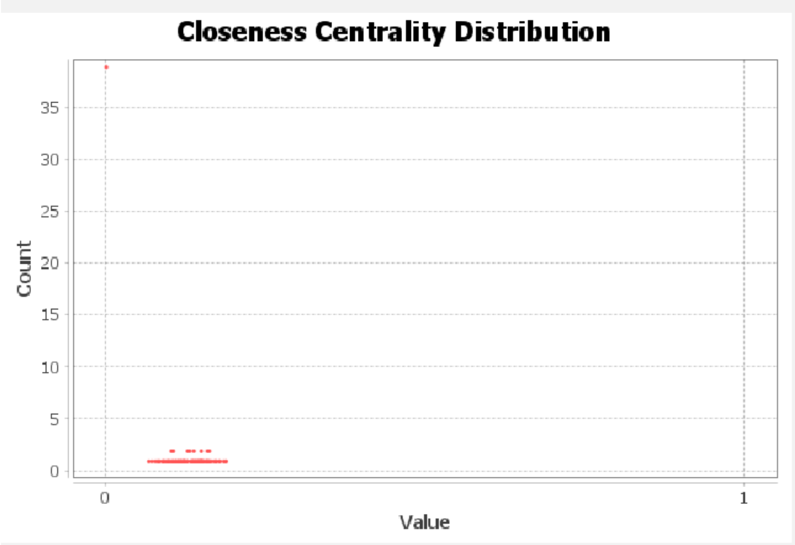


Figure 9: Closeness Centrality Distribution

Running the standard Closeness Centrality revealed a massive spike of approximately 37-38 nodes at the value 0. This result indicates that the graph is not connected, with a large portion of nodes unable to reach the main component of the network.

Consequently, using Harmonic Closeness Centrality is methodologically required, as it is designed to handle disconnected graphs. The harmonic centrality distribution confirmed an identical spike at 0, proving that these ~38 nodes are completely isolated.

The observed fragmentation (38 isolated nodes) is not an anomaly, but a direct consequence of the scope of our dataset, which is limited to metro and tram lines.

Our hypothesis is that these isolated nodes represent stations or stops that, in reality, are served exclusively by other modes of transport (for example, bus lines or regional trains) that were not included in this analysis. The complete multimodal network would very likely connect these data "islands," radically changing the overall accessibility structure.

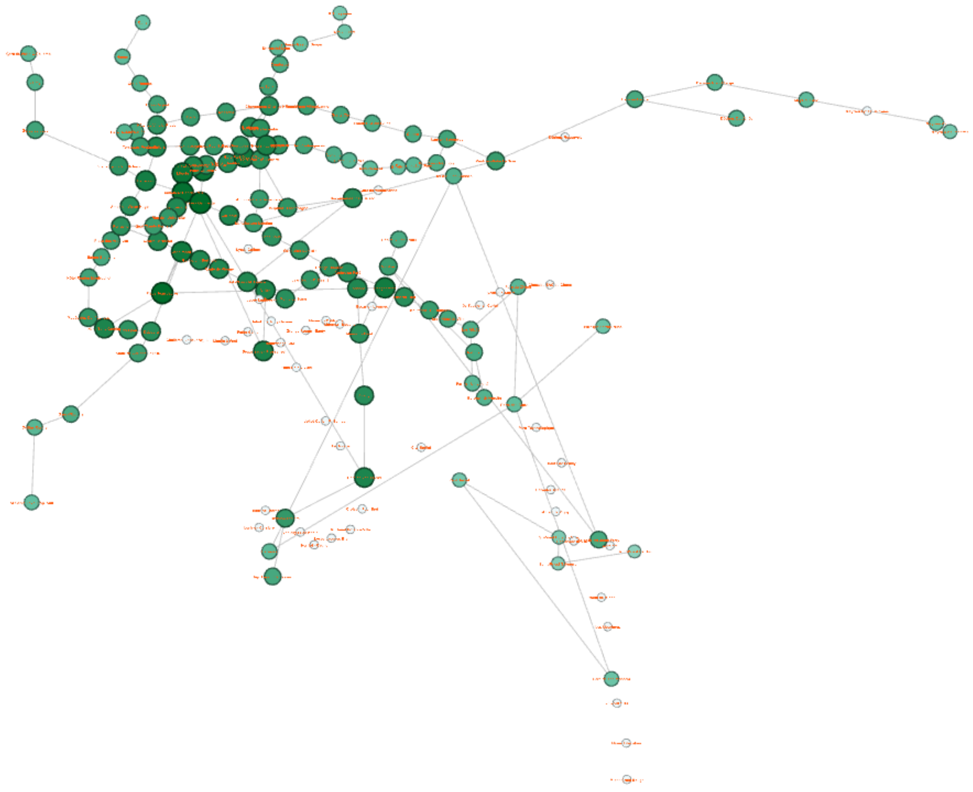


Figure 10: Graph Presentation based on Closeness Centrality

Focusing on the main component (the visible connected network), the visualization is relevant. The large, dark-colored nodes (mainly in the central-left cluster) represent the most accessible stations. These function as the efficient core of the network, offering the fastest average travel times to all other connected stations. Conversely, the smaller, lighter nodes on the peripheral branches are the least accessible.

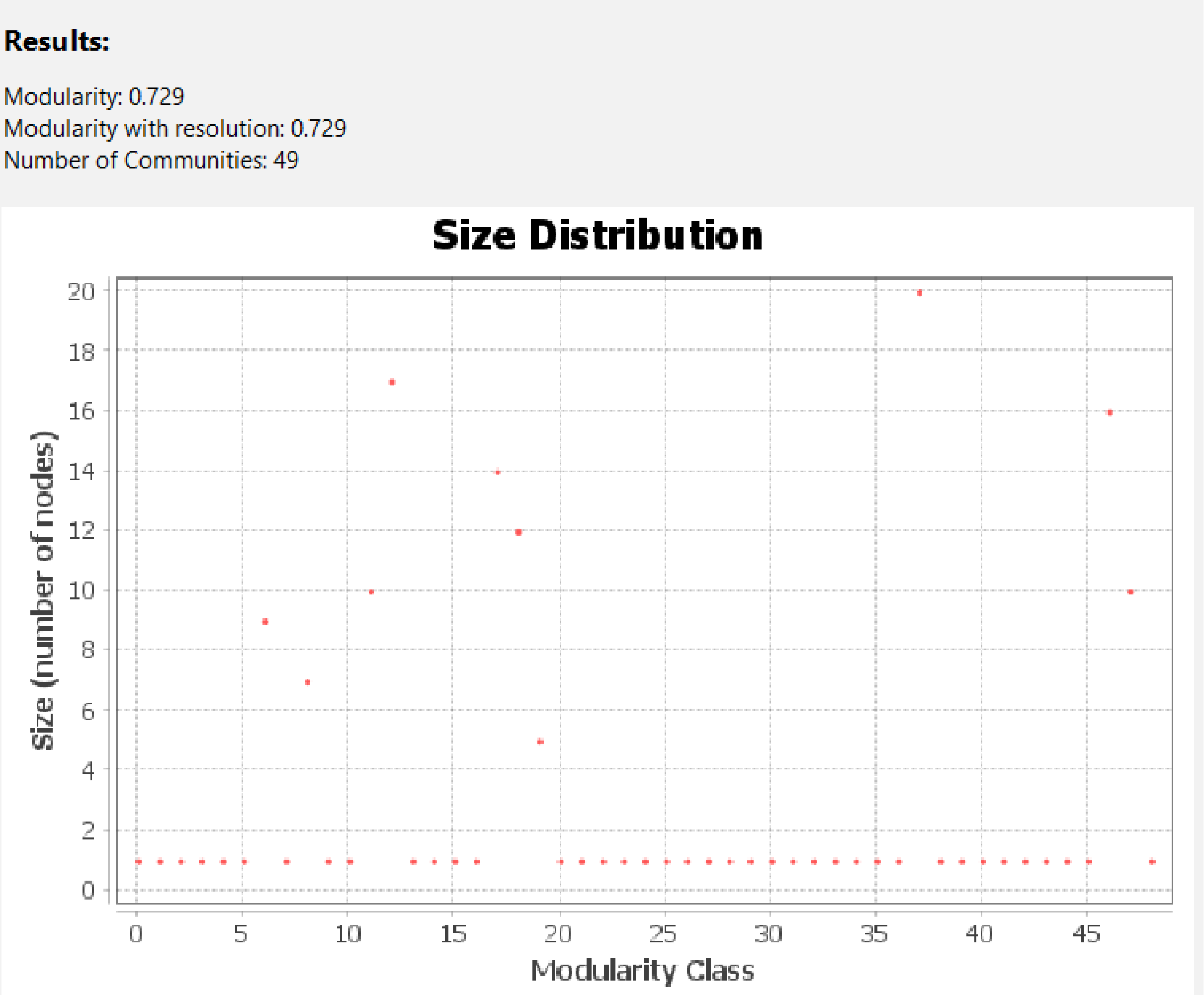
f. Small World

Given that the average shortest-path distance of the network is **8.012** while **log (number of nodes) = 2.20**, the graph does **not** exhibit the Small-World property.

The absence of the Small-World property confirms that the metro/tram sub-network is structurally rigid. This implies low resilience: a failure on a single segment can force major detours and significantly increase path lengths, due to the lack of local alternative connections.

To ensure high availability of service and maintain user satisfaction, the integration of the bus network is essential. Buses provide local redundancy and help preserve efficient average travel times across the full TCL transportation system.

g. Modularity



Results:

Modularity: 0.729
Modularity with resolution: 0.729
Number of Communities: 49

Figure 11: Modularity

The calculated **high Modularity score of 0.729** indicates that the metro/tram sub-network is **highly partitioned into distinct, rigid communities**, with the analysis identifying **49 separate groups**. This fragmentation is visibly reinforced by the presence of numerous single-node communities, which include the 38 isolated vertices. This structural characteristic confirms the network's **non-Small World status** and its low clustering coefficient, as connectivity is overwhelmingly favored *within* each line (community) rather than *between* lines. For the TCL project, this result highlights the **critical vulnerability** of the system: the integrity of the entire network relies heavily on a very small number of transfer hubs (like Charpennes and Thiers-Lafayette) to bridge these nearly independent communities, making the network prone to systemic failures when these few inter-line connections are disrupted.

4. Conclusion

This project successfully modeled the TCL metro and tram network using graph theory, demonstrating how a unified representation of stops can support efficient route planning, transfer detection, and structural analysis. Treating all stops that share the same physical location as single nodes proved to be an intuitive and realistic approach, closely matching how passengers navigate the system. The resulting graph offers a clear view of how stations connect, how routes overlap, and where the network flows most effectively.

The analysis shows that while the metro and tram network operates efficiently within each individual line, it remains weakly connected on a global scale and depends heavily on only a small number of transfer stations. Because there are limited alternative paths between lines, the system is structurally rigid and becomes vulnerable to disruptions or maintenance events. This study therefore highlights the importance of integrating additional transportation modes (buses…) to improve connectivity, reduce fragmentation, and increase resilience for passengers. Overall, the graph-based approach provided valuable insight into the strengths and weaknesses of the current network and offers a solid foundation for future multimodal expansion.