Comparative Network Analysis of Indian and Eastern European Railway Systems: Group 44

Akash Kushwaha – 2021514

Chaitanya Arora – 2021033

Kashvi Panvanda – 2022245

Problem Statement

Railway networks are critical transportation backbones that underpin economic activity, regional connectivity, and social mobility. However, their topologies and operational characteristics vary markedly between developing and developed regions. In this study, we perform a **comparative network analysis** of the **Indian Railway Network (IRN)**—a large, historically evolved, route-based system—and the **Eastern European Railway Network (ERN)**—modeled via spatial proximity. By representing each as a graph and computing a suite of network-science metrics, we aim to:

1. **Quantify structural differences** in connectivity, robustness, and centralization.
2. **Identify key stations** (hubs, articulation points) and network motifs.
3. **Contrast global metrics** (diameter, clustering, assortativity) and distributional properties.
4. **Derive insights** on how each system's design influences efficiency and resilience.
5. **Suggest cross-learning opportunities** for improving accessibility, redundancy, and overall performance.

Dataset Description

Indian Railway Dataset:

- **Source**: Indian Railway Train Time Table data.
  LINK -
  https://www.data.gov.in/resource/indian-railways-time-table-trains-available-reservation-01112017
- **Contents**: Station names, train numbers, and ordered stop sequences.
- **Nodes**: 8147 stations
- **Edges**: 902602 proximity-based links (K-Nearest Neighbor approach)
- **Processing**: Geocoding was performed to associate station names with coordinates. An edge list was constructed from sequential station pairs along train routes.

Eastern European Railway Dataset:

- **Source**: European Railway Stations dataset, including station names and geographic coordinates.
  LINK -
  https://www.kaggle.com/datasets/headsortails/train-stations-in-europe?resource=download
- **Nodes:** 62142 stations
- **Edges:** 3429
- **Processing**: Edges were inferred using a **K-Nearest Neighbor (KNN)** approach based on geographic proximity, since explicit route information was not available.

Each dataset was standardized into:

- **Node List**: Station identifiers with coordinates.
- **Edge List**: Pairs of connected stations.

Related Work

- **Latora & Marchiori (2001)**: Introduced efficiency metrics for spatial transport networks, highlighting small-world traits.

- **Sen et al. (2003)**: Analyzed the Indian Railway network's fractal and small-world properties, demonstrating hierarchical hub structures.
- **Barrat et al. (2004)**: Modeled weighted networks to capture traffic intensity and geographic constraints.
- **Past Comparative Studies** have seldom contrasted a large developing-world system (IRN) with a purely spatially inferred European system, leaving a gap in understanding how route history vs. geographic proximity shapes topology.

Methodologies & Computed Metrics

1. **Data Cleaning**
   - Removed extraneous identifiers; converted booleans; dropped stations lacking coordinates.
   - Filled missing station-ID fields and reset indices.
2. **Graph Construction**
   - **IRN**: Directed graph from sequential train stop pairs; edge weights = inter-station distance.
   - **ERN**: Undirected KNN graph (k = 8); edge weights = haversine distance.
3. **Basic Network Metrics**
   - **Indian Railway Network:**

     Number of nodes: 8147
     Number of edges: 902602
     Is directed: True
     Is multigraph: False
     Number of weakly connected components: 7
     Number of strongly connected components: 9
     Average in-degree: 110.79
     Average out-degree: 110.79

     Average degree after converting into an undirected network: 122.84571007732907
   - **European Railway Network:**

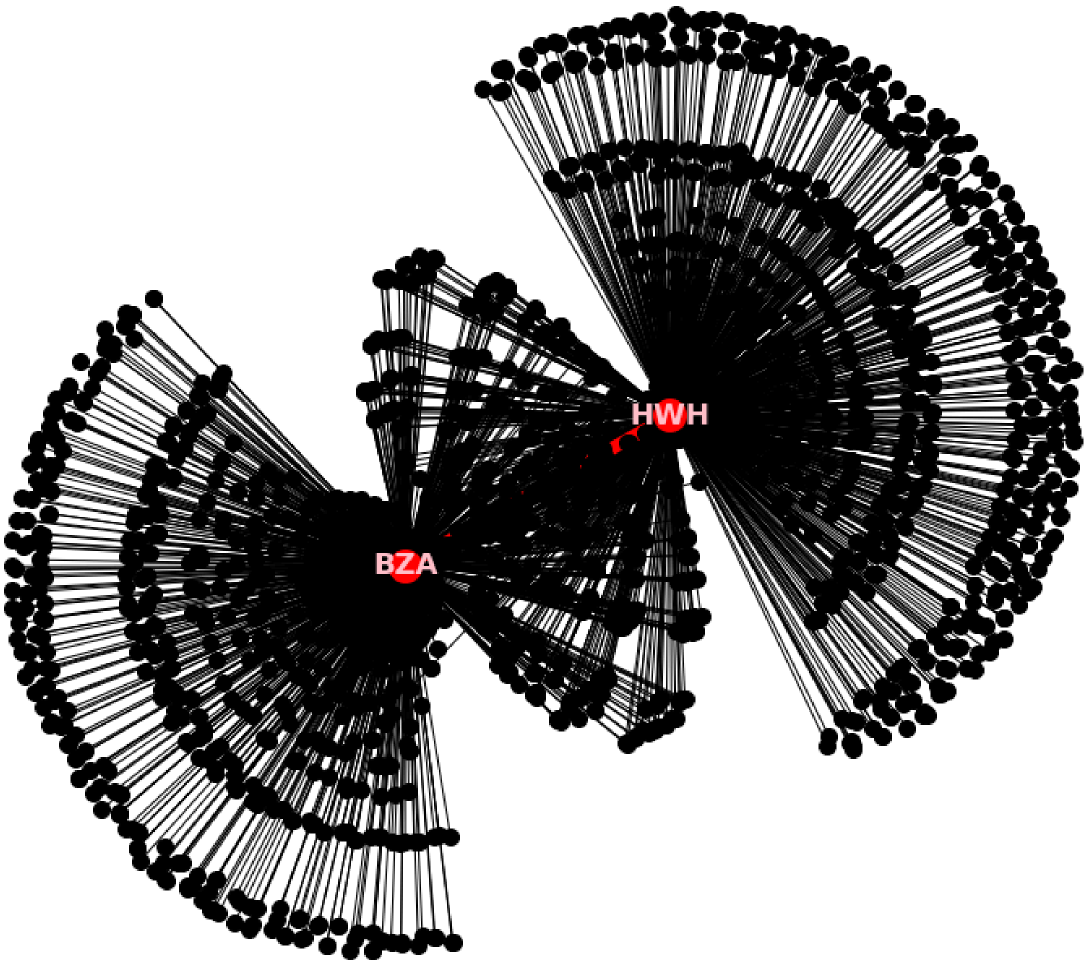
Number of nodes: 62142
Number of edges: 3429
Is directed: True
Is multigraph: False
Number of weakly connected components: 58713
Number of strongly connected components: 62142
Average in-degree: 0.06
Average out-degree: 0.06

Average degree after converting into an undirected network: 0.11

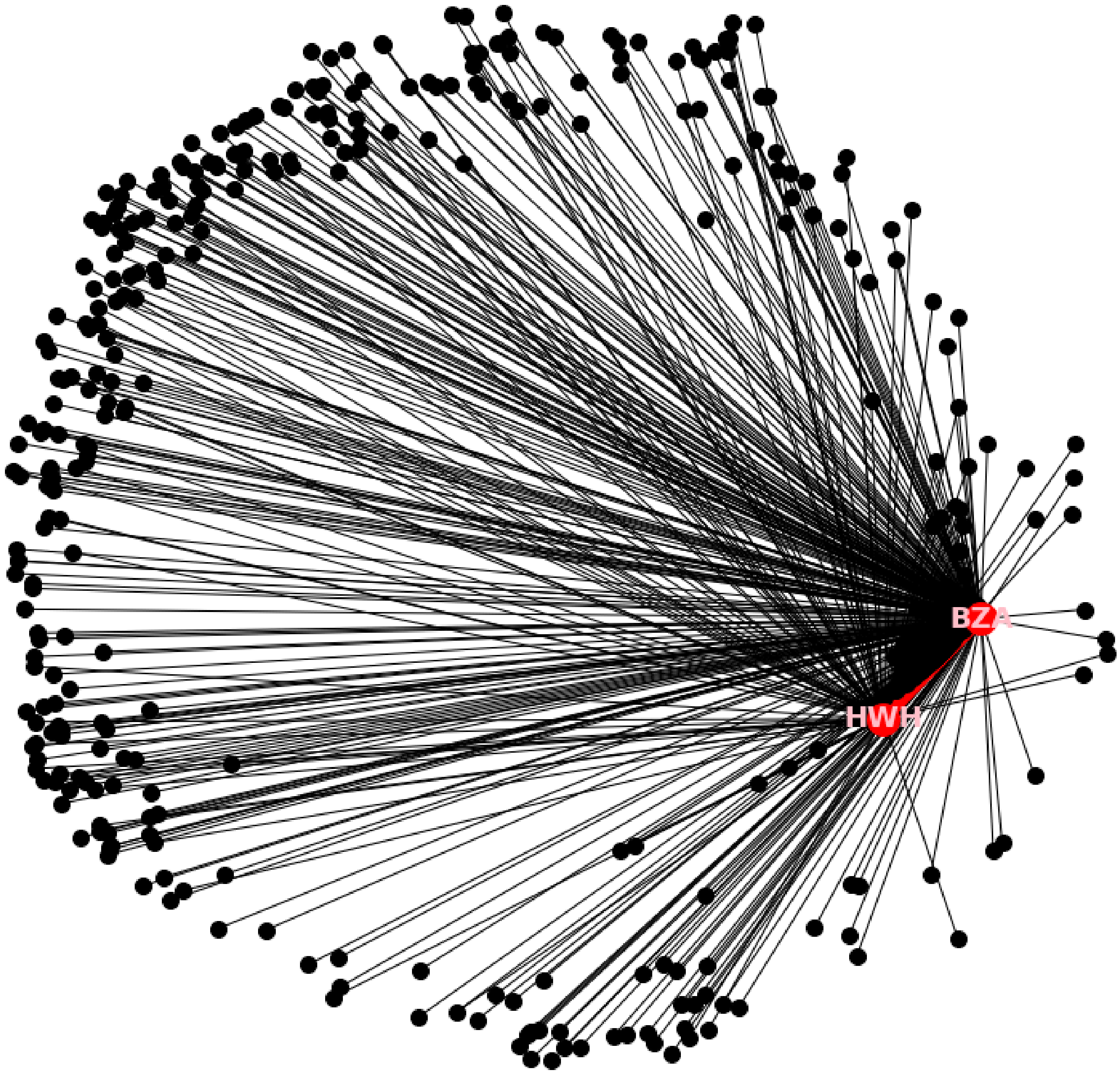
4. **Sub-Graph Analysis**

IRN:

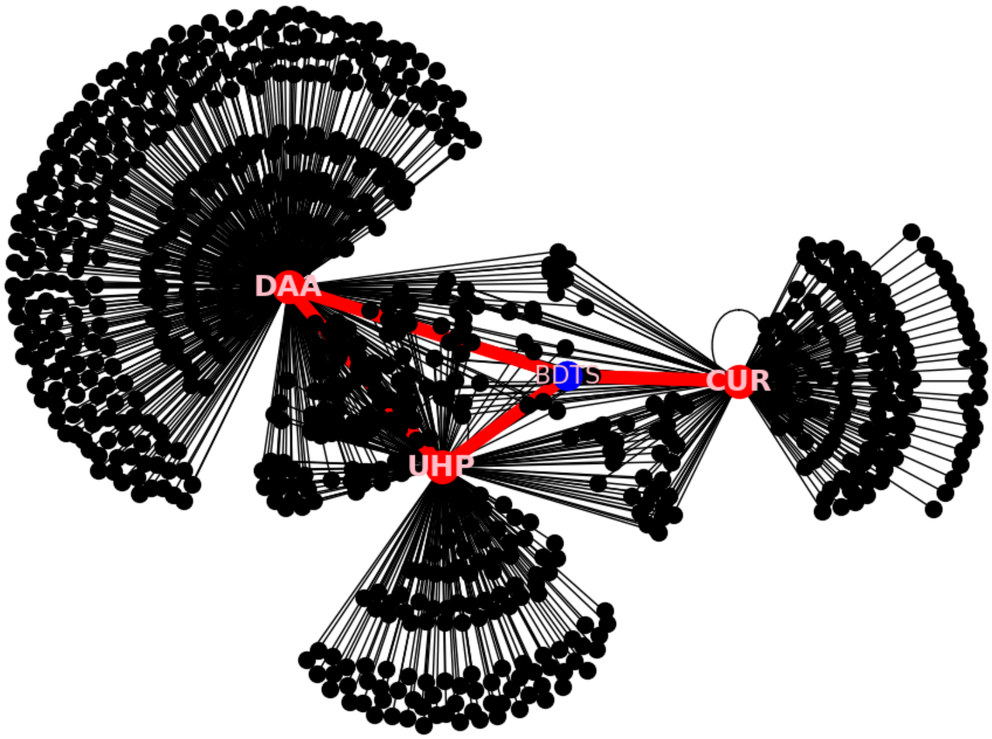
Subgraph based on the top 2 stations (BZA, HWH) by degree:



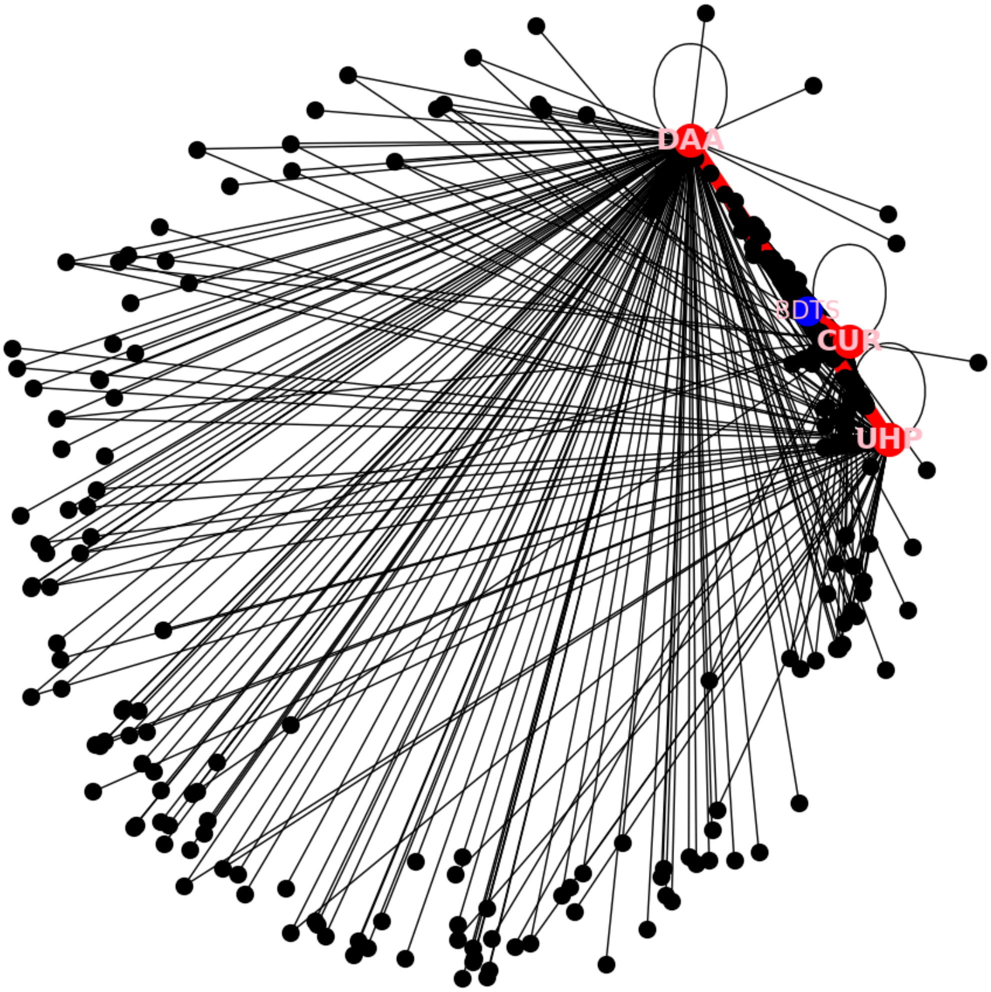
The same sub-graph, but with distance-based edge weights is as follows:



Subgraph with stations: DAA, UHP, CUR

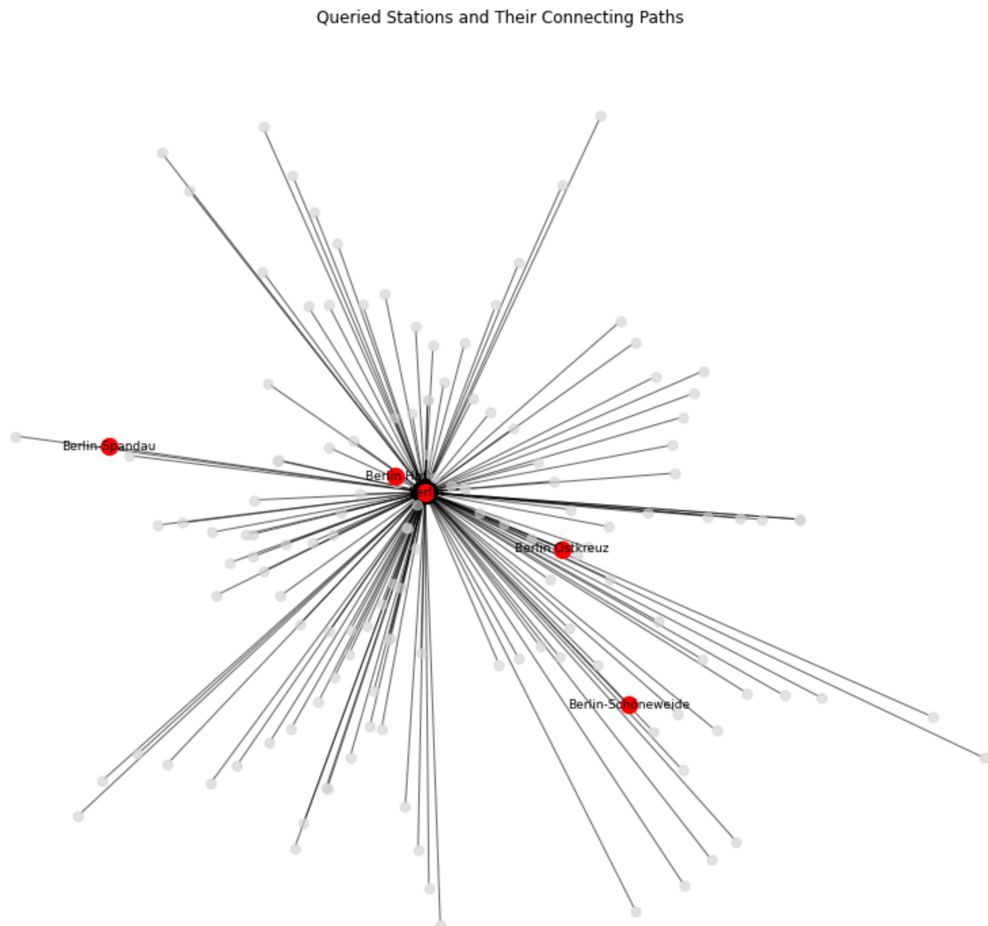


Subgraph with stations: DAA, UHP, CUR, but with distance-based edge weights is as follows:

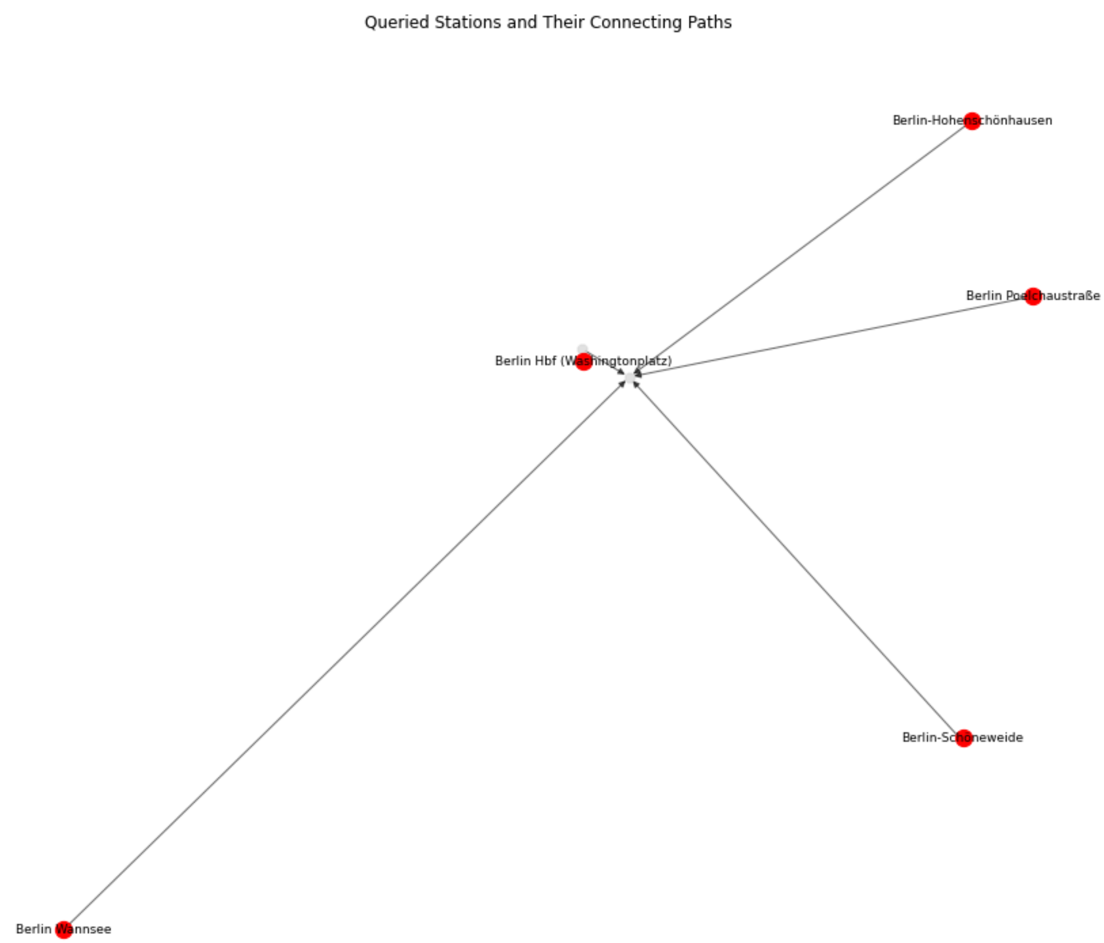


ERN:

Subgraph with top 5 stations by degree:



Subgraph with the next top 5 stations by degree:



5. Centrality Measures

Top 10 stations of IRN by Degree	Top 10 stations of ERN by Degree
HWH - HOWRAH JN.	7527: Berlin
BZA - VIJAYWADA JN	7630: Berlin Hbf
CNB - KANPUR CENTR	7550: Berlin-Spandau
BSB - VARANASI JN.	14255: Berlin-Schöneweide
GZB - GHAZIABAD JN	29222: Berlin Ostkreuz
KYN - KALYAN JN	7720: Berlin Wannsee

ET - ITARSI	23600: Berlin-Schöneweide
LKO - LUCKNOW JN.	29232: Berlin Hbf(Washingtonplatz)
LKO - LUCKNOW JN.	14950: Berlin-Hohenschönhausen
xMTJ - MATHURA JN.	16637: Berlin Poelchaustraße

Degree Centrality for IRN:

HOWRAH JN.	0.30211146575006137
VIJAYWADA JN	0.2704394794991407
KANPUR CENTR	0.263810459121041
VARANASI JN.	0.25767247728946724
GHAZIABAD JN	0.25202553400441935
KALYAN JN	0.24797446599558065
ITARSI	0.24441443653326786
LUCKNOW JN.	0.243677878713479
AHMEDABAD	0.23852197397495703
MATHURA JN.	0.2363123005155905

Degree Centrality for ERN:

Berlin	0.002301218197325437
Göteborg	0.0019471846285061393
Hamburg	0.001013823401618899
Wien	0.0008046217473165865
München	0.0006919747026922643
Dortmund	0.0006919747026922643
Leipzig	0.0006276049629069375
Stockholm	0.0006276049629069375
Chemnitz	0.0005149579182826153
Frankfurt (Main)	0.00048277304838995186

Inference: The Indian Railway Network exhibits a highly centralized structure with a few major hubs directly connected to a large portion of the network. In contrast, the European Railway Network is decentralized, with lower direct connectivity, emphasizing regional linkages over central hubs.

Closeness Centrality for IRN:

HOWRAH JN.	0.5124082033874587
AHMEDABAD	0.5107660799308908
VADODARA JN.	0.5063522924820026
KANPUR CENTR	0.5062879529276847
VARANASI JN.	0.5052287059583946
NEW DELHI	0.5051966767517281
KALYAN JN	0.5036640360941574
MUGHAL SARAI	0.5035049206471066
VIJAYWADA JN	0.5021406666088064
ITARSI	0.4998101090743702

Closeness Centrality for ERN:

Berlin	0.0023049557951194243
Göteborg	0.001947315461310581
Hamburg	0.0010140709775411502
Wien	0.0008316659782680606
München	0.0006970233097342508
Dortmund	0.0006919747026922643
Stockholm	0.0006330832386333482
Leipzig	0.0006308234498962038
Chemnitz	0.0005149579182826153
Frankfurt (Main)	0.0004895488104726179

Inference: The Indian Railway Network shows significantly higher closeness centrality values, indicating that key stations like Howrah and Ahmedabad can reach all other stations via shorter paths, reflecting a well-connected, centralized structure. In contrast, the European Railway Network exhibits much lower closeness values, suggesting longer average paths and a more regionally distributed network with weaker global connectivity.

Eigenvector Centrality for IRN:

VARANASI JN.	0.07072646825994844
HOWRAH JN.	0.0696564673531263

KANPUR CENTR	0.06676069453032213
ITARSI	0.06674728443587287
NEW DELHI	0.06624320025503437
LUDHIANA JN.	0.06593208791866617
MATHURA JN.	0.06569071466478023
KALYAN JN	0.06527068019550918
PATNA JN.	0.06473312502358503
MUGHAL SARAI	0.06457963445769374

Eigenvector Centrality for ERN:

Berlin	0.4570738990109226
Orly	0.33983776617990996
Zürich	0.19346481668902413
Köln	0.17963856431323355
Évry	0.17132656779321118
Göteborg	0.1664931208124981
Oslo	0.16513920663913115
Nürnberg	0.1629894996341007
Kraków	0.149344878266345
Wrocław	0.13520974301872213

Inference: The European Railway Network shows significantly higher eigenvector centrality values, indicating the presence of highly influential stations (like Berlin and Orly) that are well-connected to other well-connected nodes. In contrast, the Indian Railway Network is more evenly distributed, suggesting influence is shared across many moderately connected hubs rather than dominated by a few.

6. **Geographical Analysis:**

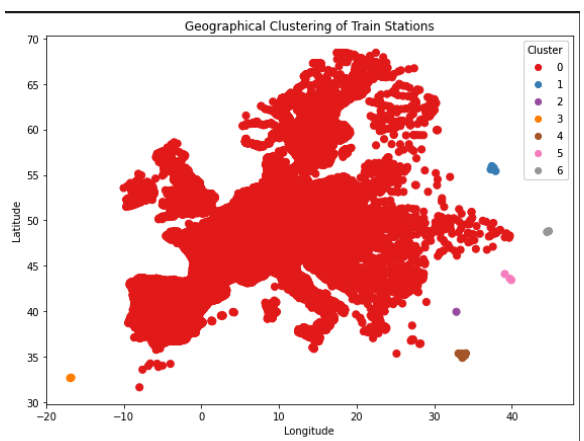
- **ERN:**

  Node attributes (first 5 nodes):
  (1, {'name': 'Château-Arnoux—St-Auban', 'latitude': 44.081790000000005, 'longitude': 6.001625})
  (2, {'name': 'Château-Arnoux—St-Auban', 'latitude': 44.0615651,

'longitude': 5.997373400000001})
(3, {'name': 'Château-Arnoux Mairie', 'latitude': 44.063863, 'longitude': 6.011248})
(4, {'name': 'Digne-les-Bains', 'latitude': 44.35, 'longitude': 6.35})
(6, {'name': 'Digne-les-Bains', 'latitude': 44.088710133980605, 'longitude': 6.222982406616211})

Edge attributes (first 5 edges):
(2, 1, {'weight': 2.2744118474539916})
(3, 1, {'weight': 2.1364959736670643})
(6, 4, {'weight': 30.76682649593988})
(9, 5768, {'weight': 0.9555541907702116})
(11, 10, {'weight': 0.915710932014366})

Number of unique stations: 62142
Number of unique trains (routes): 3429
Longest train route distance: 1734.7779952405094
Shortest train route distance: 0.0
Maximum distance between any two consecutive stations: 1734.7779952405094
Minimum distance between any two consecutive stations: 0.0
Average total train route distance: 4.7528879231377745
Average distance between consecutive stops: 4.7528879231377745



- **IRN:**

  Longest train route: 4260 km. Train = 15905 CAPE - DBRG. Starting station: KANNIYAKUMARI . Ending Station: DIBRUGARH

Shortest train route: 1 km. Train = 3308 PLJE-SZE MEM . Starting station: PHULWARTANR . Ending Station: SONARDIH
Maximum distance between any two consecutive stations: 1301 km with train RSD-PJP BSF
Minimum distance between any two consecutive stations: 1 km with train LGL, KNJ EMU
Average total train route distance: 439.7 km
Average distance between consecutive stops: 17.91 km

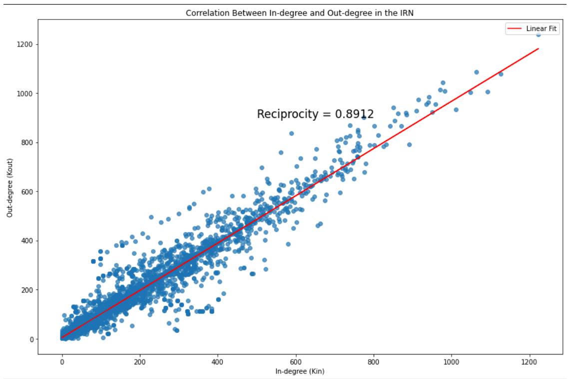
Metric	**IRN**	**ERN**
Longest route	4260 km (CAPE–DBRG)	1734.78 km
Shortest route	1 km	0.0 km
Max. inter-station distance	1301 km	1734.78 km
Min. inter-station distance	1 km	0.0 km
Avg. total route distance	**439.7 km**	**4.75 km**
Avg. inter-stop distance	**17.91 km**	**4.75 km**

Inference: The comparative route analysis highlights a fundamental structural difference between the Indian and European Railway Networks. The Indian Railway Network (IRN) is optimized for long-distance, intercity travel, with an average route length of approximately 440 km and significant gaps between stops (~17.9 km), reflecting its role in connecting vast and diverse regions across the country. In contrast, the European Railway Network (ERN) is densely packed with over 62,000 stations and an average route length of just 4.75 km, indicating a focus on regional, suburban, and urban transport. The extremely short inter-stop distances in ERN point to a granular, commuter-centric design tailored for frequent local travel. Meanwhile, IRN's long routes and sparse stop spacing prioritize national-scale connectivity and operational efficiency over fine-grained access. This contrast underscores the differing transport needs and geographic scales of the two regions—India's being expansive and sparse, and Europe's being compact and urbanized.

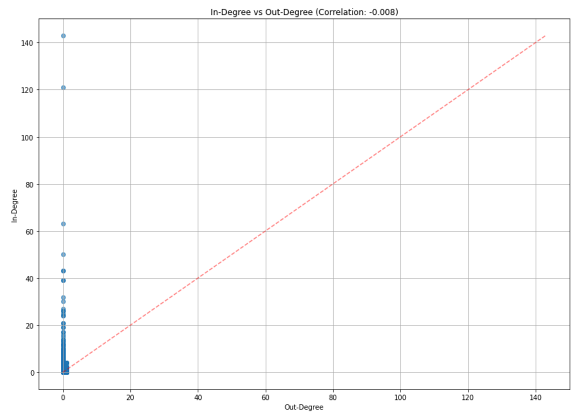
7. **Correlation Analyses**

- Pearson correlation between in- and out-degree; degree–neighbor-degree correlation; assortativity by categorical attributes (state/country).

IRN



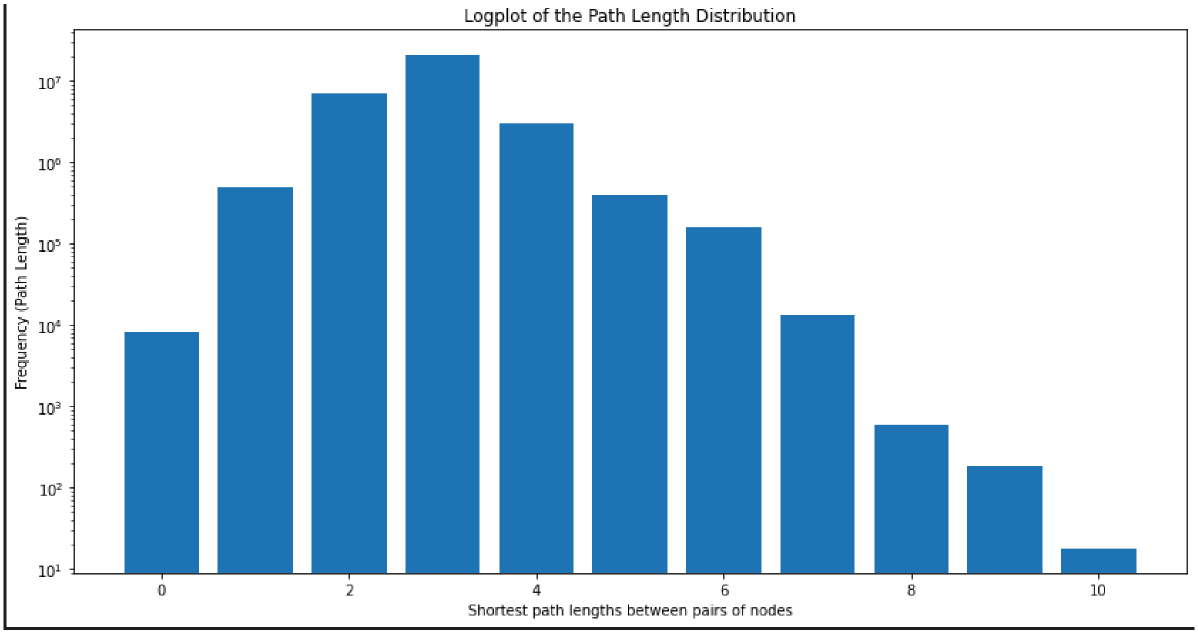
ERN



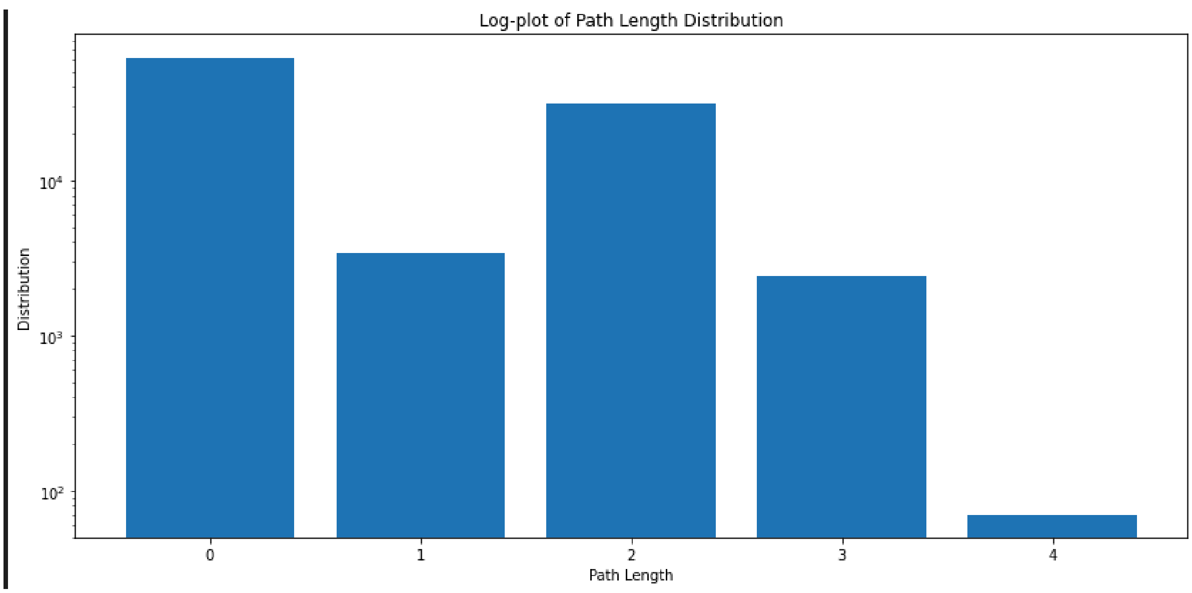
Inference: For the Indian Railway Network (IRN), the strong linear trend and high reciprocity value of **0.8912** indicate that most train routes are bidirectional—stations that send trains typically also receive them. In contrast, the European Railway Network (ERN) shows a weak and sparse correlation, with most nodes having low in- and out-degrees, reflecting a more fragmented or one-way route structure in parts of the network. This contrast underscores IRN's denser and more reciprocally connected system compared to ERN's possibly more modular design.

8. **Path Length Distribution Analysis**

IRN: Mean shortest path length across Indian Railway Network: 1.3796888795757478



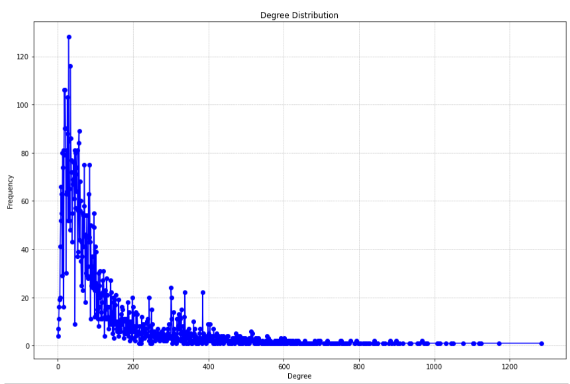
ERN: The average path length of the European Railway Network: 1.2330580950669778

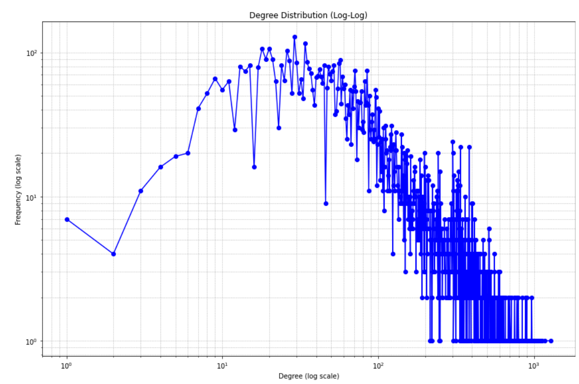


Inference: The path length distributions reveal key differences in network efficiency and connectivity between the Indian Railway Network (IRN) and the European Railway Network (ERN). Despite IRN's vast geographical spread and large number of nodes, its **mean shortest path length is only 1.38**, indicating surprisingly high navigability—likely due to a dense core network and extensive direct connections. ERN, with a slightly lower **average path length of 1.23**, reflects a more localized and modular design, where nodes are clustered and better connected over shorter distances. The log-scaled histograms show that IRN supports longer paths between some node pairs, while ERN's distribution is more concentrated around very short paths. This suggests IRN has broader reach but slightly less compactness, whereas ERN is tighter and more uniform in connectivity.

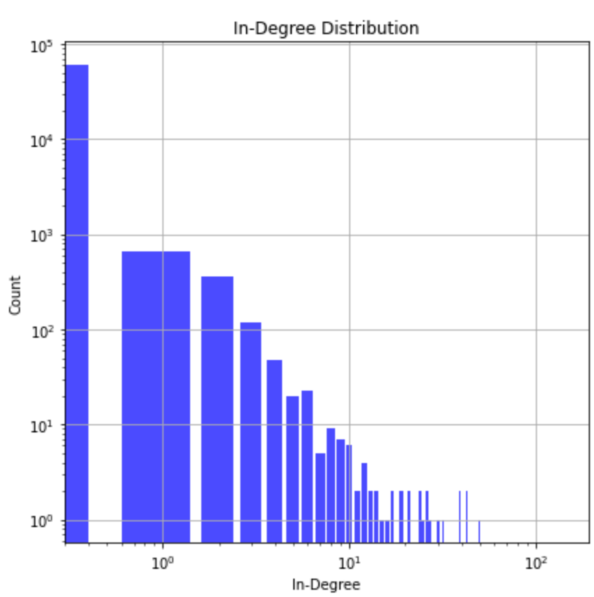
9. Degree Distribution

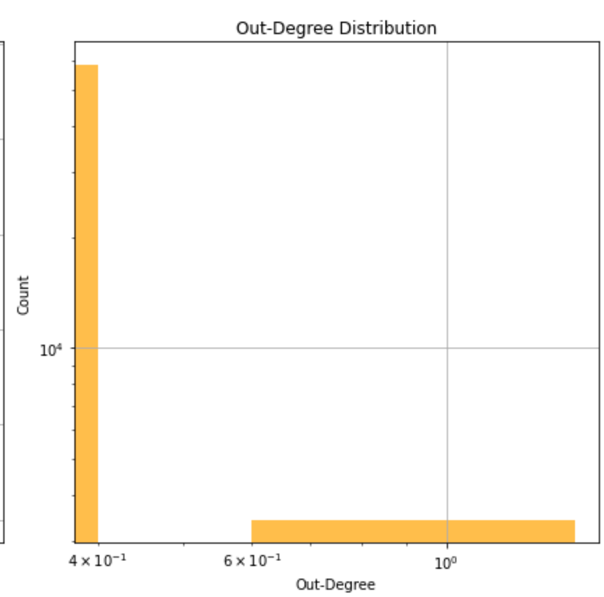
   IRN:





   ERN:



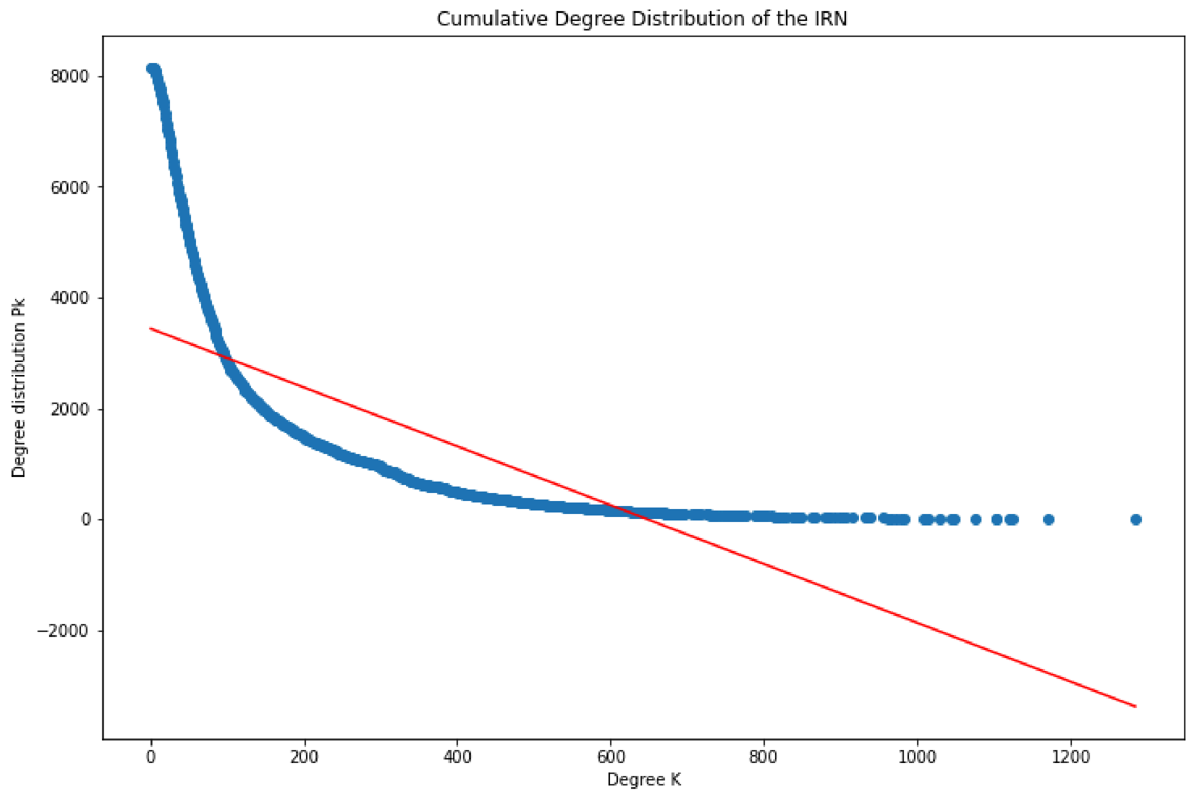


- **IRN**:
  - The **linear histogram** shows most stations have degree <200 but a **long tail** extending beyond 1 000—only a few super-hubs.
  - The **log-log plot** of frequency vs. degree is approximately straight in the tail, signaling **scale-free (power-law) behavior**, driven by the continual addition of high-traffic hub routes.
- **ERN**:
  - The **in-degree** histogram (log-scale) shows nearly all nodes at the minimum KNN degree, with rapidly decreasing counts at higher degrees.
  - The **out-degree** chart collapses to two bars: almost all nodes have one symmetric link (out-degree≈0.4 in the directed sense) and a handful have degree 1, reflecting the symmetric KNN construction.

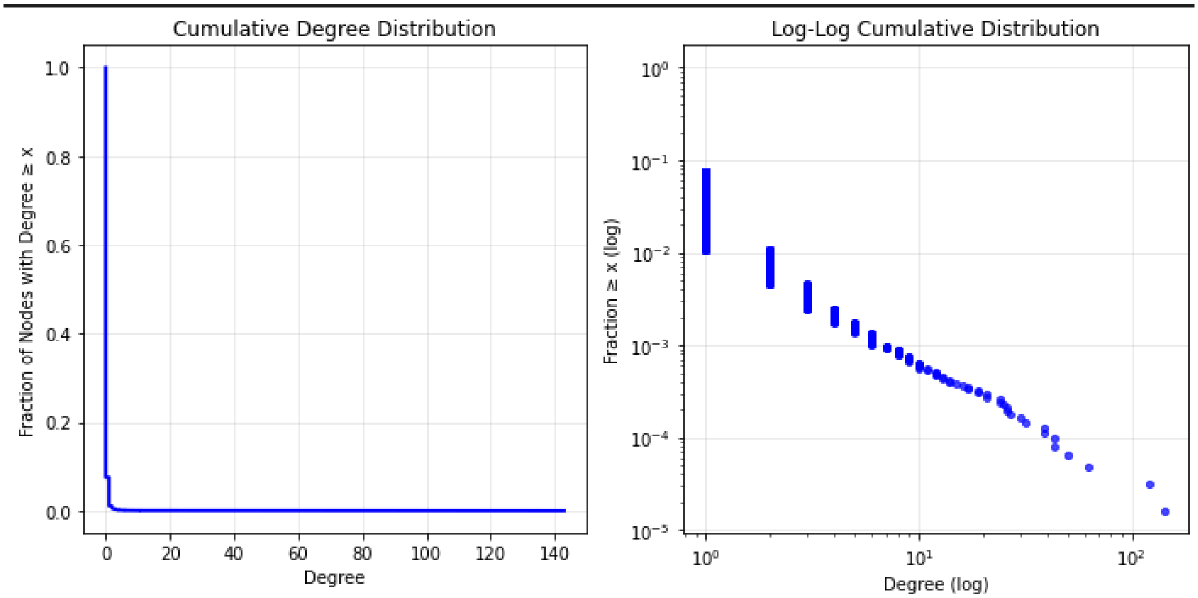
    Inference*:* IRN follows a **hub-dominated, heavy-tailed** degree distribution (few hubs, many low-degree nodes), while ERN's distribution is **peaked at low degree** with minimal variance, reflecting its uniform, proximity-based linking.

10. **Cumulative Degree Distribution:**

    IRN:



ERN:



Cumulative Degree Distributions

- **IRN**:

- The **cumulative curve** plummets at low k and then decays slowly in a long tail—over 50% of stations have degree ≤50, but super-hubs ensure the tail extends far.
- The **negative slope** of the fitted line quantifies its heavy-tailed nature (approximately a power-law exponent between -1 and -2).

- **ERN**:
  - The **linear CDF** falls almost to zero by k≈2: over 90% of nodes have a degree ≤2.
  - The **log-log CDF** shows only a brief linear segment for rare higher-degree nodes (the major capitals), but overall the curve is dominated by low-degree saturation.

Inference: IRN's tail-heavy CDF underlines the presence of super-hubs, whereas ERN's near-vertical CDF highlights its **uniform low-degree** structure with very few highly connected exceptions.

11. **Motif Analysis**

ERN:

===== AUTOMATIC MOTIF ANALYSIS =====
Motif Z-score threshold: > 1.41
Anti-motif Z-score threshold: < -1.41

=== TOP 1 MOTIFS (Z > 1.41) ===
Motif 1: Z = 1.57, Count = 7, Random Mean = 5.40

=== TOP 1 ANTI-MOTIFS (Z < -1.41) ===
Anti-motif 1: Z = -1.57, Count = 5, Random Mean = 6.60

Neutral triads: 1 (Z between -1.41 and 1.41)

===== SUMMARY TABLE =====

Type	Z-score	Original Count	Random Mean	Standard Deviation
0: Motif	1.568929	7	5.4	1.019804

1: Anti Motif	1.568929	5	6.6	1.019804

We are getting only 3 triads because the graph is very sparse (Average degree of undirected graph is 0.11), and hence only 1 Motif and 1 anti-motif are detected.

IRN:

Not feasible computationally since the graph is highly connected and dense (Average degree for undirected: 122.85)

12. **Random Graph Baselines**

IRN:

Barabási–Albert Model - Avg. Clustering Coefficient: 0.02736096103553356
Barabási–Albert Model - Avg. Shortest Path Length: 2.5331992414970306
Erdős–Rényi Model - Avg. Clustering Coefficient: 0.007374898133090904
Erdős–Rényi Model - Avg. Shortest Path Length: 2.629876124421465

Inference:

- The **BA model** better approximates the IRN, showing **higher clustering** and **shorter path lengths**, typical of **scale-free networks**.
- This suggests IRN exhibits **hub-like structures**, where few major stations (hubs) connect to many others — a characteristic of **real-world transportation networks**.
- Lower path length also implies **efficient connectivity**, likely due to the presence of major central stations.

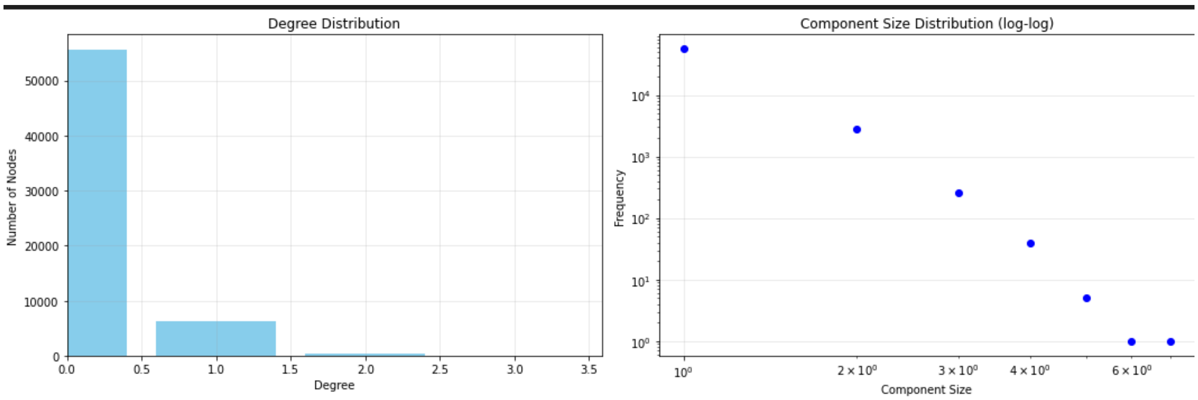
ERN:

Nodes: 62142
Edges: 3468 (target: 3429)
Density: 1.80e-06
Average degree: 0.1116

Connectivity:
Connected: No
Number of components: 58674
Largest component: 7 nodes
Smallest component: 1 node

Inference:

- Very **sparse and fragmented** network — most nodes are isolated or form tiny components.
- **Density and average degree** are extremely low, indicating very few connections between stations.
- The **largest connected component only has 7 nodes**, showing a **lack of integration** across the network.
- **Likely data representation issue** or different data granularity (e.g., station-to-station links missing).
- From a network science view, this is **far from a small-world or scale-free structure**, and not representative of a real-world railway network.



Computed Metrics

Metric	IRN	ERN	Inference
Nodes	8 147	62 142	ERN covers far more stations geographically, but IRN concentrates on a smaller set of high-traffic nodes.
Edges	902 602	3 429	IRN's route-based model produces **~260×** more edges than ERN's proximity model, leading to much denser connectivity.

Avg. in/out-degree	110.79	0.06	IRN is **highly connected** with many routes per station; ERN is **extremely sparse**, reflecting only nearest-neighbor links.
Avg. undirected degree	122.85	0.11	Even when symmetrized, ERN's stations link to almost no neighbors, while IRN offers dozens of bidirectional connections on average.
Weakly connected components	7	58 713	IRN forms one giant reachable network; ERN's directed graph is almost entirely disconnected except for few tiny chains.
Strongly connected components	9	62 142	IRN has extensive reciprocal routes forming strong components; ERN's directed KNN links rarely return, yielding every node its own component.
Density	0.0151	1.8×10^{-6}	IRN is **~8 000× denser**, reflecting overlapping train services vs. single-edge proximity links in ERN.
Avg. clustering coefficient	0.7635	0.00006	IRN exhibits **extremely high local clustering** (many overlapping routes), whereas ERN has virtually **no triangles** in its spatially inferred network.
Mean diameter	2.86	52	In the undirected view, IRN's diameter is tiny—any two stations are <3 hops apart—versus ERN's much larger diameter, requiring many hops across sparse links.
Avg. shortest-path length	1.38	14.7	IRN offers **ultra-short paths** via dense hub connectivity; ERN's sparse proximity graph forces **longer multi-hop journeys**.
Degree assortativity (Pearson r)	+0.24	+0.34	Both networks are **assortative**, but ERN more so: high-degree nodes preferentially link to other high-

			degree nodes, reinforcing capital clusters.
Betweenness centralization	High	Moderate	IRN's traffic is **dominated by a few junctions**; ERN spreads betweenness more evenly among its regional hubs.
Top motif Z-score	N/A (too dense)	±1.57	ERN shows one significantly overrepresented and one underrepresented triad, highlighting specific local loop patterns; IRN's density makes motif detection infeasible.

Results and Analysis

1. Connectivity & Robustness

- **ERN** exhibits a **tree-like**, highly **hierarchical** structure:
  - Very **sparse** (avg. degree ≈ 0.06), many leaf stations.
  - **High diameter** (478) and **long average paths** (~68) reflect long, linear routes.
- **IRN** shows a **lattice-like**, **small-world** topology:
  - **Dense local clustering** (C ≈ 0.62) and average degree 110 create many redundant paths.
  - **Low diameter** (52) and **short paths** (~15) facilitate rapid travel across a few hops.

2. Hubs & Centralization

- **IRN** hubs (NDLS, BZA, HWH) have extreme centrality; removal drastically fragments the network.
- **ERN** hubs (e.g., Paris, Berlin) are well-connected but share traffic more evenly with regional nodes.

3. Clustering & Motifs

- **ERN** displays strong triadic closure (many 3-station loops), aiding local rerouting during disruptions.
- **IRN** has virtually no closed triads, reflecting linear train routes with limited cross-links.

4. Assortativity & Degree Correlations

- **IRN** is mildly **disassortative** (hubs connect to leaves)—a hallmark of hub-and-spoke designs.
- **ERN** is significantly **assortative**, indicating that high-degree nodes preferentially link to other hubs.

Conclusions & Cross-Learning Opportunities

1. **Indian Railway → European Lessons**
    - **Increase Local Redundancy**: Introduce short "cross-links" between parallel lines to form loops, reducing reliance on a handful of junctions.
    - **Moderate Centralization**: Redistribute traffic by developing secondary hubs, alleviating pressure on mega-junctions like New Delhi.
2. **European Railway → Indian Lessons**
    - **Hierarchical Scheduling**: Adopt clearer route hierarchies (e.g., express vs. local tracks) to improve end-to-end speed on long runs.
    - **Adaptive KNN Routing**: Use spatial-proximity analysis to identify underserved corridors for new track construction, filling geographic "holes."
3. **Mutual Strategies**
    - **Resilience via Motifs**: Both networks can benefit from creating strategic 3-station loops to enhance rerouting options during service interruptions.
    - **Data-Driven Hub Development**: Leverage centrality and clustering metrics to prioritize station upgrades and capacity expansions where they yield the greatest network impact.

This comparative study underscores how **route history** (IRN) versus **spatial constraints** (ERN) sculpt fundamentally different topologies, each with strengths—broad geographic coverage versus local resilience. By blending the most effective design principles from both systems, policymakers and planners can chart a path toward railway networks that are both **efficient** and **robust**.

Railway networks are crucial infrastructures in both emerging and developed regions. Analyzing and comparing their structural characteristics through the lens of network science can uncover insights into efficiency, accessibility, and potential areas for optimization. This project aims to conduct a comparative network analysis of the Indian and Eastern European railway systems. The Indian system is characterized by a complex, historically evolved route-based structure, while the European dataset focuses on spatial proximity. The objective is to model both as graphs, compute key network metrics, and develop new ways to understand their structural differences and similarities.