**Indian Railway network analysis**

**Shaunak Deshpande1 , Ritik Agarwal 1** , **Azam Siddiqui1**

**(19BCE1310) (19BCE1063) (19BCE1092)**

*1 Vellore Institute of Technology, Chennai*

**Abstract**

Indian Railways have been functional from 1836. Indian Railways is among the world's largest rail network, and its route length network is spread over **67,956 kms**, with **13,169 passenger trains** and **8,479 freight trains**, plying **23 million** travelers and **3 million tonnes** (MT) of freight daily from 7,349 stations.

Railways provide the cheapest and most convenient mode of passenger transport both for long distance and suburban traffic. For any country railways are one of the primary sources to carry goods and material. Indian railways manages the fourth largest national railway system in the world by size.

We perform a Railway Analysis on the Indian Railway network and study the relationships between stations, their influence on the network, their connectivity in terms of passenger transport and freight transport.

We further perform a brief analysis for 17 railway zones of the country, identifying most significant stations and routes among each of the zones.

**Introduction**

Railways provide the cheapest and most convenient mode of passenger transport both for long distance and suburban traffic. For any country railways are one of the primary sources to carry goods and material. Indian railways manages the fourth largest national railway system in the world by size, with a total route length of 126,511 km (78,610 mi) as of 31 December 2021. This proves the importance for railway network analysis in India. Our project can help identify most significant stations and routes which could further help railway officials to maintain the network properly. Our project also helps to classify railway stations and routes according to significance level. We also aim to analyze different railway zones and corridors in the country

**Literature Review:**

### [1] [Statistical analysis of the Indian railway network: A complex network approach](http://www.facweb.iitkgp.ac.in/~niloy/PAPER/Polonica_IRNpaper.pdf): S. Ghosh, A. Banerjee, N. sharma

They analyzed IRN as a weighted complex network of stations, where the edge-weights represent the amount of traffic between two stations.

The major stations (high-degree nodes) tend to be linked among themselves and most of the traffic in the IRN flows among these high-degree nodes. High strengths are limited to two specific regions — in the states of Uttar Pradesh

and western parts of Madhya Pradesh.

The analysis was done on a very old dataset (2011), and based on 2 factors only, strength, degree of stations.

[2] [Complex Network Analysis of Indian Railway Zones](https://arxiv.org/pdf/2004.04146.pdf): Nikhil Rajput, Harshit Arora, Bhavya Grover, Piyush Badola

They compared zonal networks based on self links, in-degree, out-degree of zones, inter and intra zonal connectivity, betweenness, cliques, based on the passenger flow from one zone to another.

[3] [Empirical analysis of dependence between stations in Chinese railway network](https://doc.rero.ch/record/12564/files/zhou_ead.pdf): Yong-Li, Tao Zhou, Jian-Jun, Jian Wanga, Da-RenHe

A new metric is proposed to quantify the dependence between pairs of stations, which is shown to follow a shifted power-law distribution. In addition, we compare the resource-allocation method and the well-known multiple-edge method, and the results indicate that our proposed method is more reasonable.

Resource Allocation method, Statistics of Dependence (weight matrix)

**Proposed Work**

1. **Railway Network Analysis:** In this module we will be analyzing the whole network created using our dataset. We’ll use a dataset of Indian Railways with around 100,000 records. It contains information such as train names, ids, stations names, source/ destination stations, arrival/departure time etc. Based on this information, we plan to firstly perform an analysis of stations based on **measures of centrality, find degree (indegree and outdegree distribution) , betweenness, closeness, significance, clustering coefficient, and cliques**. These measures will help us achieve our first goal and also give a good idea about the network. These measures will be further used in the project.
2. **In-depth study of each of the 17 railway zones:** Since each zone has distinct connections, features, we identify the most significant points in each zone, suggest how connectivity in weak zones can be improved (by adding nodes/ stations).
3. Find **small stations (nodes) of high significance** but poor infrastructure, this will help the railway officials to maintain them properly.

**Modules:**

1. Data Extraction:

The first module involves searching for datasets. We tried different comprehensive datasets involving the train network, connections based on train destination, and the degree of stations as nodes. The final dataset chosen has data of ~70,000 records of trains consisting, train number, train name, station code, station name, Departure time, distance, Source station code, destination station code. This meta data of the Indian Railway network is being utilized to form different sub-datasets keeping stations as nodes, and trains traveling between 2 stations as edges. We are also forming Hypergraphs based on an edge connecting 2 stations.

1. Data Cleaning:

The data involved some N.A values, the first step in pre processing was cleaning the data. After preparing a proper annotated dataset we move on to the next module.

1. Generate Sub data from metadata:

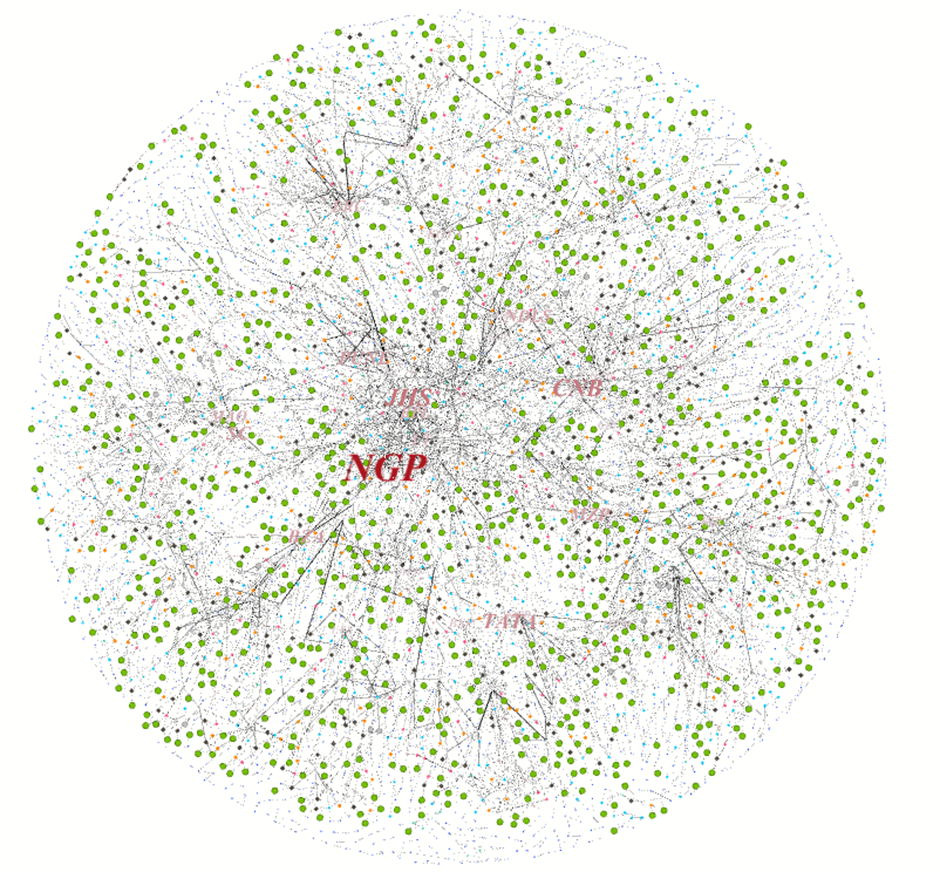
We use a python script to generate sub datasets, such as weight to station, weight to trains, simple graph metrics, page graph metrics, hypergraph metrics. These scripts are used to calculate various measures of centrality, generate hypergraphs, as discussed in the next module.

1. Perform analysis using Gephi:

We will be using Gephi for visual network analysis, Gephi is an open source visualization software. We will be visualizing a simple and a hyper network using Gephi. Metrics such as betweenness centrality, diameter, modularity, degree distribution, clustering coefficient and eigenvector centrality for both simple and hyper graphs will be calculated and visualized using this software. Later we also plan to find page rank.

**Simple Graph:**

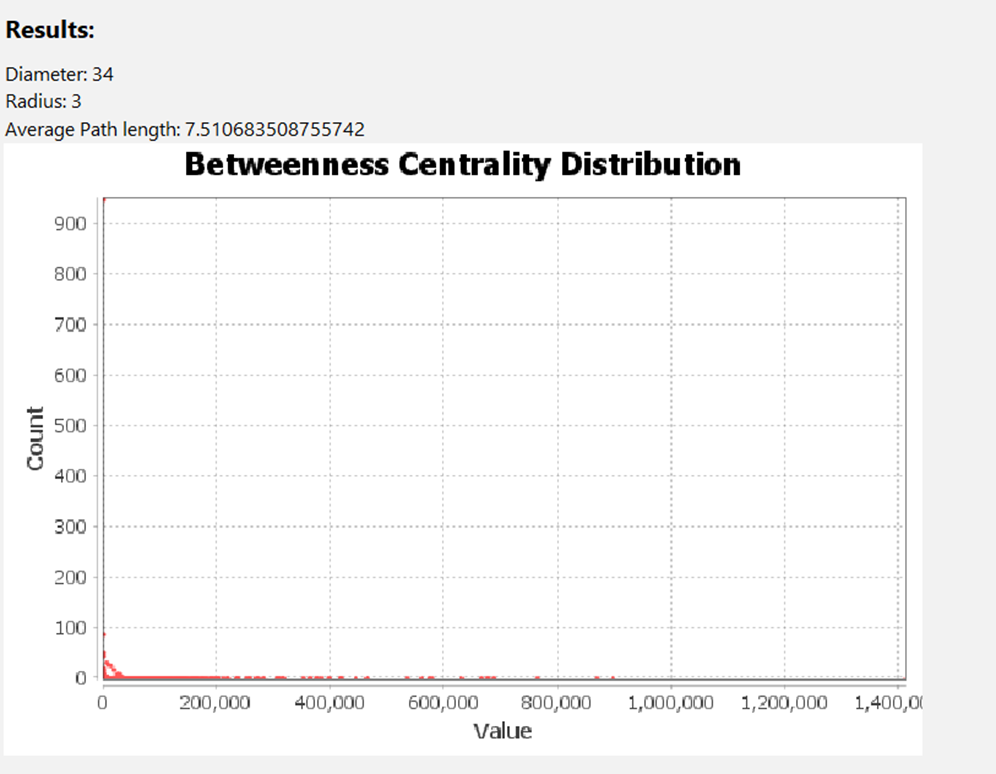
For this project we did a network analysis of Indian railways using various measures and metrics. Railway networks have been modeled as simple graphs and hypergraphs and both kinds of graphs have been analyzed. In simple graphs, stations are represented as nodes and two nodes have an edge if they are connected by at least one train. In hyper graphs, a set of nodes are connected if they have a common train which passes through all these stations. For analysis, the projection of hyper graphs as a simple graph is considered in which the trains are nodes and two nodes are connected if they have a station in common.



**Analysis of simple graph**

1. **Betweenness Centrality:**

The betweenness centrality for each node is the number of shortest paths that pass through the node. In other words, it measures how often a node appears on shortest paths between nodes in the network. Maximum betweenness centrality is observed for NGP(Nagpur). This implies that Nagpur is a very important station as many shortest paths pass through it.

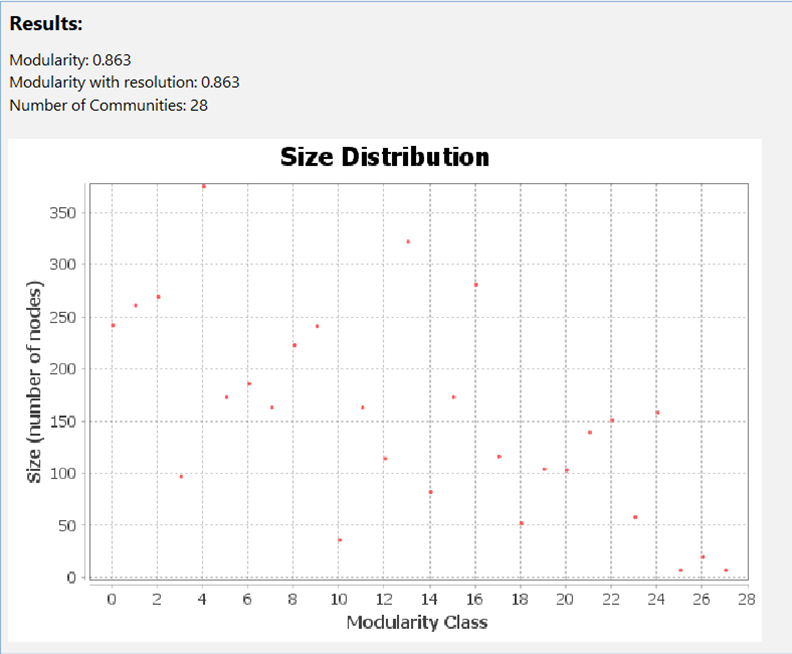


1. **Diameter:**

The greatest geodesic distance between any pair of vertices is the diameter of the graph. Diameter of the graph is found to be 34. This implies that the maximum number of stations to be crossed in order to travel between any two stations in India is 34.

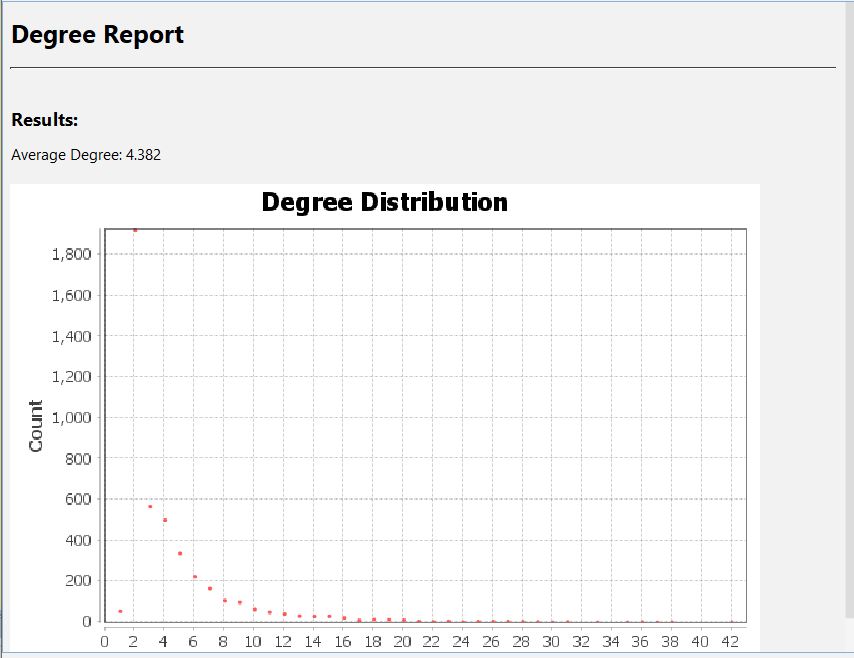
1. **Modularity (Community detection):**

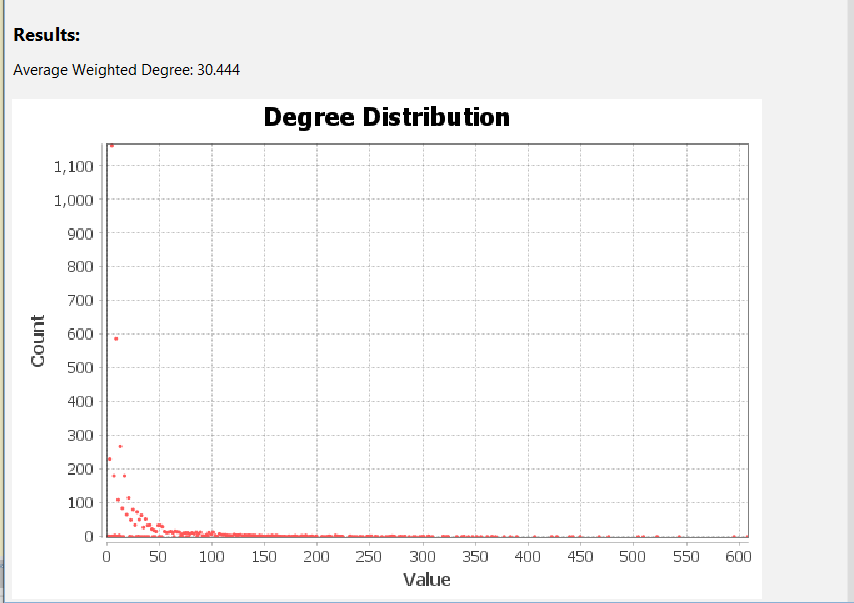
The number of communities have been found out on the basis of modularity. We used the implementation of an algorithm available in Gephi. The number of communities have been found out to be 28 and the modularity is observed to be 0.863.



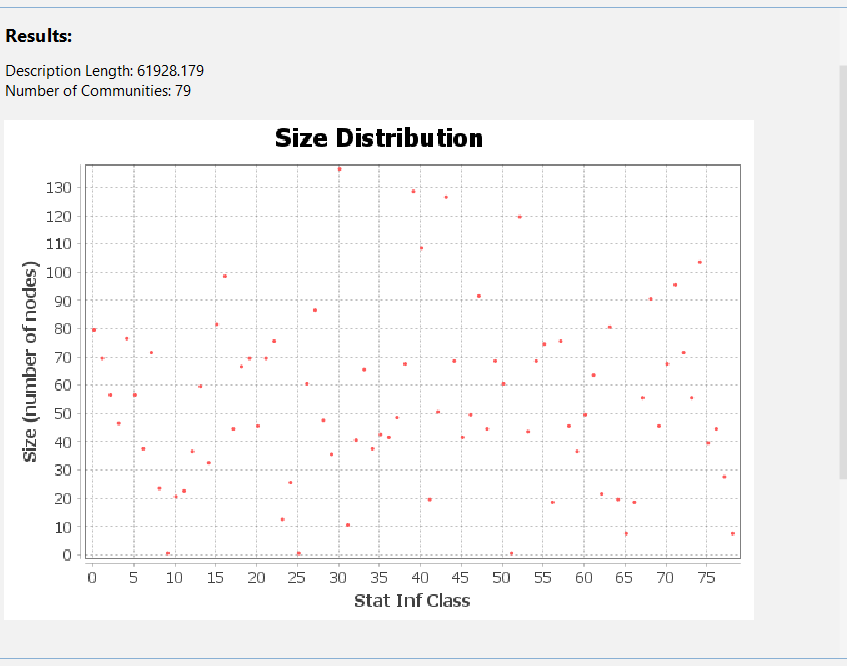
1. **Degree distribution:**

Degree of a node in a graph is the number of edges the node has to other nodes. The degree distribution of the network is further defined as the fraction of the nodes in the network with degree k. For our graph we found that the average degree came out to be 4.382. Nodes having degree between 2 and 4 were the majority. This implies that on an average a station in india is connected to 4 other stations.



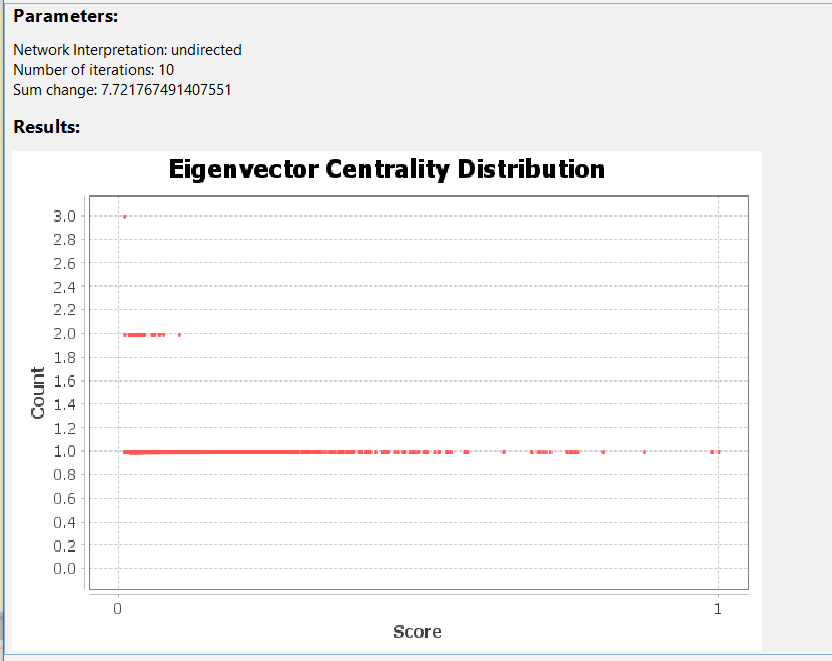


**Statistical Inference Report:**



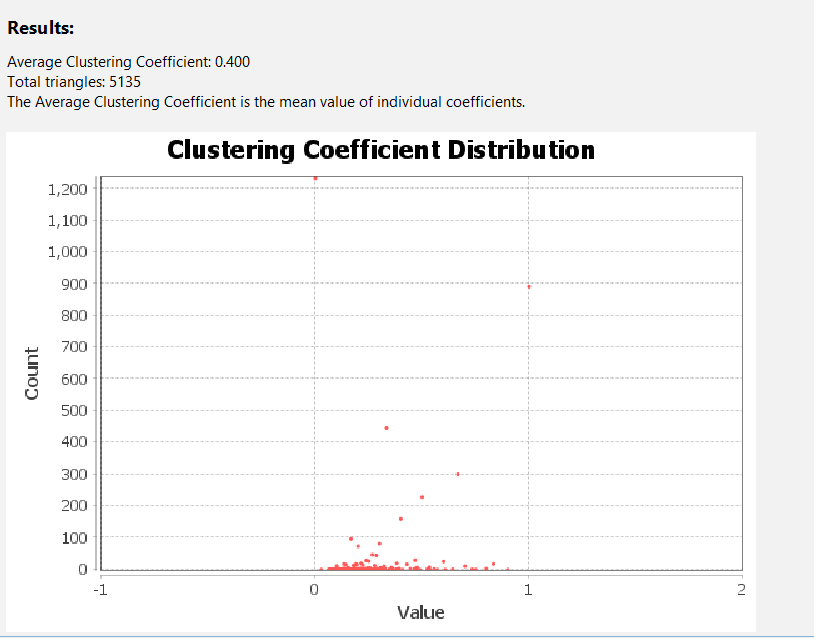
**Eigenvector Centrality:**

Eigenvector Centrality is an algorithm that measures the transitive influence of nodes. Relationships originating from high-scoring nodes contribute more to the score of a node than connections from low-scoring nodes. A high eigenvector score means that a node is connected to many nodes who themselves have high scores. The eigenvector centrality distribution for our dataset is given below:

****

**Clustering Coefficient Distribution:**

Clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. For our graph the average clustering coefficient came out to be 0.4. The clustering distribution is given below:

****

**Results and Discussion**

In the end we aim to find :

* The top 10 Most significant stations will be identified based for both passenger and freight transport.
* For each of the 17 zones, the most significant will be identified.
* 5 small stations which serve important purposes/ are significant in the network but lack infrastructure/ could be scaled for more involvement will be identified.
* We will try to come up with our own metric, and use other measures of centrality and other parameters to compare networks.

**References**

1. Ghosh, Saptarshi & Banerjee, Avishek & Agarwal, Sanket & Ganguly, Niloy & Bhattacharya, Saurav & Mukherjee, Animesh. (2011). Statistical analysis of the Indian Railway Network: A complex network approach. Acta Physica Polonica B, Proceedings Supplement. 4. 10.5506/APhysPolBSupp.4.123.
2. Rajput, Nikhil & Badola, Piyush & Arora, Harshit & Ahuja, Bhavya. (2020). Complex Network Analysis of Indian Railway Zones. Physics and Society, arXiv:2004.04146
3. Yong-Li Wang, Tao Zhou, Jian-Jun Shi, Jian Wang, Da-Ren He,

Empirical analysis of dependence between stations in Chinese railway network, Physica A: Statistical Mechanics and its Applications.

1. Jia Feng, Xiamiao Li, Baohua Mao, Qi Xu, and Yun Bai. Weighted complex network analysis of the Beijing subway system: Train and passenger flows. Physica A: Statistical Mechanics and its Applications, 474:213–223, 2017.
2. Sara Nadiv Soffer and Alexei Vazquez. Network clustering coefficient without degree-correlation biases. Physical Review E, 71(5):057101, 2005.
3. Zhongzhi Xu; Qingpeng Zhang; Dingjun Chen; Yuxin He; IEEE 2019; Characterizing the Connectivity of Railway Networks

# Haitao Hu; Haidong Tao; Frede Blaabjerg; Xiongfei Wang; Zhengyou He; Shibin Gao; IEEE 2017; Train–Network Interactions and Stability Evaluation in High-Speed Railways

1. Sai Nageswar Satchidanand; Siddharth Kumar Jain; Amit Maurya; Balaraman Ravindran; IEEE 2014; Studying Indian Railways Network using hypergraphs.