Analysis of Trenitalia's railway network

Project Report for Social Network Analysis

Francesco Vece, Computer Science, 0001100490
Riccardo Spini, Computer Science, 0001084256

Contents

1 Introduction

Trenitalia [1] is the main railway operator in Italy and manages one of the largest railway networks in Europe. This network offers a wide range of services including high-speed trains, intercity trains, and regional trains, covering the entire national territory from the Alps to Sicily. Thanks to this extensive coverage, Trenitalia can effectively connect the main Italian cities, medium-sized urban centers, and rural areas, providing an essential service for millions of commuters, travelers, and tourists.

The size and extent of Trenitalia's railway network represent both an invaluable resource and a continuous challenge. Managing such a vast network requires constant monitoring and in-depth analysis to ensure that the service remains efficient, reliable, and capable of meeting the ever-evolving needs of its users.

2 Problem and Motivation

A detailed analysis of the railway network is essential for:

- **Understanding the current state of the network**: Assessing operational efficiency, identifying critical areas, and analyzing the overall performance of the system.
- **Identifying areas for improvement**: Determining where infrastructure changes, operational management improvements, and service optimizations are needed.
- **Planning for the future**: Developing strategies for the expansion and modernization of the network, considering future needs in terms of sustainable mobility and economic development.

Trenitalia's ability to provide diversified services on such an extensive network also involves the need to address complex issues, such as congestion at some main *hubs*, the obsolescence of certain infrastructures, and the need to improve train punctuality and reliability. Furthermore, the geographical and demographic diversity of Italy means that solutions must be tailored to meet the specific needs of different regions.

A detailed and systematic network analysis not only provides an accurate snapshot of the current state of the railway network but also offers valuable insights on how to improve the service. Through this analysis, it is possible to:

- **Optimize connectivity**: Improve connections between different areas of the country, reducing travel times and increasing the frequency and reliability of services.
- **Increase operational efficiency**: Identify and resolve bottlenecks, improve traffic management, and reduce delays.
- **Promote sustainability**: Increase the use of public transport and reduce environmental impact, contributing to national and European sustainability goals.

In summary, analyzing Trenitalia's railway network is crucial to ensure that this critical resource continues to meet the country's needs efficiently and sustainably, both today and in the future.

[1] `https://www.trenitalia.com/it.html`

3 Datasets

To create the dataset, we used the **Trenitalia application API** [2] to collect information on trains running for a week (27/05/2024 to 02/06/2024) on the Italian railway network. This data collection was subsequently compared with the weekly schedule publicly available on the Trenitalia portal [3] to ensure the consistency and accuracy of the obtained data. The collected train data was stored in a `csv` file containing information on the routes of various national trains.

The structure of the `csv` file includes the following columns:

Name	**Description**
categoria	The category of the train (e.g., regional, intercity, high-speed).
stazPart	Departure station of the train.
stazArr	Arrival station of the train.
Regione_part	Region of departure of the train.
Provincia_part	Province of departure of the train.
Comune_part	Municipality of departure of the train.
Longitudine_part	Longitude of the departure station.
Latitudine_part	Latitude of the departure station.
Regione_arr	Region of arrival of the train.
Provincia_arr	Province of arrival of the train.
Comune_arr	Municipality of arrival of the train.
Longitudine_arr	Longitude of the arrival station.
Latitudine_arr	Latitude of the arrival station.

For the cartographic representation of the various cities included in our analysis, we used the geographical coordinates provided by **Garda Informatica** [4]. These coordinates, organized in a `csv` file, allowed us to accurately visualize the Italian railway network and effectively represent the connections between cities in Section 6.

The `csv` file of geographical coordinates includes the following information:

[2] `https://github.com/TrinTragula/api-trenitalia`

[3] `https://www.trenitalia.com/it/informazioni/orari_regionali_digitali.html`

[4] `https://www.gardainformatica.it/`

Name	Description
SIGLA_PROVINCIA	Province code (e.g., RM for Rome).
CODICE_ISTAT	ISTAT code of the municipality.
DENOMINAZIONE_ITA_ALTRA	Another name of the municipality in Italian, if exists.
DENOMINAZIONE_ITA	Official name of the municipality in Italian.
DENOMINAZIONE_ALTRA	Name of the municipality in a language other than Italian, if exists.
FLAG_CAPOLUOGO	Indicates whether the municipality is a provincial capital (yes/no).
CODICE_BELFIORE	Belfiore code of the municipality.
LAT	Latitude of the municipality.
LON	Longitude of the municipality.
SUPERFICIE_KMQ	Surface area of the municipality in square kilometers.
CODICE_SOVRACOMUNALE	Supramunicipal code of the municipality, if applicable.

This data allowed us to create detailed maps showing the exact location of each city and the railway connections between them.

3.1 Data Preprocessing

We chose to focus on cities rather than individual railway stations (for example, Rome has two main stations, Roma Termini and Roma Tiburtina). This methodological choice was driven by the interest in examining national connections between different Italian cities, rather than the specific details of internal connections within individual cities. Therefore, in our analysis, we considered each city as a single node, ignoring the presence of multiple stations within it. This approach allowed us to obtain a clearer and more concise view of the national railway connections, facilitating the identification of the main routes and relationships between different cities. Additionally, it simplified the network analysis, making it easier to compare the various routes.

We then performed preprocessing on the datasets to obtain the following two representations:

Stazione Partenza	Stazione Arrivo

Table 1: Dataset containing train journeys (treni_solo_città.csv).

Città	Longitudine	Latitudine

Table 2: Dataset containing cities and coordinates (merged_data.csv).

3.2 Building the Networks

The constructed network is based on a **monopartite graph with undirected and unweighted edges**, where the nodes represent the cities in the railway network and the edges represent the connections between the various cities linked by direct trains. In this representation, each node of the graph corresponds to a specific city within the railway system, and each edge indicates

a railway route that allows direct travel between two cities. This modeling using a monopartite graph allows for a clear and intuitive visualization and analysis of the railway network. Thanks to this structure, it is possible to easily identify which cities are directly connected to each other, determine the number of existing connections for each city, and identify any alternative routes to reach a destination.

In the graph, we decided to highlight the cities with the highest degree of centrality by making the nodes red.

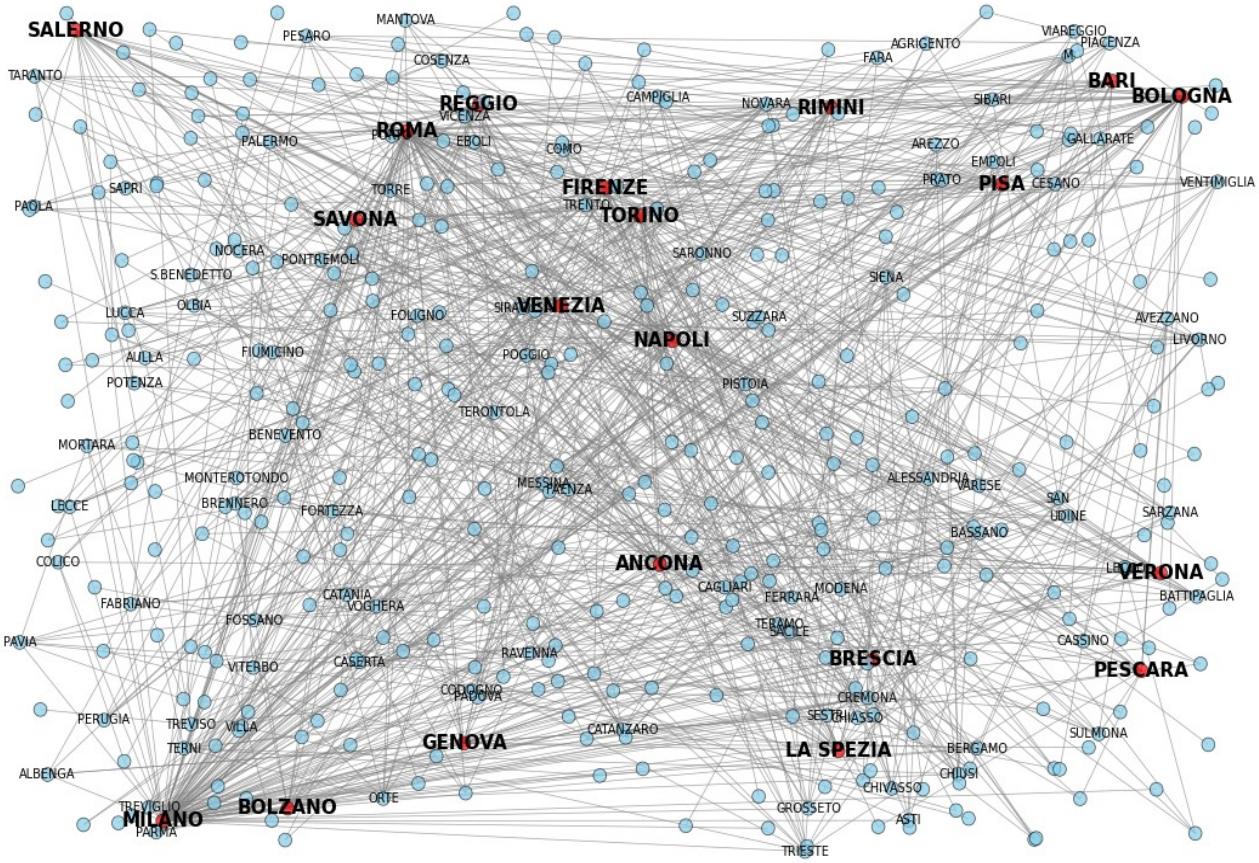


Figure 1: Graph with cities and scheduled journeys.

3.3 Python Script

To enable the re-execution of the code and thus the analysis, the Python script and the various datasets used can be found at the GitHub link.

4 Validity and Reliability

The validity of the research should be of a good level. The schedules and cities exported from the Trenitalia portal ensure the reliability of the information related to the trains. Trenitalia is one of the main Italian railway companies and provides accurate and up-to-date data on its services. By using this data, we can be confident that the connections between the cities

represented in the graph faithfully reflect the actual railway network. The schedules and train routes are regularly updated on the Trenitalia portal, thus ensuring that our data is always current and precise.

Similarly, Garda Informatica provides comprehensive geographical data that is meticulously verified and maintained. With continuous updates and validation processes, the dataset offers highly precise and detailed geographic information. This accuracy allows us to accurately position cities in our graph, thereby enhancing the reliability of our model.

Re-proposal of the Analysis

The use of tools like `Pandas` for data management, `NetworkX` for graph creation and analysis, and `Matplotlib` for visualization makes the entire analysis process repeatable and transparent. These tools are widely used in both scientific and industrial communities for their reliability and flexibility.

- **Pandas**[5]**:** Provides powerful data structures and analysis tools, allowing easy data manipulation and cleaning. The code used to load and preprocess data can be easily reused or adapted for future datasets, ensuring analysis repeatability.
- **NetworkX**[6]**:** Offers a wide range of algorithms for graph construction, manipulation, and analysis. The creation of the railway graph and the calculation of centrality metrics can be faithfully reproduced with any updates to the data, ensuring consistency in results.
- **Matplotlib**[7]**:** Enables clear and customizable data visualization, facilitating the presentation of results.
- **Folium**[8]**:** Is a Python library used to create interactive maps that supports loading and displaying GeoJson data, which are common formats for geographical data.

Reliability of the Procedure

The procedure described not only ensures the accuracy of the data but also allows for easy updating and rerunning of the analysis. Whenever new data is provided by Trenitalia are available, the entire analysis can be rerun with a few simple steps. This enables keeping the model up-to-date and adaptable to changes in the railway network or geographical information.

In summary, the combination of reliable data sources and robust analysis tools ensures that our research is valid, accurate, and repeatable. The methodology adopted can be easily replicated, ensuring that the results obtained are consistent and verifiable over time.

[5] `https://pandas.pydata.org/docs/`
[6] `https://networkx.org/documentation/stable/reference/index.html`
[7] `https://matplotlib.org/`
[8] `https://python-visualization.github.io/folium/latest/`

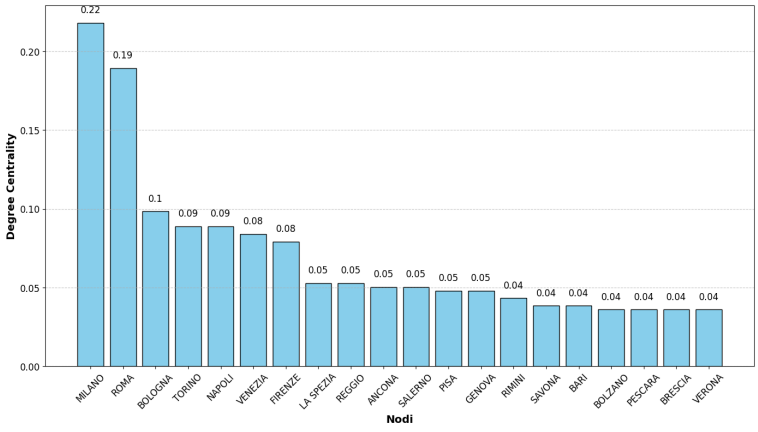
5 Measures and Results

In the following section we sum up and explain the results of all the measures we applied in this study.

5.1 Centrality

5.1.1 Degree Centrality

In the case of Trenitalia's railway network, **Degree Centrality** can help identify cities that have the highest number of direct connections with other stations. This can indicate the most important cities in terms of railway traffic and their relevance in the national network.



Working on an undirected graph, we used the following formula to calculate degree centrality:

$$C_D(v) = \frac{d(v)}{N-1} \tag{1}$$

Where:

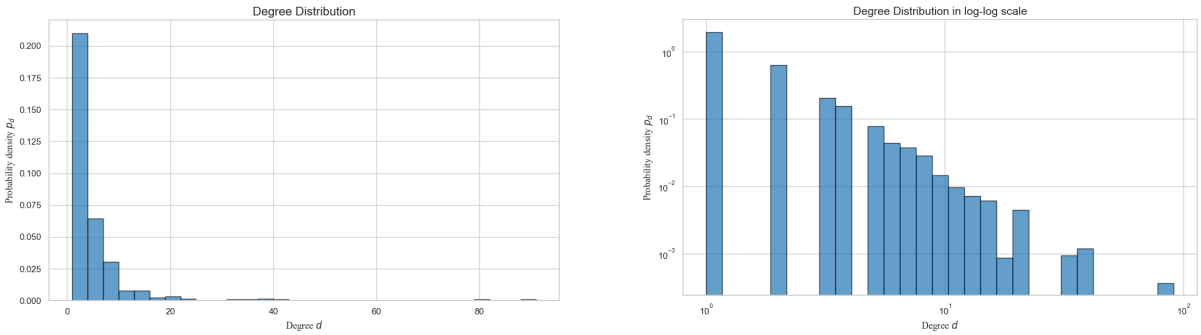
- $C_D(v)$ is the degree centrality of node v.
- $d(v)$ is the degree of node v (number of edges connected to v).

5.1.2 Degree Distribution

We analyzed the distribution of nodes in our dataset, revealing that it follows a heavy tail pattern. This means there are some nodes with a high number of connections, called *hubs*.
Heavy tail distributions have the property of being *scale-free*, which means that some characteristics of the network are independent of the observation scale. Here are some features of *scale-free* networks:

- **Degree distribution**: The number of connections (degree) of the nodes in the network follows a **power law**. This implies that there are few highly connected nodes (*hubs*) and many nodes with few connections.

- **Resilience to errors**: *Scale-free* networks are generally more resilient to random errors compared to other networks. The removal of random nodes does not significantly interrupt the network, while the removal of *hub* nodes can have a dramatic impact. This behaviour is observable in Section 6 with the removal of random nodes.



(a) Linear distribution of degrees.

(b) Distribution log-log scale.

Figure 2: Degree distribution in two different scale.

Distribution results

We then applied the power law formula $p_d = Cd^{-\alpha}$ where:

$$P_d = C \sum_{d'=d}^{\infty} d'^{-\alpha} \approx C \int_d^{\infty} d'^{-\alpha}\, \partial d' = \frac{C}{\alpha - 1} d^{-(\alpha-1)} \tag{2}$$

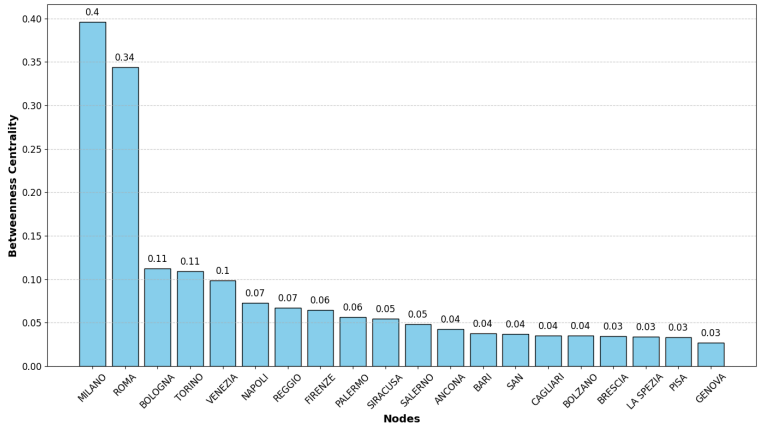
- $\boldsymbol{\alpha = 2.9728548470}$: this value represents the exponent of the power law. An α of around 3 indicates that the degree distribution in the graph follows a power law with a fairly heavy tail, suggesting the presence of some nodes with very high degrees (*hubs*) compared to the general degree distribution.

- $\boldsymbol{p - value = 0.0601123750}$: the *p-value* associated with the power law measures how well the data fit the proposed power law. A *p-value* of around 0.06 suggests a good fit of the data to the power law, although not perfect.

These results confirm that the analyzed network exhibits characteristics typical of *scale-free* networks, with significant implications for its structure and resilience.

5.1.3 Betweenness Centrality

Betweenness Centrality can reveal which cities or stations are crucial for train transit between different regions or for national connectivity. It identifies the nodes that, if removed, could significantly impact the overall connectivity of the network.



We used the following formula to calculate it:

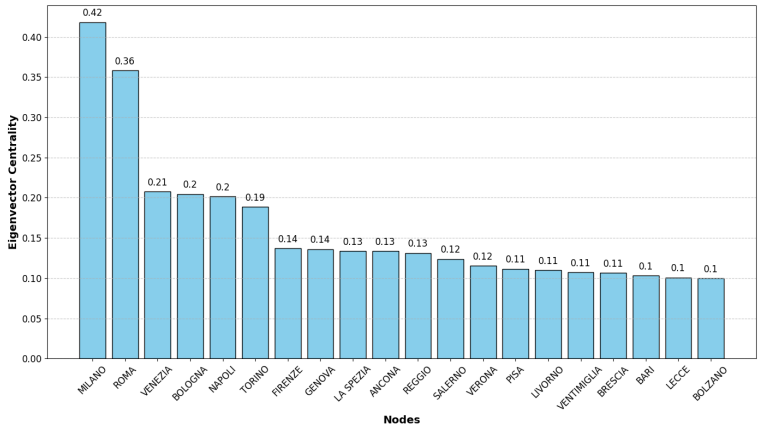
$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{3}$$

Where:

- $C_B(v)$ is the betweenness centrality of node v.
- σ_{st} is the total number of shortest paths from node s to node t.
- $\sigma_{st}(v)$ is the number of those shortest paths that pass through node v.

5.1.4 Eigenvector Centrality

Eigenvector Centrality can help identify cities (or stations) that are central not only in terms of the number of connections but also in terms of the quality of the connections they provide. This measure can highlight cities that serve as *hubs* for the most important and strategic railway routes.



We used the following formula to calculate it:

$$x_i = \frac{1}{\lambda} \sum_{j \in \mathcal{N}(i)} A_{ij} x_j$$

Where:

- x_i is the eigenvector centrality of node i.
- λ is the largest eigenvalue associated with the adjacency matrix A.
- $\mathcal{N}(i)$ is the set of neighbors of node i.
- A_{ij} is the element of the adjacency matrix representing the connection between nodes i and j.

5.2 Density

The **density** indicates the completeness or saturation of the network in terms of connections between cities. A high density may indicate a well-developed and interconnected network, while a low density could highlight areas of the network that are less developed or less accessible.

Misuration	Value
Density	0.01084300024

The formula used to calculate density is:

$$D = \frac{2E}{N(N-1)} \tag{4}$$

Where:

- D is the density of the graph.
- E is the number of edges in the graph.
- N is the number of nodes in the graph.

We obtain a density that is not particularly high, which is justified by the fact that in a railway network, it is not common for all nodes to be directly connected. Instead, expecting that nodes are reachable through paths that pass through other intermediate nodes is more realistic.

5.3 Connected Component

The **connected components** help identify groups of cities (or stations) that are interconnected with each other. This measure is useful for understanding whether the network forms a single large component or if there are isolated groups of stations that may require greater integration.

Misuration	Value
Connected Component	1

We find a single connected component, despite being islands, because **Trenitalia offers trips that include ferries** (for connections from Sardinia to the mainland) **or ferries on which locomotives can be embarked**(crossing the Strait of Messina).

The result of this measurement, although it may seem obvious, is crucial to ensure that any point in the network can be reached from another point via a more or less long path.

5.4 Degree Assortativity Coefficient

The **degree assortativity coefficient** can reveal whether stations with a high number of connections tend to connect to each other or if there are specific preferences in the formation of railway routes. This information is useful for understanding the structure of relationships between cities and for optimizing existing connections.

Misuration	Value
Degree Assortativity Coefficient	- 0.138304541791

The formula used to calculate the degree assortativity coefficient is:

$$r = \frac{\sum_{ij}(ij - q_i q_j)}{\sigma^2} \tag{5}$$

where:

- i and j are the degrees of nodes connected by an edge in the network.
- q_i is the fraction of edges that connect to nodes of degree i.
- σ^2 is the variance of the degree distribution of nodes.

This result tells us that stations with many connections (high-degree nodes) tend to connect with stations with few connections (low-degree nodes) and vice versa. This reflects a typical structure of transportation networks, where major *hubs* (large stations) are connected to many peripheral stations.

5.5 Bridges

Identifying **bridges** in the Trenitalia railway network is crucial for understanding key connections between different regions or groups of stations. This information helps identify critical points in the network where maintenance or expansion could have a significant impact on overall connectivity and the efficiency of the railway system.

Misuration	Value
Bridges	140

The number of bridges (also identified on GitHub) suggests that certain stations certainly have a more impactful role on the entire network, as we will further explore in the simulation in Section 6.

5.6 Modularity and Community Detenction

Modularity can help identify natural groups of stations or cities that form well-defined clusters or communities within the network. This information is useful for understanding how different parts of the network are organized and for identifying areas that may benefit from specific improvements or require greater integration.

Misuration	Value
Modularity	0.6542929929173315

Modularity Q is defined as:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{6}$$

where:

- A_{ij} is the element of the adjacency matrix between the nodes i and j.
- k_i and k_j are the node degrees of i and j.
- m is the number of the edges in the graph.
- c_i is the community to which node i belongs.
- $\delta(c_i, c_j)$ is a Kronecker delta function that equals 1 if $c_i = c_j$, and 0 otherwise.

Having a relatively high modularity value allows us to understand how within the entire network, there are more tightly connected communities with a higher internal density. To identify them, we applied community detection (also identified on GitHub). This allowed us to observe how the identified communities tend to revolve around larger cities or those with greater centrality in the network.

For example, Milan has been identified as part of sub-community 0, with cities like Monza, Lodi, and Varese. This illustrates how communities also have a geographical context in our case, nearby cities are grouped into common sub-communities.

6 Simulations

To better understand the robustness and resilience of the Trenitalia railway network, we conducted a **simulation** where **5 random cities are removed from the network**. This simulation allows us to assess the network's ability to endure malfunctions. To ensure statistical robustness, we repeated the simulation **100 times**, calculating the average of relevant measurements.

For a more comprehensive analysis, we also simulated the worst-case scenario: **disabling the 5 nodes with the highest degree centrality** (Milan, Rome, Bologna, Turin, and Naples). Even in this extreme scenario, we calculated and evaluated the relevant measurements, allowing us to fully understand the vulnerabilities and strengths of the network.

For the following representations we use the python library `folium`:

(a) Connection **before** simulation.

(b) Connection **after** simulation.

Figure 3: Comparison of connections before simulation and after in the **worst case**.

Results of simulations

Graph	Density	Connected Component
Default	0.0108430002	1
Random Cities	0.0095101231	1.16
Worst Case	0.0078516185	49

As we can see from the final results, **the Trenitalia railway network demonstrates a remarkable ability to handle the absence of randomly selected cities**, maintaining a good average density and a number of connected components close to default network values.

In the **worst-case scenario**, however, we observe a significant increase in the number of connected components, confirming **the crucial role of cities with the highest centrality as connection points in our network**. The removal of these highly central cities leads to significant fragmentation of the network, creating numerous isolated subgraphs.

This fragmentation effect also explains the significantly reduced density: with fewer direct connections between the remaining nodes and a more disconnected network, the total number of edges decreases compared to the maximum possible number of connections, in this way reducing the overall density of the network.

These results underline the strategic importance of cities with high centrality in maintaining the integrity and connectivity of the railway network.

7 Conclusions

This study clearly highlights how the Trenitalia railway system is extremely extensive and, at the same time, well connected. As expected, within this network, larger cities play a crucial role as principal *hubs* that facilitate connections between various communities. These major urban centers act as distribution and transit points, enabling a continuous flow of railway traffic.

Furthermore, our analysis has shown that the Trenitalia railway network can be characterized as a *scale-free* network. This type of structure is known for its robustness and resilience, as most connections are concentrated on a relatively small number of major *hubs*, while peripheral nodes are less connected. This configuration gives the network significant capacity to resist localized issues, such as maintenance interruptions, without compromising the overall operation of the network.

This resilience feature was further demonstrated through simulations in the random case in Section 6, confirming the network's ability to maintain high levels of functionality even in the face of partial disruptions. The simulations allowed us to observe how, despite the temporary removal of certain connections or nodes, the system effectively reorganizes itself. This not only attests to the efficiency of the current railway network design but also its adaptability to emergency situations or maintenance, ensuring operational continuity.

8 Critique

A relevant criticism of this analysis concerns the dataset of journeys used. Currently, the analysis focuses exclusively on the starting and destination cities, leave out the intermediate cities and train stops. Including these cities could provide a more comprehensive view of the interconnections within the railway system.

Using the API to obtain data on every train stop presents significant challenges, including managing large volumes of data and the need for accurate mapping of railway routes. Additionally, manually digitizing information for each stop would be an extremely expensive task.

Therefore, a future improvement in the analysis could involve developing more advanced automated tools for data acquisition and processing, in way facilitating the inclusion of intermediate cities in the dataset. This could imply adopting **machine learning technologies for automatic data extraction** from available sources and creating a richer and more detailed database.