Network Science
Project Report

Modeling the Indian Railway as a Network: Route Clustering and Zone Analysis

Group 46
Kritika Gupta 2021395 | Prajna Vohra 2021345 | Siddharth Gupta 2021355

Objective

The goal of this project is to study the Indian Railway Network through the lens of network science by modeling stations and train routes as a directed graph. Each station is treated as a node, and every direct train movement between two stations is a directed edge. The aim is to understand the underlying topology, temporal behavior, and zonal clustering of the railway system using network science principles. Finally, we exhibit the small world property of the Indian Railway network using an interactive interface.

Deliverable 1: Network Construction and Structural Analysis

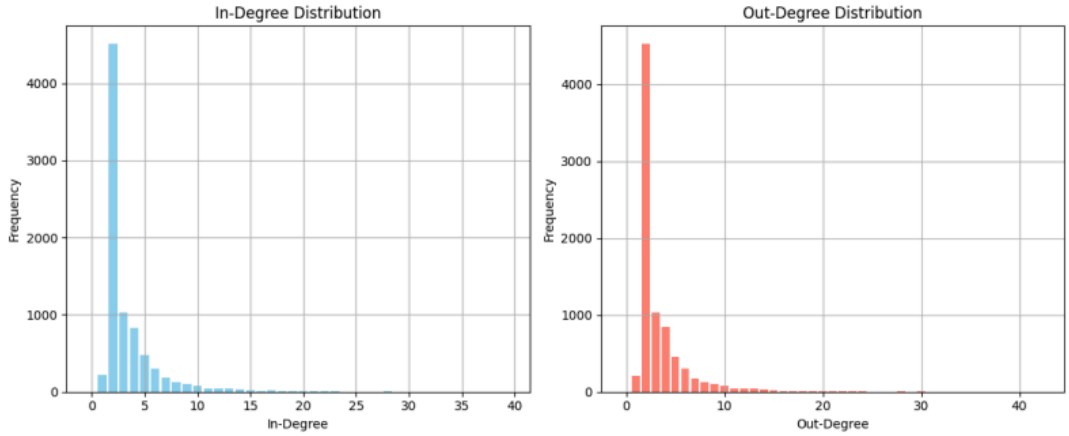
Data Source: https://www.kaggle.com/datasets/arihantjain09/indian-railways-latest

Data Preprocessing:

- Cleaned the Indian Railway schedule and train info datasets.
- Ensured consistency in train numbers and station codes.
- Constructed edge lists using origin-destination pairs based on direct train routes.

Graph Construction:

- Modeled the railway as a directed graph (DiGraph) using NetworkX.
- Each station was a node; an edge from A → B implied a train route from A to B.
- Weight represents the number of trains that directly run between two consecutive stations. So the weight reflects the "traffic frequency" or "route redundancy" between a pair of stations — higher weight means more trains use that segment.
- Network constructed using NetworkX with **8,147 nodes** and **28,179 directed edges.**



Inference from Degree Distributions:

- The in-degree and out-degree distributions are right-skewed, indicating that:
  - Most stations are connected to only a few others.
  - A few stations act as major hubs with high connectivity.
- This suggests the network may exhibit **scale-free properties**, typical of real-world transport systems:
  - **Robust** against random failures.
  - **Vulnerable** to targeted attacks on hub stations.
- The shape of the distribution and presence of hubs also support the **small-world property**, implying:
  - Short paths between most station pairs.
  - Efficient navigability across the network.

Structural Metrics Computed:

- **In-degree and Out-degree Distributions:** Counted how many trains arrive and depart from each station.
- **Average Clustering Coefficient**: Computed as 0.354.
- **Strongly and Weakly Connected Components:** Identified core interconnected subnetworks. Largest WCC Size- 8059 nodes and 27993 edges

Top 10 Stations by Betweenness Centrality — Inference:

Betweenness centrality captures the importance of a station as a bridge within the network — i.e., how often it appears on the shortest paths between other station pairs. Higher values indicate greater influence in ensuring the flow of traffic across regions.

Rank	Station Code	Full Station Name	Betweenness Centrality	Interpretation
1	MGS	Mughalsarai Jn (Pt. Deen Dayal Upadhyay Jn)	0.2754	Most critical node — connects eastern, northern, and central zones.
2	BZA	Vijayawada Jn	0.2214	Key hub in south-east corridor; connects Chennai, Hyderabad, and Kolkata.
3	BPQ	Balharshah Jn	0.1790	Crucial mid-point in central India; links north-south and east-west corridors.
4	BSL	Bhusaval Jn	0.1665	Connects western India to north-central routes; important for west-east travel.
5	CNB	Kanpur Central	0.1566	Prominent junction in north-central India; heavy traffic flow across belts.
6	HBJ	Habibganj (Bhopal)	0.1420	Central node in Madhya Pradesh region.
7	NDLS	New Delhi	0.1406	National capital; a key administrative and interzone transit hub.
8	HWH	Howrah	0.1161	Eastern metropolitan gateway; high inter-zonal transfer activity.
9	STA	Satna	0.1145	Connects central and northern routes; strategic for east-central movement.
10	VZM	Vizianagaram	0.1139	Coastal node bridging Odisha and Andhra Pradesh corridors.

Insight: The top stations by betweenness are **geographically dispersed** and highlight India's **key transit corridors**. MGS and BZA consistently appear as high-traffic connectors, underlining their structural and strategic significance in maintaining national rail connectivity.

Deliverable 2: Temporal Network Slicing and Visualization

Approach:

- Created daily subgraphs for each day of the week (Monday to Sunday).
- Used **Days_of_Operation** data to filter the train schedule accordingly.
- Visualized each daily subgraph to observe how train operations change with the day.

Inferences of Temporal Network Analysis

- **Variation in Network Size across Days**

  The Indian Railway network shows notable variation in scale throughout the week:

Day	Stations (Nodes)	Routes (Edges)	Avg Degree
Friday	6,077	12,828	4.22
Wednesday	6,006	12,705	4.23
Sunday	6,017	12,634	4.20
Saturday	6,029	12,526	4.16
Tuesday	6,072	12,597	4.15
Thursday	5,923	12,120	4.09
Monday	5,905	11,706	3.96

  Insight: The network is most active on **Fridays and mid-week (Wednesday)**, suggesting a higher volume of scheduled services during these periods — likely to

accommodate end-of-week and mid-week passenger demand. **Mondays** are comparatively quieter. The subgraph visualizations for Monday and Friday are included in the code folder.

- **Top Hub Stations by Day**

  Using degree centrality, we identify the top 5 most connected stations for each day:

  Consistently Appearing Hubs:

  1. **BZA (VIJAYAWADA JN)** – appears as a top hub **every day**, indicating its critical role in routing.
  2. **MGS (MUGHAL SARAI), CNB (KANPUR CENTR), LKO (LUCKNOW JN.) –** frequently occur, serving as **major junctions.**
  3. **KGP (KHARAGPUR)** and **KYN (KALYAN JN)** – appear several times, underlining their importance regionally.

  Conclusion: These hubs act as **key connectors** in the railway network, facilitating inter-zone travel and supporting high traffic flow. Their repeated centrality suggests that they are **structurally critical** and potential **single points of failure** in the system.

- **Centrality Dynamics**

  Stations like **VIJAYAWADA JN, KANPUR**, and **LUCKNOW** frequently top the centrality rankings. This implies:

  1. They manage a **large number of direct connections** with other stations.
  2. They likely serve as **transfer points** or **regional anchors**.
  3. Their operational performance can significantly influence network efficiency and flow.

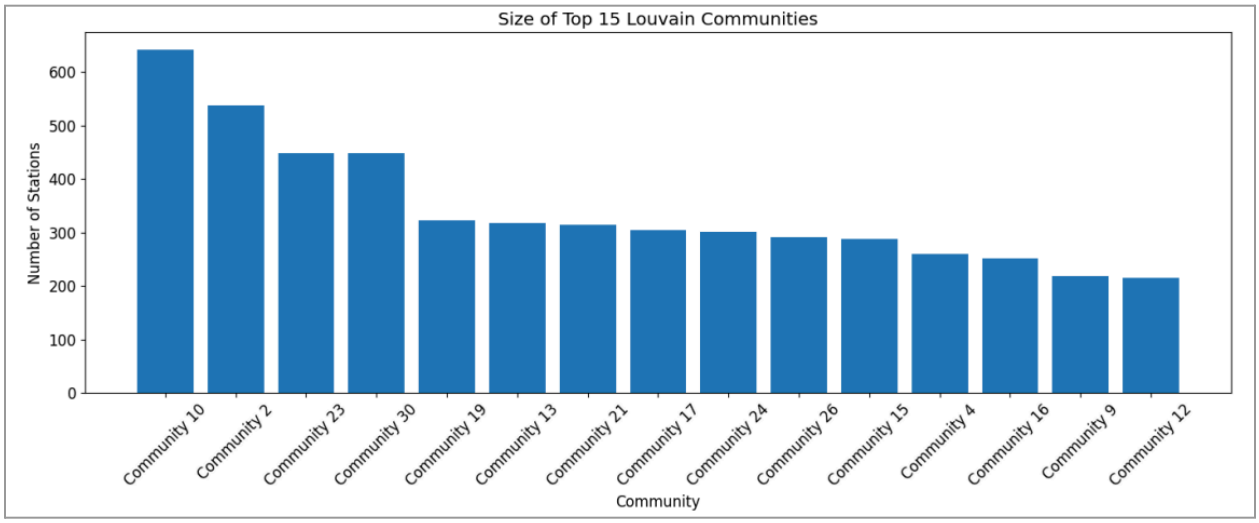
- **Summary**
  - The Indian Railway network exhibits **dynamic structure over time**, with notable differences in activity between weekdays and weekends.
  - **Temporal slicing** reveals patterns in **station importance** and **network load**.
  - This analysis aids in **resource planning**, **robustness evaluation**, and **service optimization**.

Deliverable 3: Community Detection and Functional Zoning

Approach

- Built an undirected version of the railway network to facilitate community detection algorithms with **8147 nodes** and **14461 edges**.
- Applied **Hierarchical Clustering** using network modularity and dendrogram-based grouping. As part of this, explored the **Louvain method** for scalable community detection.
- Also visualized the communities with different colors using matplotlib and networkx.

Functional Zone Analysis Results



- **Community Structure Insights:**
  - **Louvain's algorithm** identified **57 communities**, showing high modularity in the Indian railway network.
  - **Size Distribution**: Uneven; largest community (ID 10) has **642 stations (7.9%)**; top 10 communities cover **~45%** of all stations.
  - **Geographic Clustering**: Communities align well with geography, reflecting evolution based on **regional needs**.
  - **Network Density**: Average degree mostly between **3.1–3.7**; community 10 has the highest (**3.71**) showing dense internal links.
  - **Internal Structure**: Path lengths within communities range **5.6 to 7.8**; clustering coefficients **0.28–0.41**, indicating "small world" traits.

- **Key Functional Zones:**
  - **Zone 10 (642 stations)**: **North Central Region** centered around Lucknow, Varanasi, and Kanpur. This zone covers much of Uttar Pradesh, one of India's most populous states, with high station density. Its relatively low average path length (5.6) indicates efficient connectivity despite its large size. The zone serves as a critical transportation hub for the Gangetic plains.
  - **Zone 2 (538 stations)**: **Central Indian Corridor** spanning from Bhusaval to Nagpur, Itarsi, and Jhansi. This zone functions as a central connecting bridge between north, south, east, and west India. The presence of multiple junction

stations with high betweenness centrality confirms its role as a critical transit corridor.

- **Zone 23 (449 stations)**: **Rajasthan Network** centered around Jaipur, Jodhpur, and Phulera. This zone has distinctive topological characteristics with longer average paths (7.2) and lower clustering (0.28), reflecting the scattered population distribution in this desert region. The network serves as a critical connection between Delhi and western India.
- **Zone 30 (448 stations)**: **Northeast Frontier** organized around New Jalpaiguri, New Bongaigaon, and Katihar. This zone has the highest average path length (7.8), indicating challenging connectivity in this mountainous region. The zone serves as the gateway to Northeast India and parts of eastern Himalayan regions.
- **Zone 19 (322 stations)**: **Western Railway Zone** centered around Ahmedabad, Vadodara, and Surat. This commercially important zone connects major industrial cities in Gujarat and extends to Mumbai. Its relatively high clustering coefficient (0.35) indicates efficient local connectivity supporting industrial and commercial activity.
- **Zone 13 (318 stations)**: **East Coast Corridor** with key hubs at Vizianagaram, Khurda Road, and Cuttack. This zone links Odisha and northern Andhra Pradesh through the eastern coastal regions, providing critical freight connectivity for ports and industrial centers along the Bay of Bengal.
- **Zone 21 (314 stations)**: **Northern Network** connecting Ludhiana, Ambala, and Jammu. This zone extends to India's northern frontiers and has strategic importance for connectivity to Jammu & Kashmir. Its higher clustering coefficient (0.36) suggests resilient local connectivity despite challenging terrain.

- **Bridge Station Significance:**
  - **Mughal Sarai (MGS)**: Connects **8 communities**; top **super-hub** critical to national flow.
  - **Secunderabad (SC)** & **Vijayawada (BZA)**: Link **5 communities** each; key south-central bridges.
  - **New Delhi (NDLS)**: Central northern hub, connecting **5 communities**.
  - **High Betweenness Edges** (serve as critical bottlenecks):
    Habibganj–Delhi Sarai (0.1545)
    Secunderabad–Balharshah (0.1348)
    Mughal Sarai–Patliputra (0.1147)
  - **Network Vulnerabilities**: Single-link dependencies like **Howrah–Bhubaneswar** and **Salem Junction** highlight areas needing redundancy.

- **Practical Implications:**
    - **Capacity Planning**: Prioritize high-betweenness routes for congestion relief, especially **Habibganj–Delhi Sarai** and **Secunderabad-Balharshah**.
    - **Resilience**: Enhance capacity or develop alternatives near key junctions like **Mughal Sarai**.
    - **Regional Development**: Zones with long internal paths (e.g., **Zone 30**) need strategic links to reduce travel times.
    - **Zone Management**: Consider reorganizing operations around **naturally detected zones** for better efficiency.
    - **Targeted Upgrades**: Improve stations with **high network importance**, not just high footfall.
    - **Future Expansion**: Build new links to ease load on **single-edge connectors**, e.g., alternatives to **Mughal Sarai–Patliputra**.

Deliverable 4: Robustness and Critical Node Analysis

To evaluate the resilience of the Indian railway network, we simulate targeted attacks on stations (nodes) based on centrality measures and analyze the resulting impact on network connectivity and cohesion.

Methodology

We modeled the railway network as a graph where:

- Nodes represent stations (based on Station_Code).
- Directed edges represent connections between consecutive stations in a train's route.

We converted the network to an undirected graph to focus on structural robustness rather than directional flow.

Two attack strategies were simulated:

1. Degree Centrality-Based Attack:
   Iteratively removed stations with the highest degree (most direct connections).
2. Betweenness Centrality-Based Attack:
   Iteratively removed stations with the highest betweenness (key bridges on shortest paths).

For both strategies, we removed nodes in batches (typically 5 at a time) and after each step computed:

- Number of connected components.
- Size of the largest connected component.

Specialities of the two Attack Strategies

1: Degree Centrality Attack (High-degree node removal):

- Targets the most "connected" stations—those with the highest number of direct links (routes).
- Removes major hubs that are locally central but not necessarily bridging distant parts of the network.
- Gradually degrades the network—slower fragmentation but consistent reduction in size of the largest component.

2: Betweenness Centrality Attack (High-bridge node removal):

- Targets nodes that lie on the most paths between other nodes—often crucial “bridges” between different network clusters.
- Extremely effective in fragmenting the network with fewer removals.
- Causes a rapid increase in the number of components and a steep drop in the largest component's size.

Comparison between the two Attack Strategies

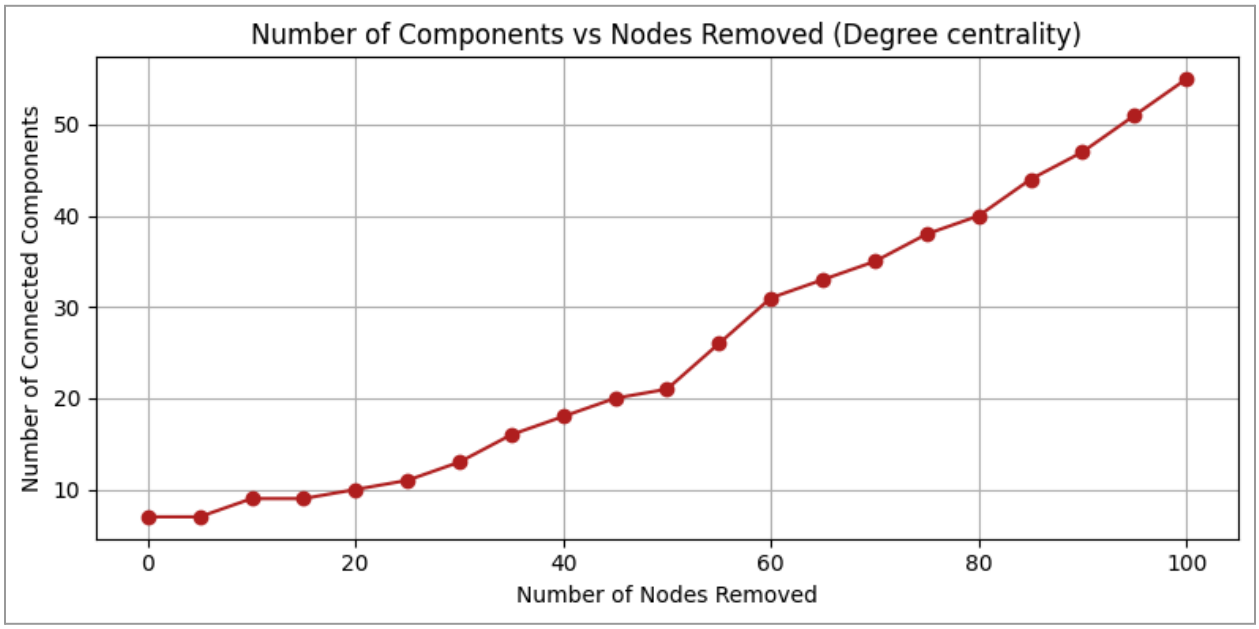
Metric	Degree Attack	Betweenness Attack
Speed of Fragmentation	Slower	Faster
Network Cohesion	Gradually weakens	Rapid breakdown
Targets	Locally connected hubs	Structurally vital bridges
Most Critical For	Traffic load analysis	Structural resilience study

Insights based on the Attack Graphs for the 2 Strategies

Degree Attack – Number of Components vs Nodes Removed:

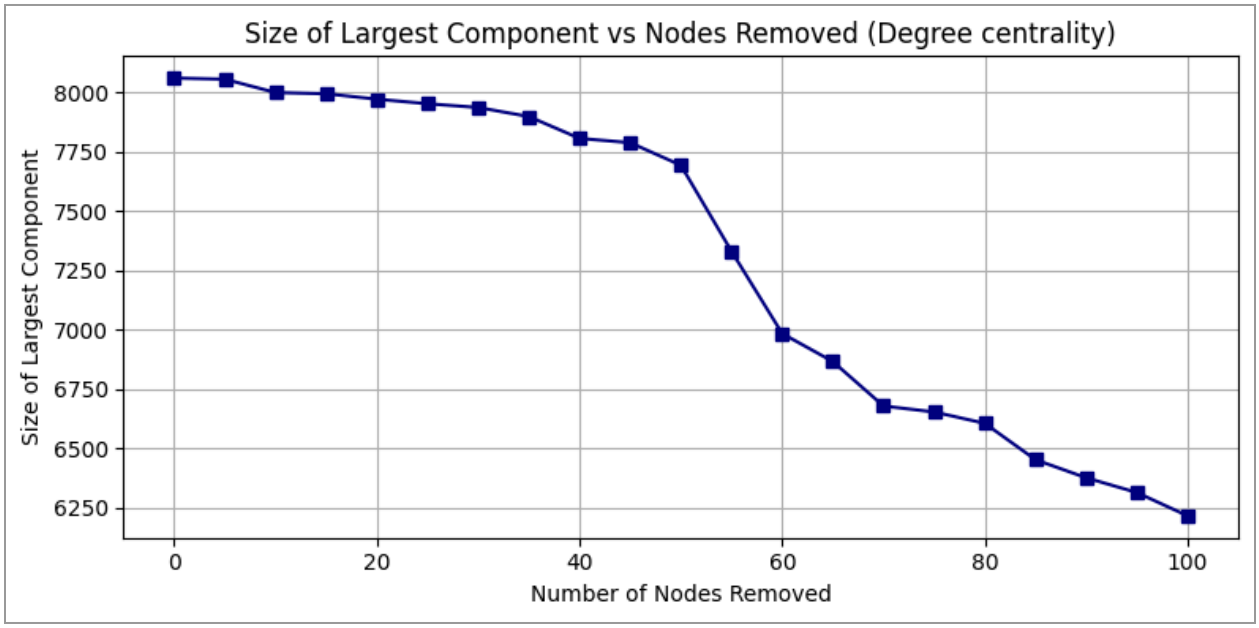
- Number of connected components increases steadily with node removals.
- Growth is nearly linear after about 30 nodes are removed.
- Indicates gradual fragmentation rather than abrupt collapse.

- Suggests that high-degree nodes are spread out, not forming critical bridges.



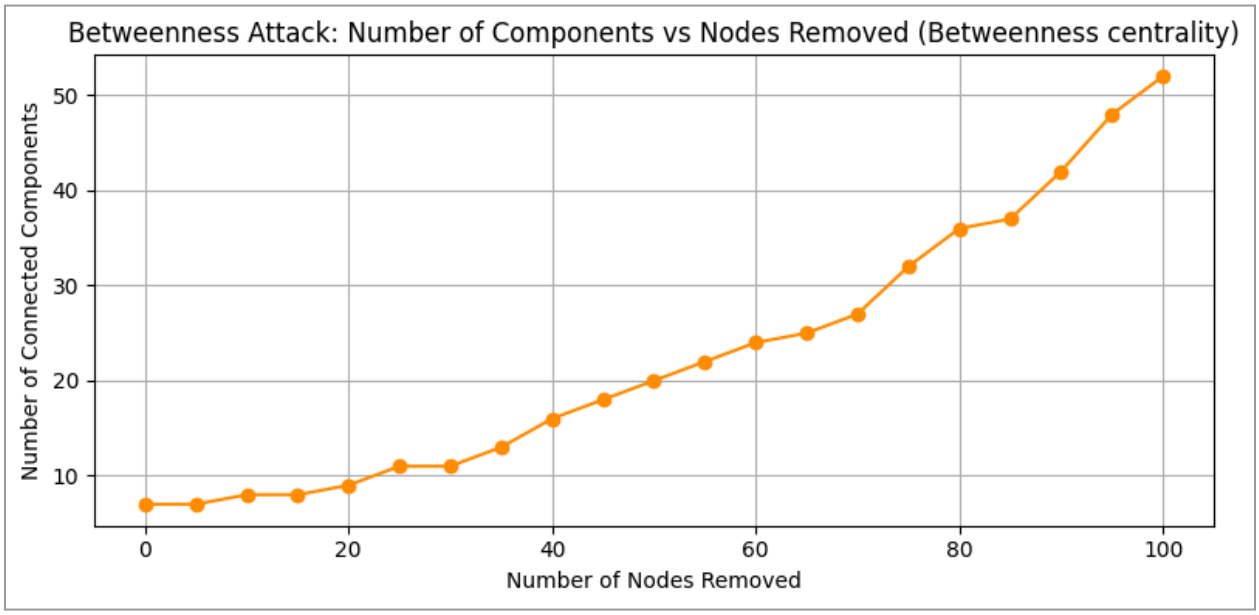
Degree Attack – Size of Largest Component vs Nodes Removed:

- More gradual reduction in component size than betweenness attack.
- Even after removing 100 highest-degree nodes, largest component retains ~6200 nodes.
- Shows that hubs are important but not as structurally critical as bridges.



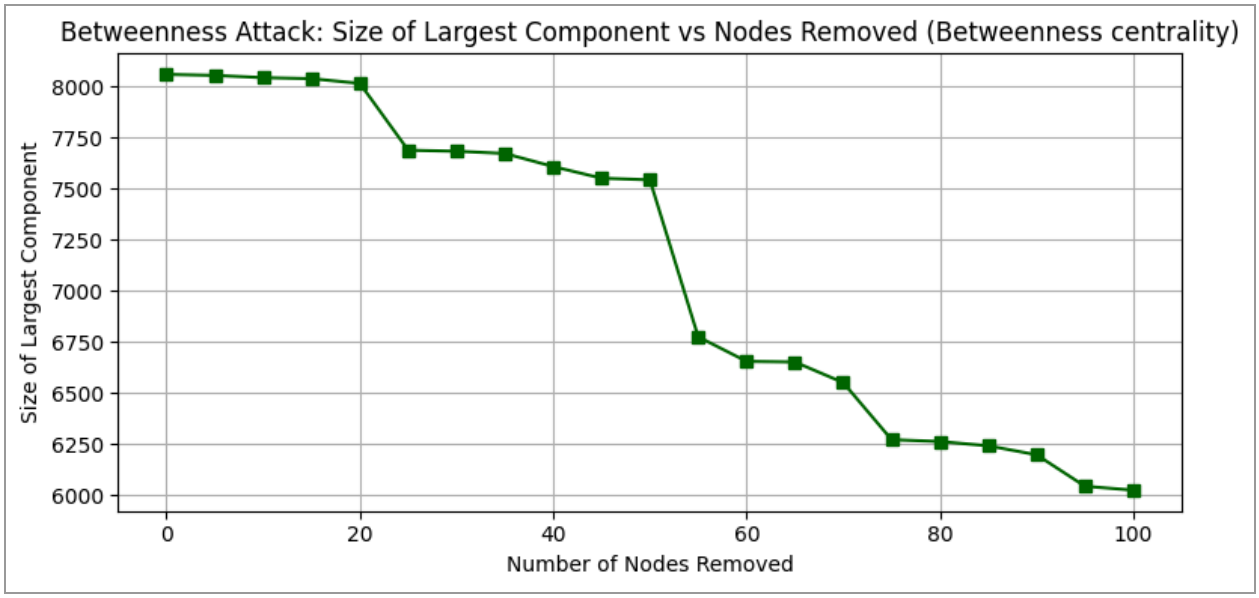
Betweenness Attack – Number of Components vs Nodes Removed:

- Starts with 6–7 components, increases rapidly as top bridge nodes are removed.
- After ~40–50 nodes, fragmentation accelerates—indicating those nodes were critical structural bridges.
- At 100 removals, the graph is highly fragmented (50+ components).
- Suggests that the Indian railway network is vulnerable to bridge disruptions.



Betweenness Attack – Size of Largest Component vs Nodes Removed:

- Largest component remains above 8000 nodes until ~20 nodes removed.
- After 50 nodes, sharp drop from ~7500 to ~6700, then gradual decline.
- Indicates that many smaller bridges don't disrupt the backbone—but a few key ones do.

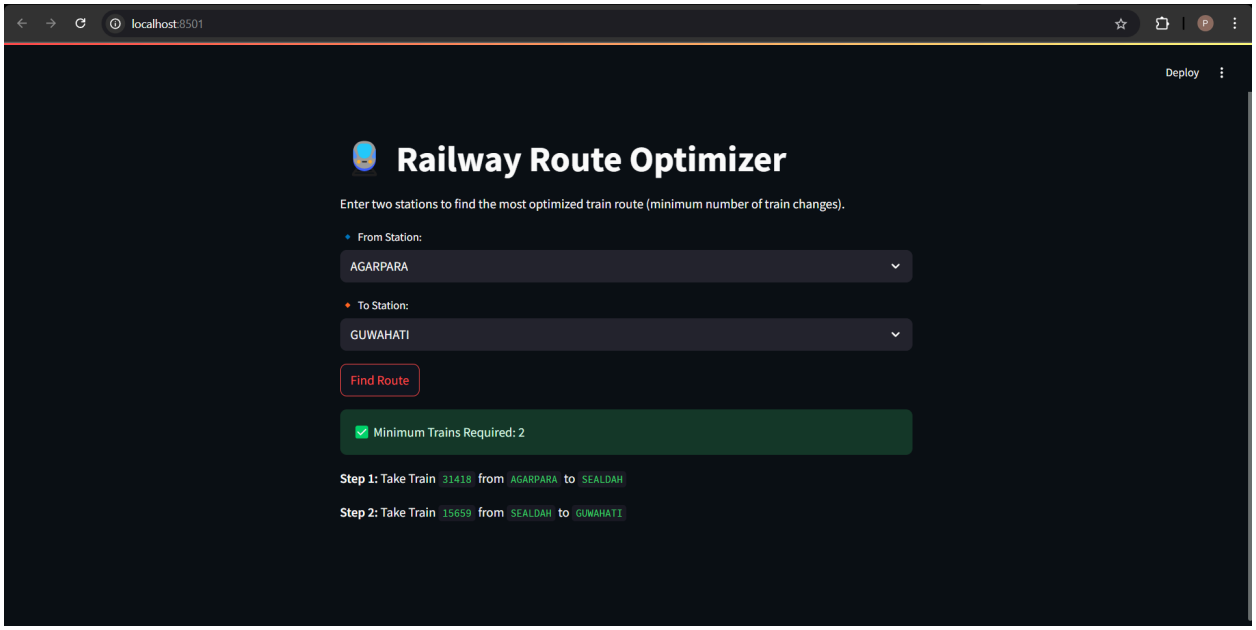


Practical Insights and Applications in real life:

- The Indian railway network exhibits high robustness against high-degree node failures but is structurally vulnerable to disruption of bridge nodes.
- Betweenness centrality reveals critical structural bottlenecks better than degree centrality.
- These findings can inform decisions on monitoring and hardening critical junctions to ensure continued national connectivity.

Interactive Interface for Small-World Property Visualisation

To enhance the interpretability of our railway network model, we developed an interactive route-finding interface using **Streamlit**, a Python-based web framework. This interface allows users to input any two station codes and visualise the optimal route with the minimum number of train changes between them. The backend leverages our directed graph model, constructed from real Indian railway data.



Steps Implemented

1. Train-to-Station & Station-to-Train Mappings: We preprocessed the data to efficiently track which trains pass through which stations and in what order.
2. BFS-Based Optimal Route Search: A modified Breadth-First Search (BFS) algorithm was used to find the shortest path (in terms of train changes) from the source to the destination.
3. Frontend Integration: Using Streamlit, the user inputs source and destination stations using drop downs and receives:
    a. A step-by-step route between the source and destination stations they entered
    b. Train numbers
    c. Stations where train changes (if any) occur

Results and Functionality

- Users can dynamically test connectivity across thousands of station pairs in real-time.
- The interface successfully handles edge cases such as:
    - No path exists
    - Source and destination are the same
    - Multiple paths with different train combinations

Small-World Property and 6 Degrees of Freedom

Average Path Length Insight: Through our BFS-based route finder, we observed that the average number of train changes or steps between any two nodes is relatively low, typically less than or equal to 5 hops, even in a network with over 8000+ stations.

This aligns with the Small-World Property, which suggests:

- Most nodes can be reached from every other by a small number of steps
- The network exhibits high clustering and low average path length

The interface reinforces the “Six Degrees of Separation” concept in the Indian railway context, showing that any two stations are often reachable within a few transitions, despite the large geographic and topological size.