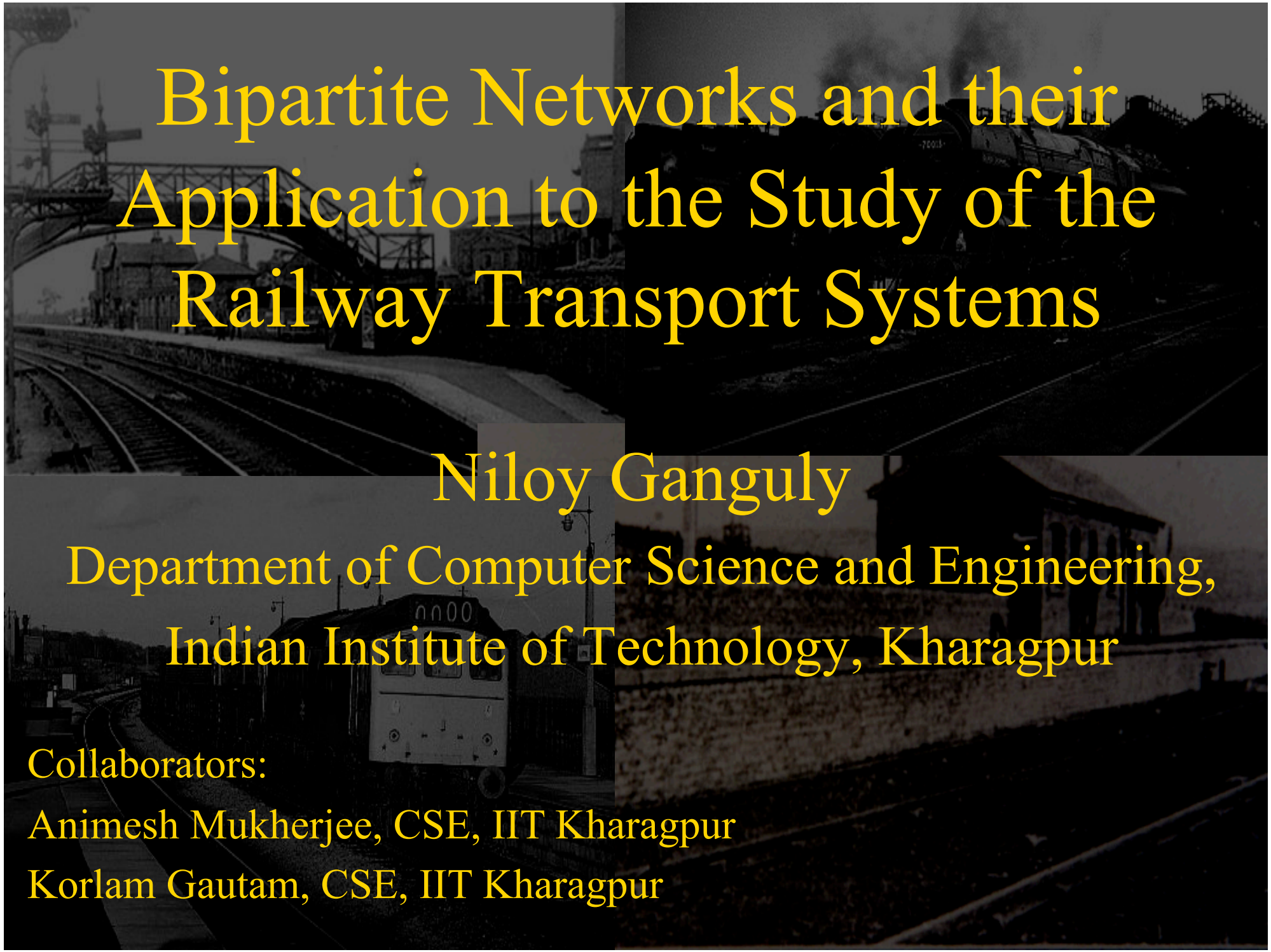


Bipartite Networks and their Application to the Study of the Railway Transport Systems

Niloy Ganguly

Department of Computer Science and Engineering,

Indian Institute of Technology, Kharagpur

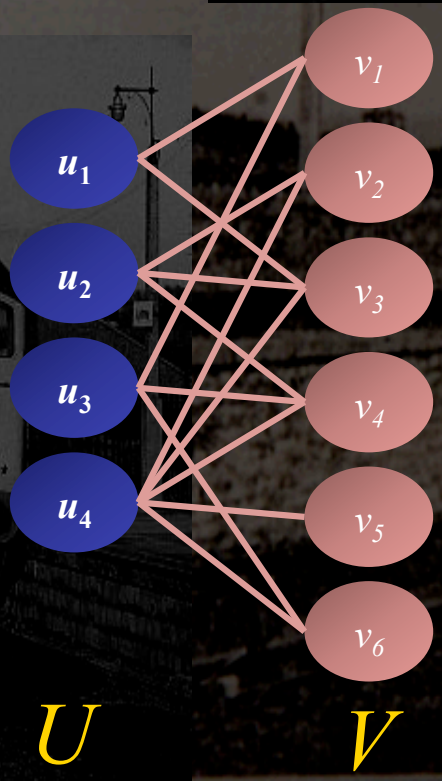
Collaborators:

Animesh Mukherjee, CSE, IIT Kharagpur

Korlam Gautam, CSE, IIT Kharagpur

Bipartite Networks (BNWs)

- A **bipartite network** (or **bigraph**) is a network whose vertices can be divided into two disjoint sets (or partitions) *U* and *V* such that every edge connects a vertex in *U* to one in *V*.



Real-World Examples

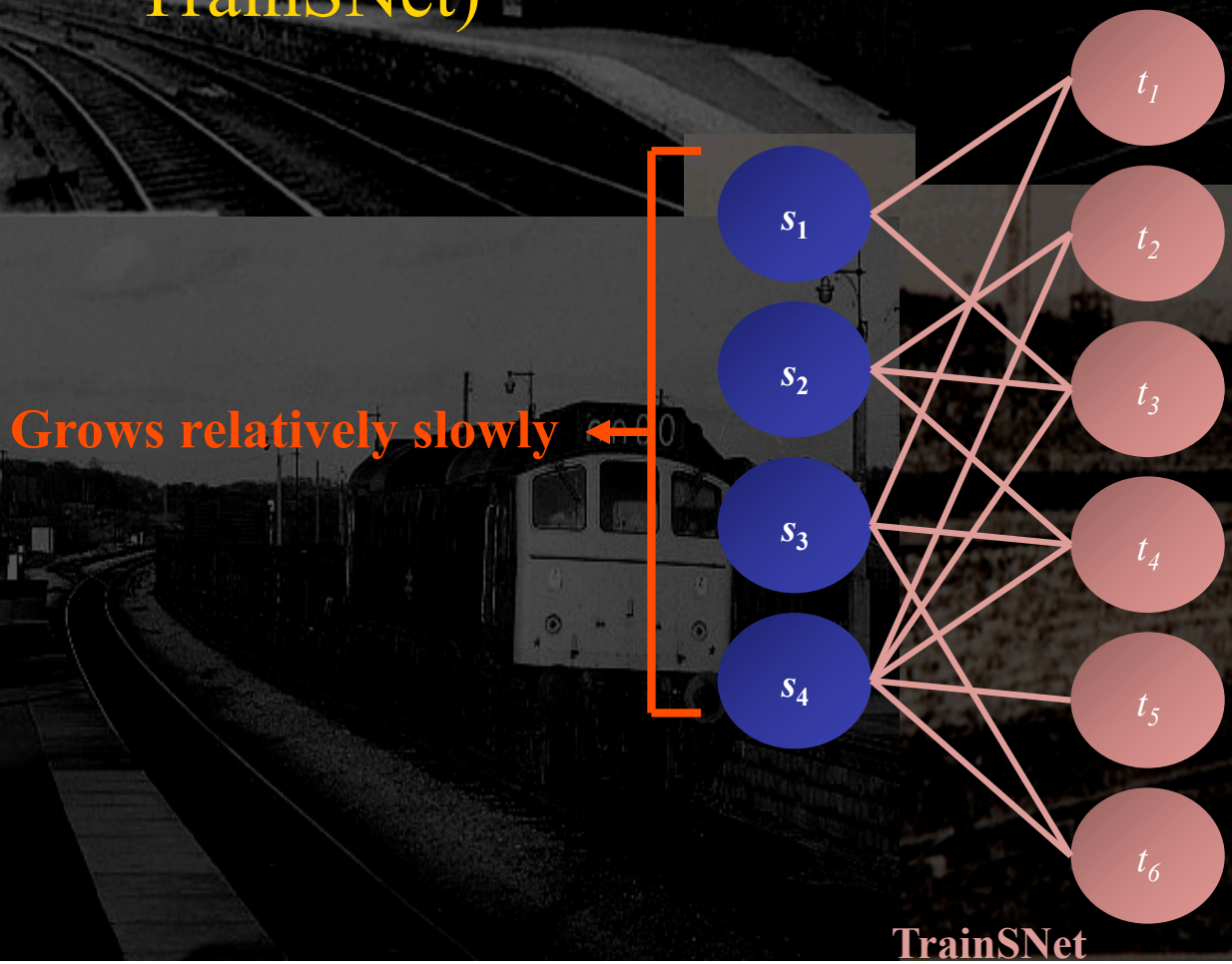
- The movie-actor network where movies and actors constitute the two respective partitions and an edge between them signifies that a particular actor acted in a particular movie.
- The article-author network where the two partitions respectively correspond to articles and authors and edges denote which person has authored which articles
- The board-director network where the two partitions correspond to the boards and the directors respectively and a director is linked by an edge with a society if he/she sits on its board.

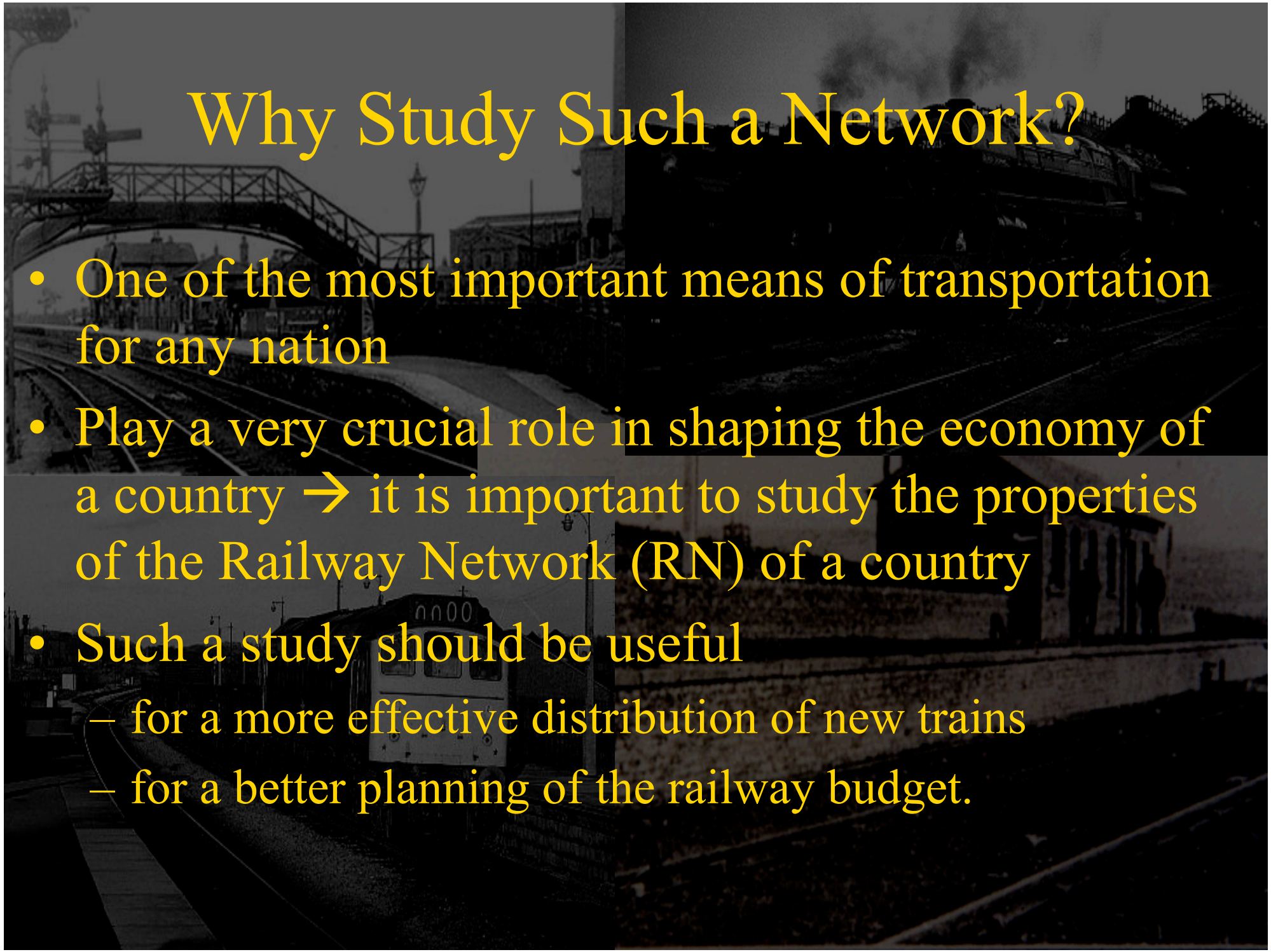
BNWs with one partition fixed

- In all the earlier examples both the partitions of the network grow unboundedly
- αBiNs → A special class of BNWs where one of the partitions does not grow (or grows at a very slow rate) with time
- Examples include
  - Gene-codon network → The two partitions are formed by codons and genes respectively. There is an edge between a gene and a codon if the codon is a part of the gene. Here codon partition remains fixed over while the gene partition grows
  - Word-sentence network → The two partitions are formed by words and sentences in a language respectively. There is an edge between a word and a sentence if the word is a part of the sentence. Here the partition of words grows at a far slower rate than the partition of sentences

Railways as an αBiN

- The two partitions are stations and trains. There is an edge between a station and a train if the train halts at that particular station (Train-Station Network or TrainSNet)





Why Study Such a Network?

- One of the most important means of transportation for any nation
- Play a very crucial role in shaping the economy of a country → it is important to study the properties of the Railway Network (RN) of a country
- Such a study should be useful
  - for a more effective distribution of new trains
  - for a better planning of the railway budget.

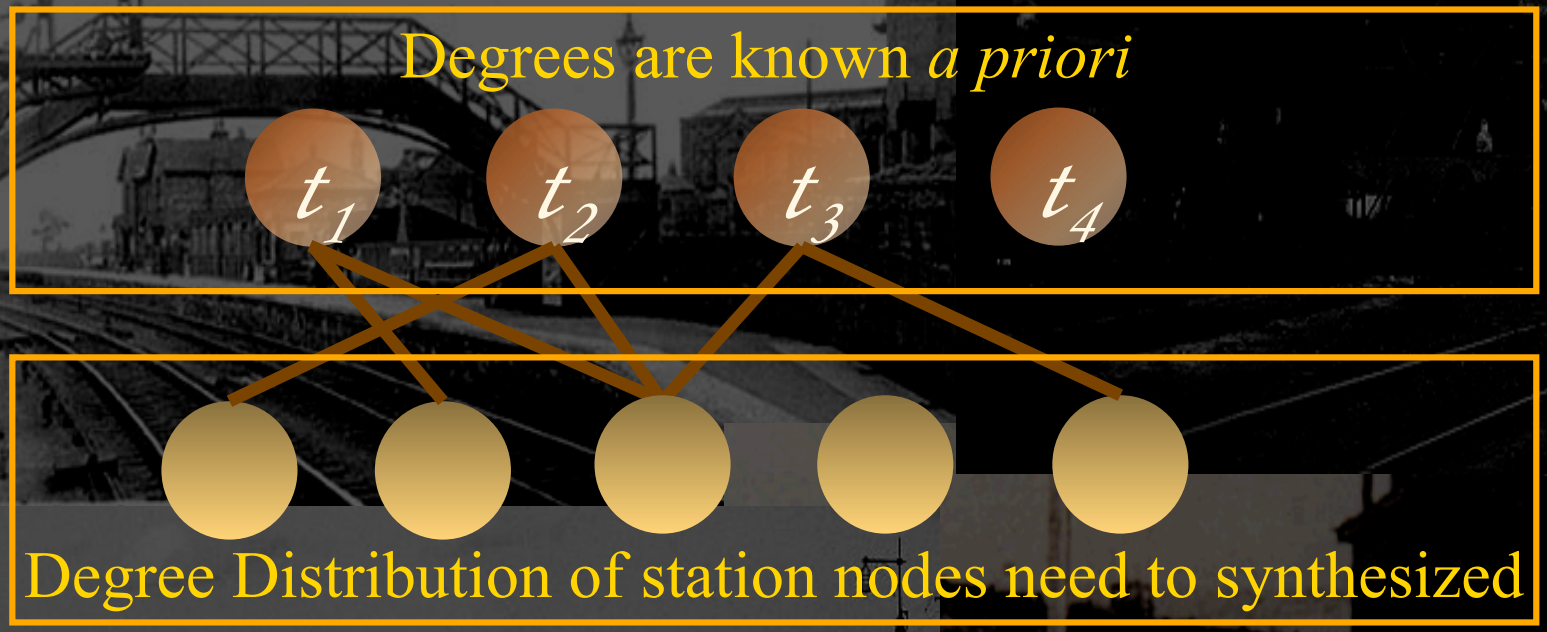
Motivation

- Some studies related to small-world properties (Sen et al. 2003, Cui-mei et al. 2007)
- However, there is no systematic and detailed investigation of various other interesting properties which can furnish a better understanding of the structure of RN
  - Degree distribution of the fixed partition of stations → How does it emerge? What is the growth dynamics?
  - Patterns of hierarchically arranged sub-structures in the network that can provide a deeper understanding into the organization of the railway transport system.
- The primary motivation for the current work is to model RN in the framework of complex networks and systematically explore various important properties

Data Source and Network Construction

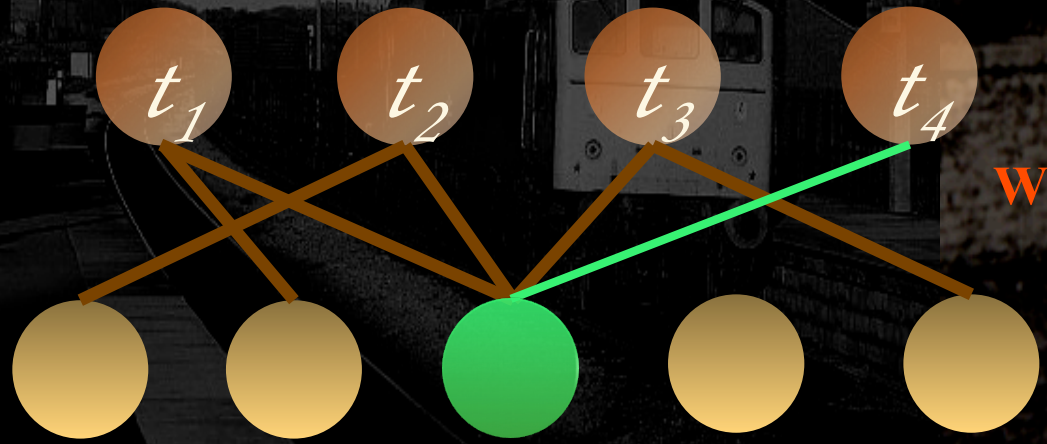
- Indian Railways (IR)
  - The data was manually collected from the http://www.indianrail.go.in, which is the official website of the Indian railways
  - 2764 stations and approximately 1377 trains halting at one or more of these stations
  - TrainSNet_{IR} and StaNet_{IR} constructed from above data
- German Railways (GR)
  - Deutsche Bahn Electronic Timetable CD
  - Had information about 80 stations (approx.) and only the number of direct trains connecting them
  - We could therefore only construct StaNet_{GR}

Growth Model for the Emergence (EPL, 2007)



After step 3

After step 4



Preference component

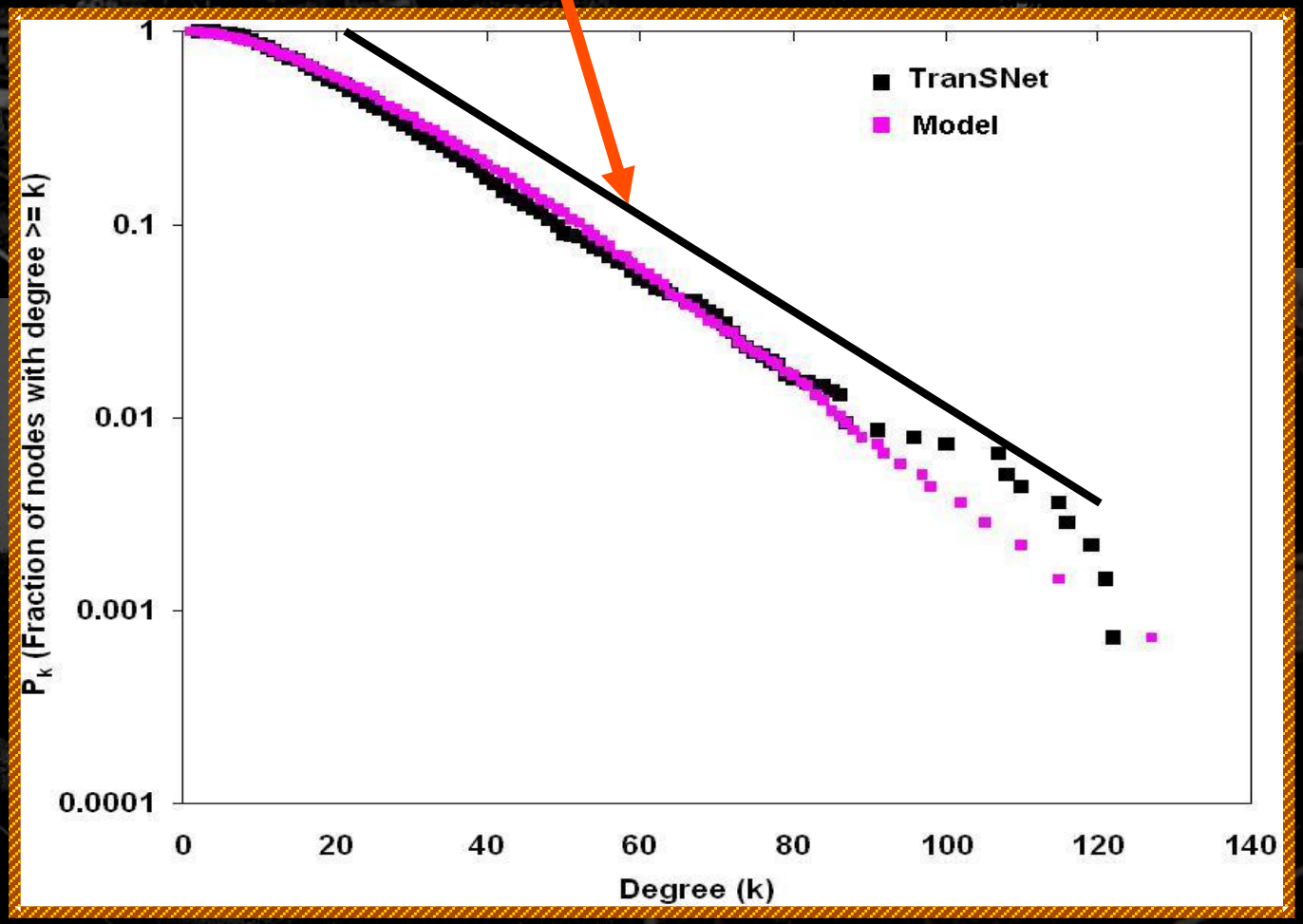
Weight of Preference

$$\frac{y_k + 1}{\sum (y_k + 1)}$$

Degree Distribution of $TrainsNet_{IR}$

Not a Power-Law !! (Fit obtained through least square regression)

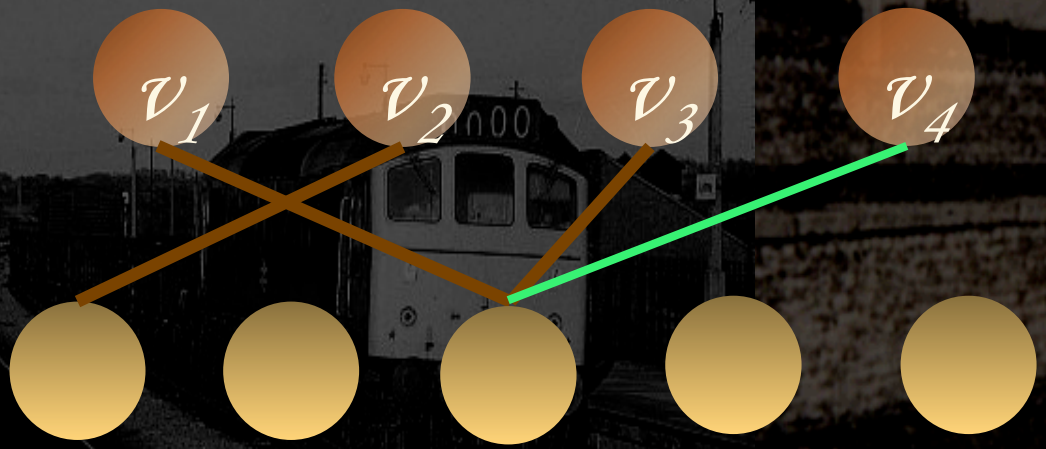
$P_k = 1.53\exp(-0.06)$



Best fit emerges at $\gamma = 0.5$

Theoretical Investigation: The Three Sides of the Coin

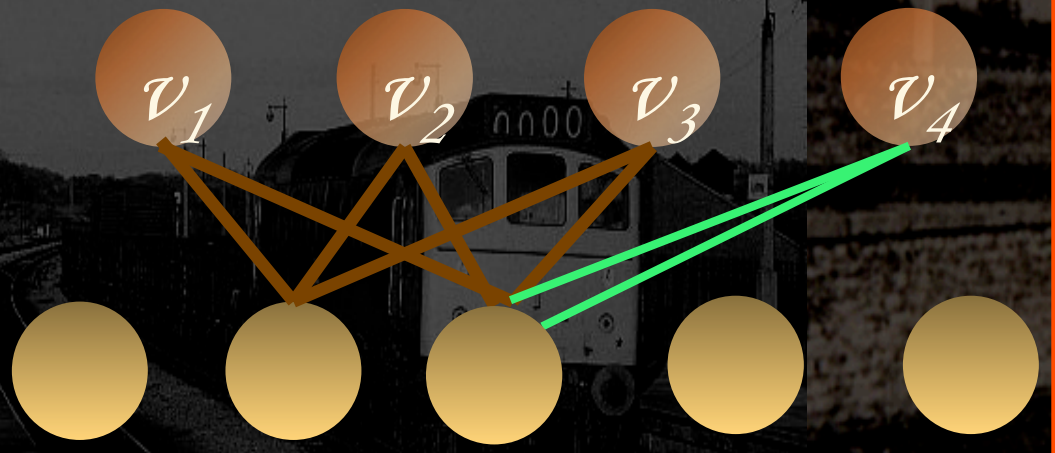
- Sequential Attachment
  - Only one edge per incoming node
  - Exclusive set-membership: Language – {speaker, webpage}, country – citizen



Each node v_i in the growing partition enters with exactly one edge

The Three Sides of the Coin

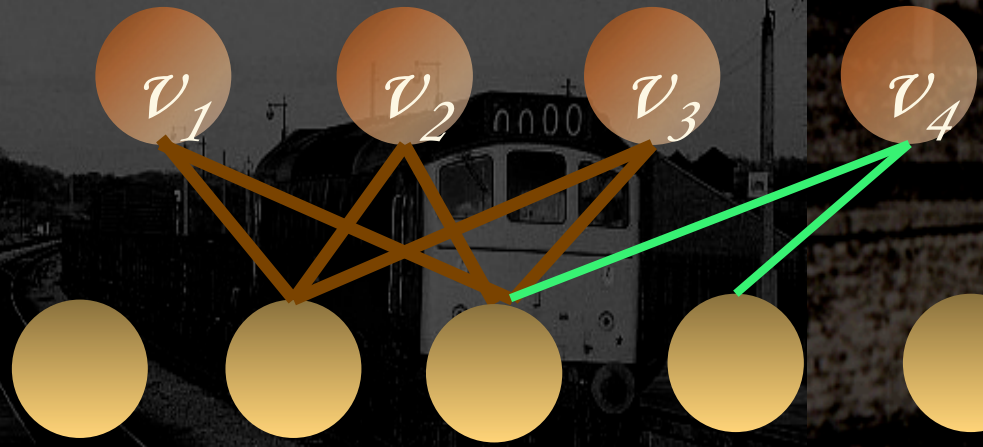
- Parallel Attachment With Replacement
  - All incoming nodes has $\mu > 1$ edges
  - Sequences: letter-word, word-document



Each node v_i in the growing partition enters with $\mu > 1$. A node may be chosen more than once in a step. Parallel edges are possible.

The Three Sides of the Coin

- Parallel Attachment Without Replacement
  - Sets: phoneme-languages, station-train



Each node v_i in the growing partition enters with $\mu > 1$. A node can be chosen only once in a time step. Parallel edges are not possible.

Sequential Attachment

Notations

t – #nodes in growing partition

N – #nodes in fixed partition

$p_{k,t} - p_k$ after adding t nodes

*One edge added per node

Markov Chain Formulation

$$p_{k,t+1} = (1 - \tilde{P}(k,t))p_{k,t} + \tilde{P}(k-1,t)p_{k-1,t}$$

$$\tilde{P}(k,t) = \begin{cases} \frac{\gamma k+1}{\gamma t+N} & \text{for} \quad 0 \le k \le t \\ 0 & \text{otherwise} \end{cases}$$

EPL, 2007

The Hard part

- Average degree of the fixed partition diverges
- Methods based on steady-state and continuous time assumptions fail

Closed-form Solution

$$p_{k,t} = \binom{t}{k} \frac{\prod_{x=0}^{k-1}(\gamma x + 1)\prod_{y=0}^{t-1-k}(N - 1 + \gamma y)}{\prod_{w=0}^{t-1}(\gamma w + N)}$$

$$p_{k,t} \approx (k/t)^{\gamma^{-1}-1}(1 - k/t)^{\eta-\gamma^{-1}-1}$$ where $\eta = N/\gamma$

EPL, 2007

Parallel attachment with replacement

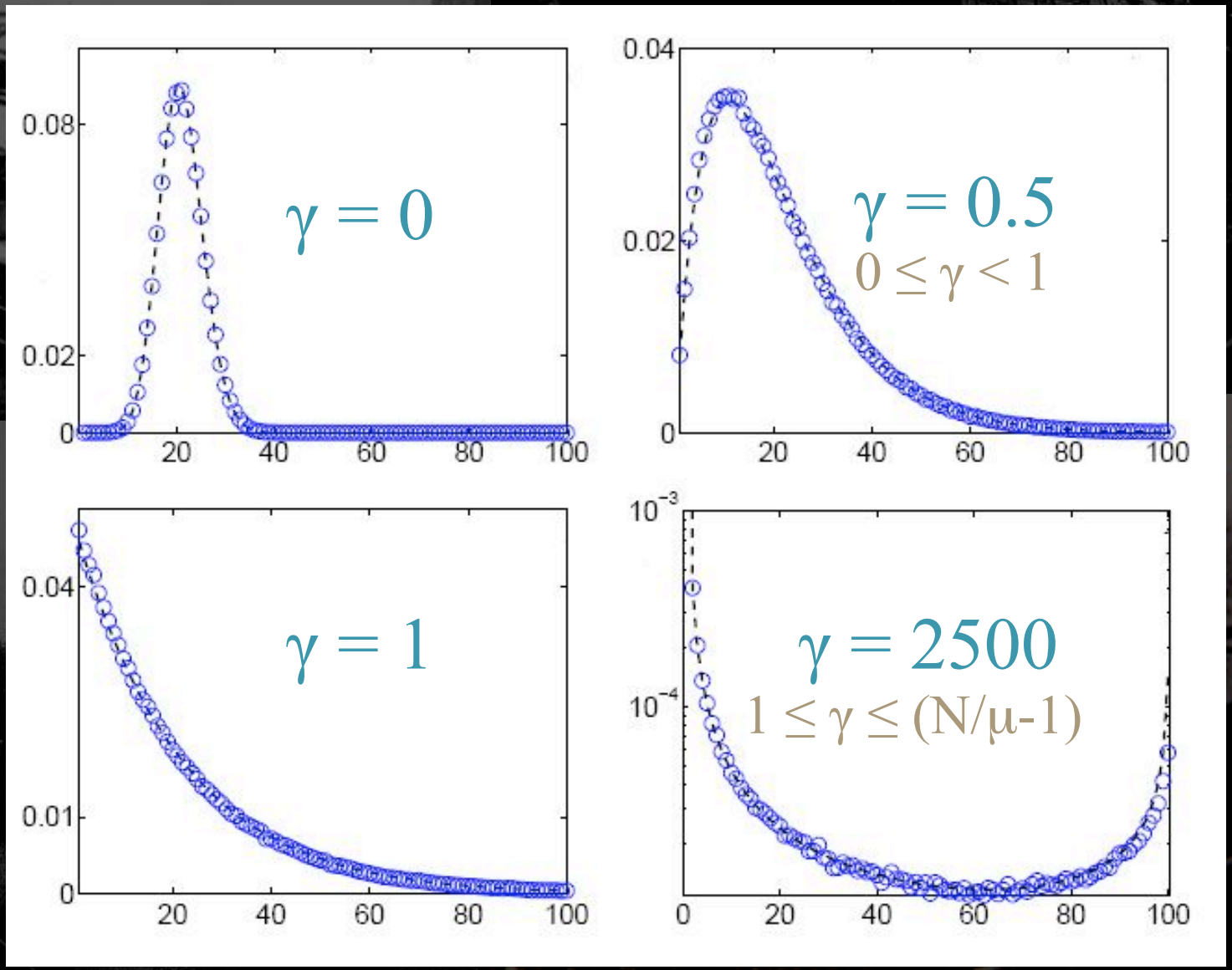
- The number of incoming edges is $\mu > 1$
- For $N >> \mu$ we can use the following approximation:

$$p_{k,t} = \binom{t}{k} \frac{\prod_{x=0}^{k-1}(\gamma x + 1)\prod_{y=0}^{t-1-k}(\frac{N}{\mu} - 1 + \gamma y)}{\prod_{w=0}^{t-1}(\gamma w + \frac{N}{\mu})}$$

- $p_{k,t} \sim \mathrm{B}(k/t;\ \gamma^{-1},\ N/(\mu\gamma) - \gamma^{-1})$

EPL 2007

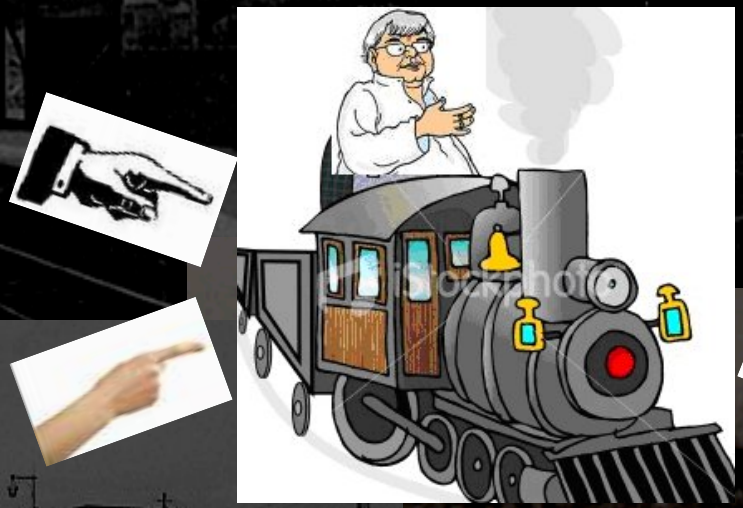
A tunable distribution



EPL, 2007

Implications

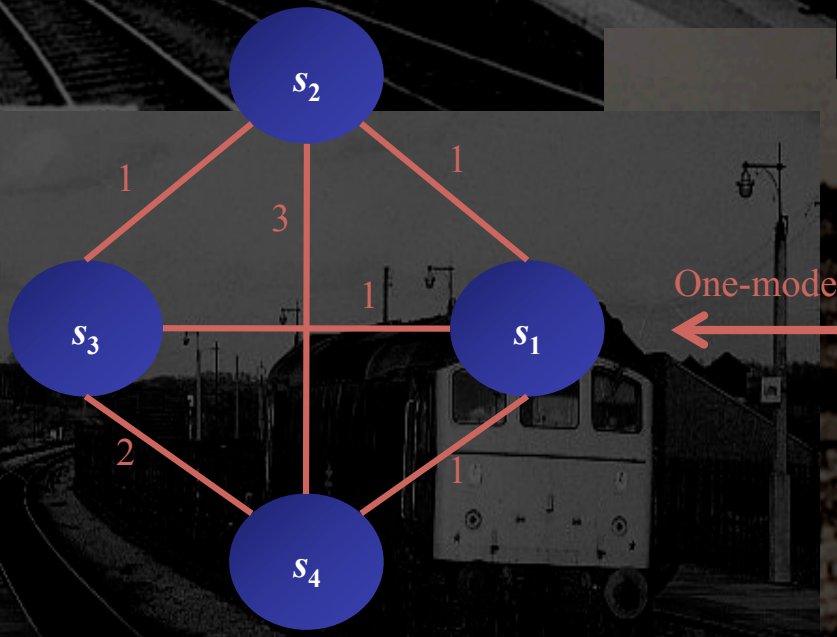
- γ (0.5) is low → preferential attachment does not play as strong a role in the evolution of Indian Railway Network as it does in case of various other social networks
- Possible reasons



  - Arbitrary change in the railway ministry & government → mainly concerned with the connectivity of the native regions of the ministers rather than the connectivity in the global scale.
  - Government has stipulated rail budgets for each of the states (possibly not very well-planned)
  - And if we don't want to blame the ministers → PA leads to a network where failure of a hub (i.e., a very high degree node) might cause a complete breakdown in the communication system of the whole country → discouraged by natural evolution

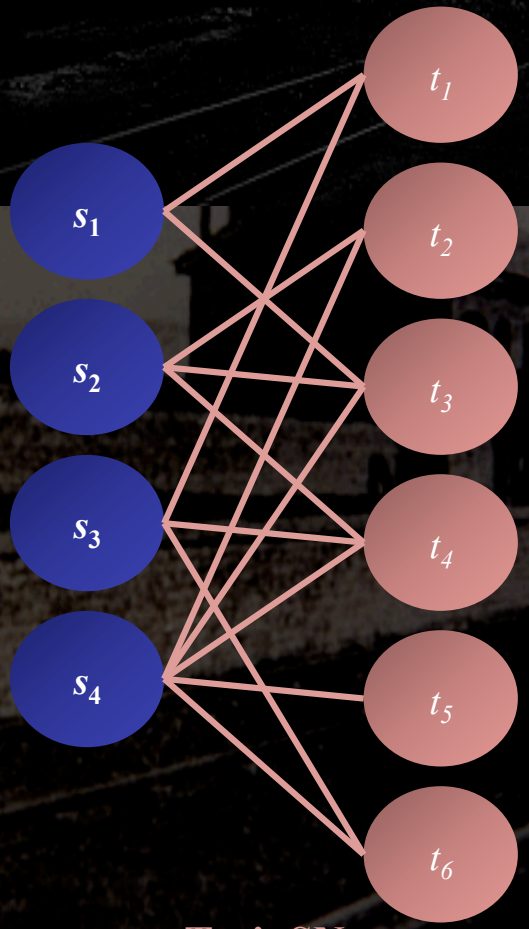
One-Mode Projection of fixed Partition

- **One mode projection onto the nodes of the fixed partition corresponds to a network of stations where stations are connected if there is a train halting at both of them. If there are *w* trains halting at both of them then the weight of the edge is *w*. We call this network StaNet or the Station-Station Network.**



StaNet

One-mode projection



TrainSNet

Small-World Properties

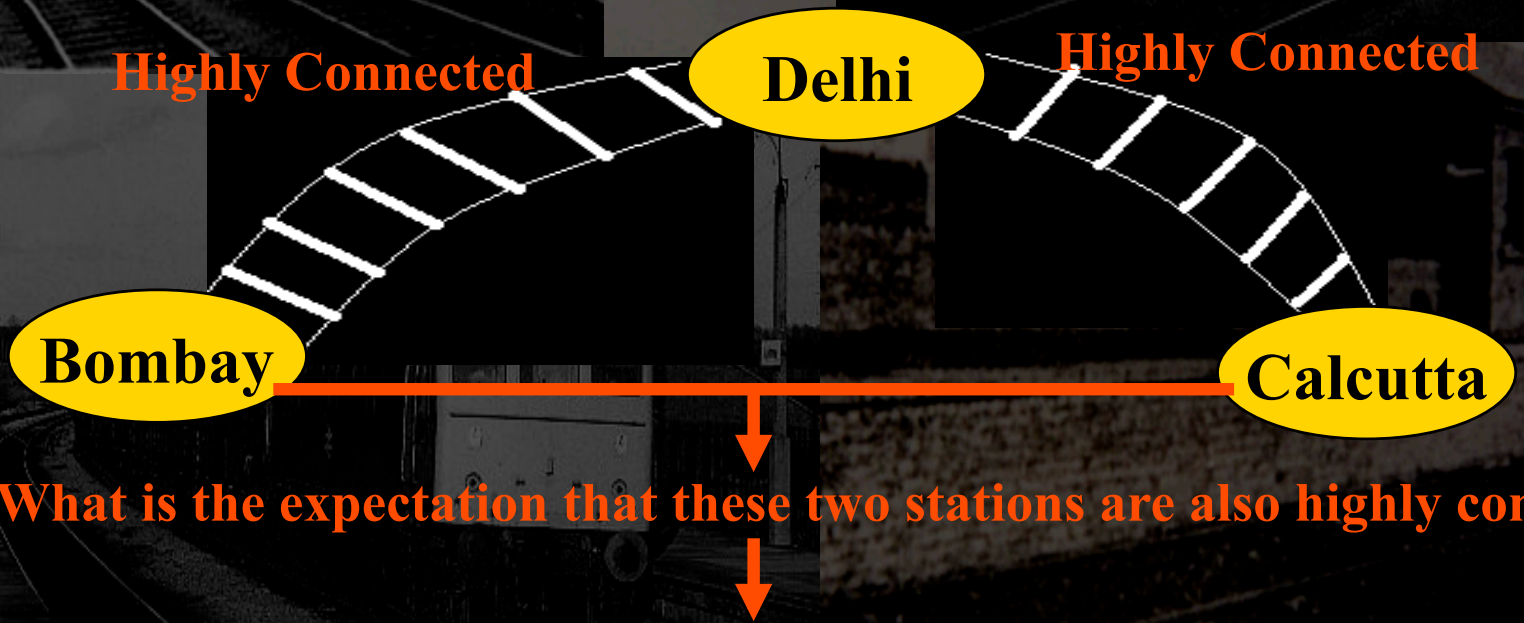
Neighboring stations of a station are also highly connected via direct trains → local connectivity high

Properties	$StaNet_{IR}$	$StaNet_{GR}$
Weighted CC	0.79	0.75
Avg. Path length, Diameter	2.43, 4.00	1.76, 3.00

Any arbitrary station in the network can be reached from any other arbitrary station through only a very few hops.

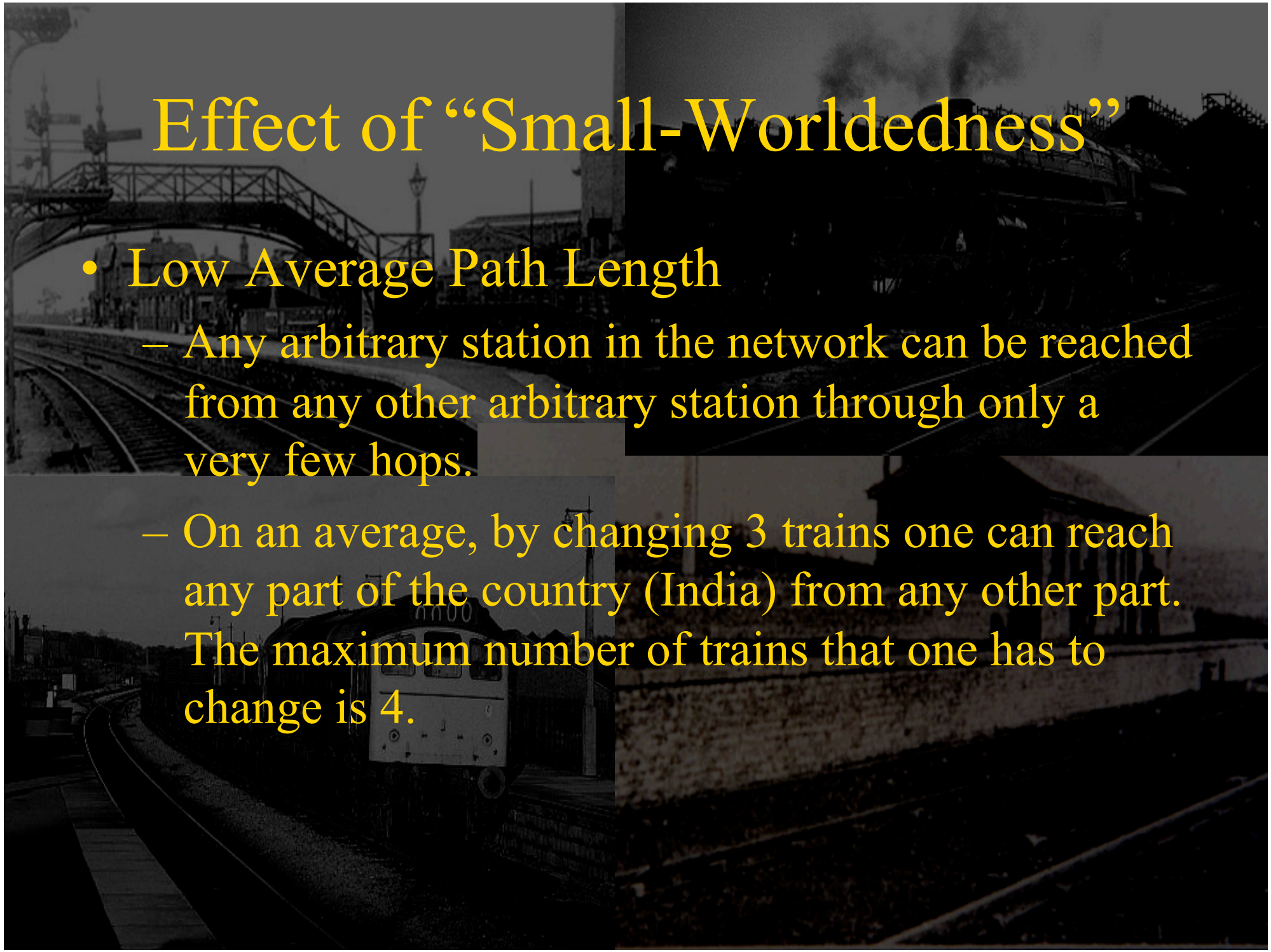
Effect of "Small-Worldedness"

- High Clustering Coefficient
  - Neighboring stations of a station are also highly connected via direct trains



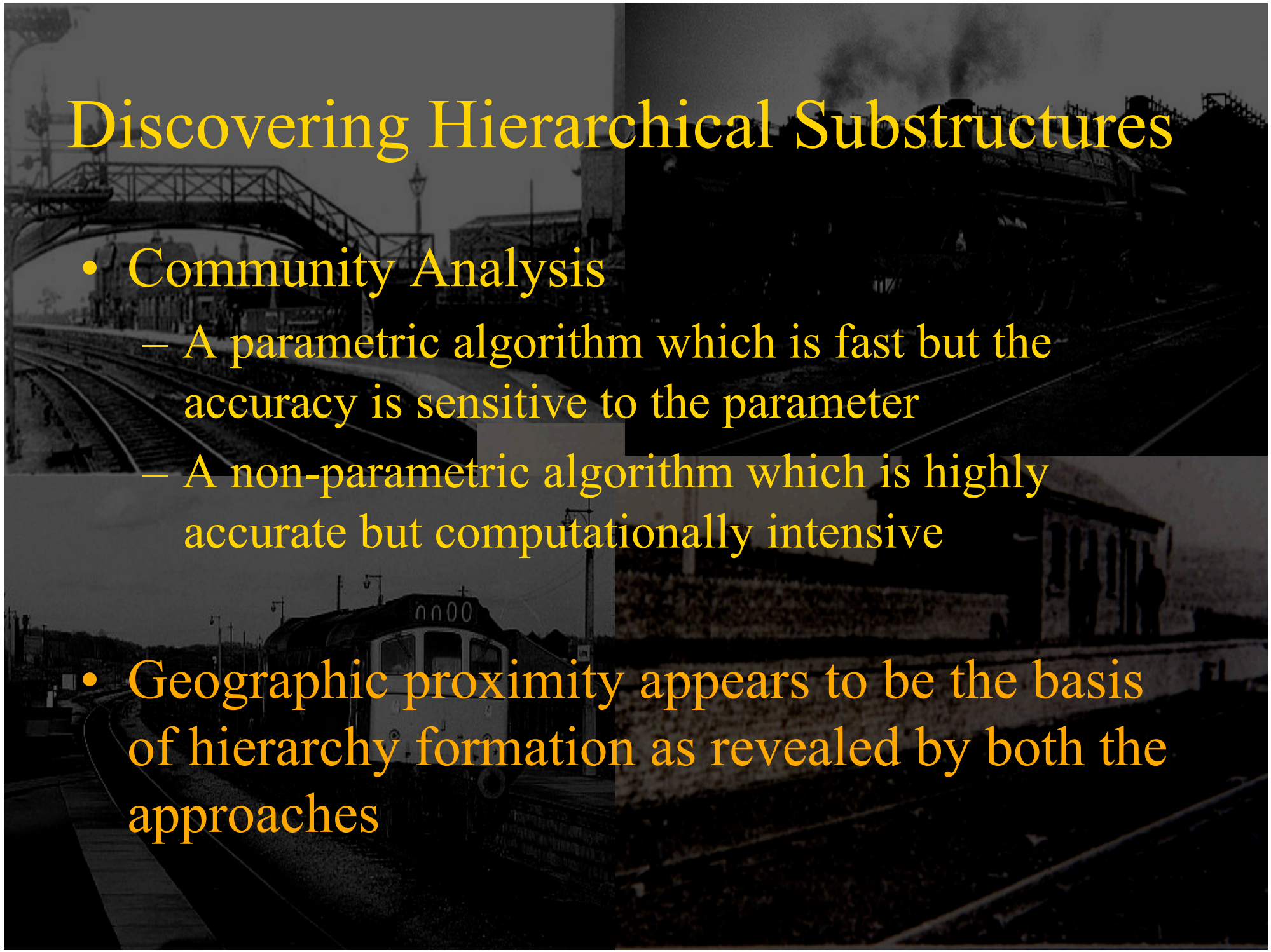
What is the expectation that these two stations are also highly connected?

This expectation is high (high CC) like many other social n/ws (Friends of friends are also friends themselves)



Effect of "Small-Worldedness"

- Low Average Path Length
  - Any arbitrary station in the network can be reached from any other arbitrary station through only a very few hops.
  - On an average, by changing 3 trains one can reach any part of the country (India) from any other part. The maximum number of trains that one has to change is 4.



Discovering Hierarchical Substructures

- Community Analysis
  - A parametric algorithm which is fast but the accuracy is sensitive to the parameter
  - A non-parametric algorithm which is highly accurate but computationally intensive
- Geographic proximity appears to be the basis of hierarchy formation as revealed by both the approaches

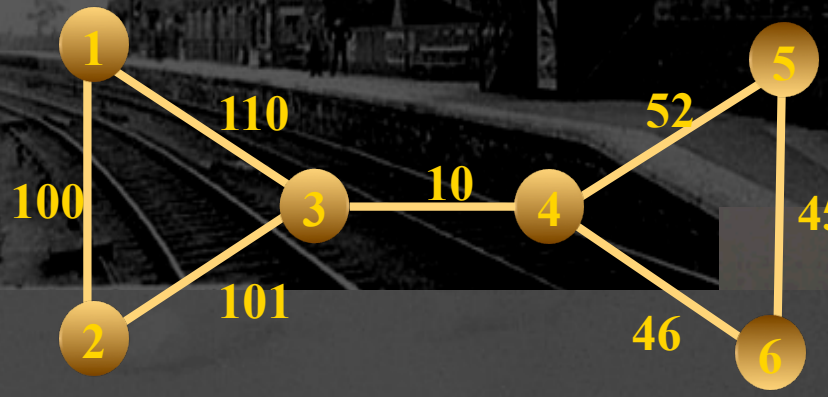
Modified Radicchi et al. Algorithm

- Radicchi et al. algorithm (for unweighted networks) – Counts number of triangles that an edge is a part of. Inter-community edges will have low count so remove them.
- Modification for a weighted network like StaNet
  - Look for triangles, where the weights on the edges are comparable.
  - If they are comparable, then the group of consonants co-occur highly else it is not so.
  - Measure strength S for each edge (u,v) in StaNet where S is,

$$S = \frac{w_{uv}}{\sqrt{\Sigma_{i \in V_c - \{u,v\}} (w_{ui} - w_{vi})^2}} \quad \text{if } \sqrt{\Sigma_{i \in V_c - \{u,v\}} (w_{ui} - w_{vi})^2} > 0 \text{ else } S = \infty$$

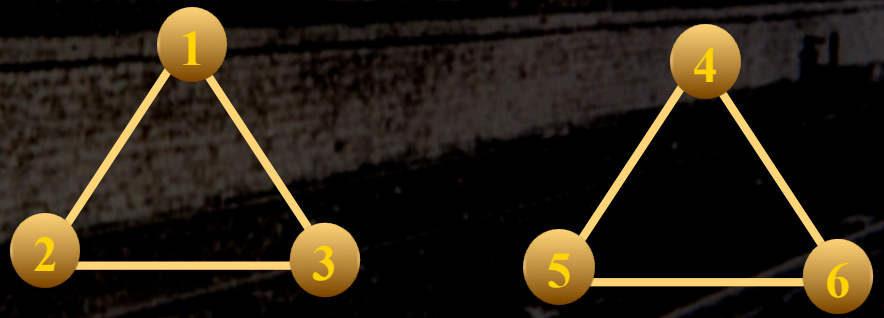
  - Remove edges with S less than a threshold η

The Process



S

η>1



For different values of η we get different sets of communities

Example Communities

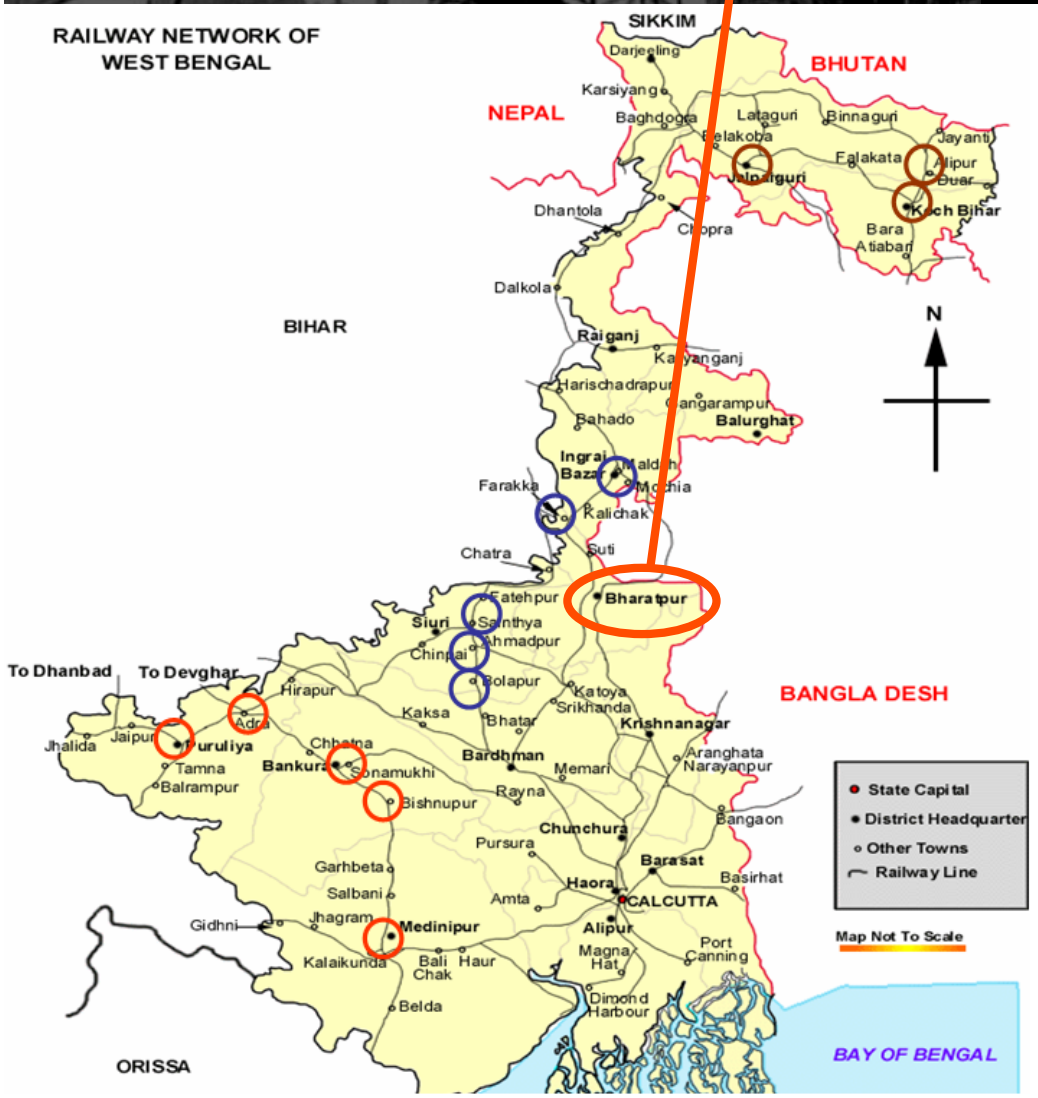
Communities from $StaNet_{IR}$	Regions	η
Adra Jn., Bankura, Midnapore, Purulia Jn., Bishnupur	West Bengal	0.42
Ajmer, Beawar, Kishangarh	Rajasthan	0.42
Abohar, Giddarbaha, Malout, Shri Ganganagar	Punjab	0.50

Parameter to which the results are sensitive

Communities from $StaNet_{GR}$	Regions	η
Bremen, Hamburg, Osnabrück, Münster	North-West Germany	0.72
Augsburg, Munich, Ulm, Stuttgart	Extreme South Germany	0.72
Diasburg, Düsseldorf, Dortmund	Extreme West Germany	1.25

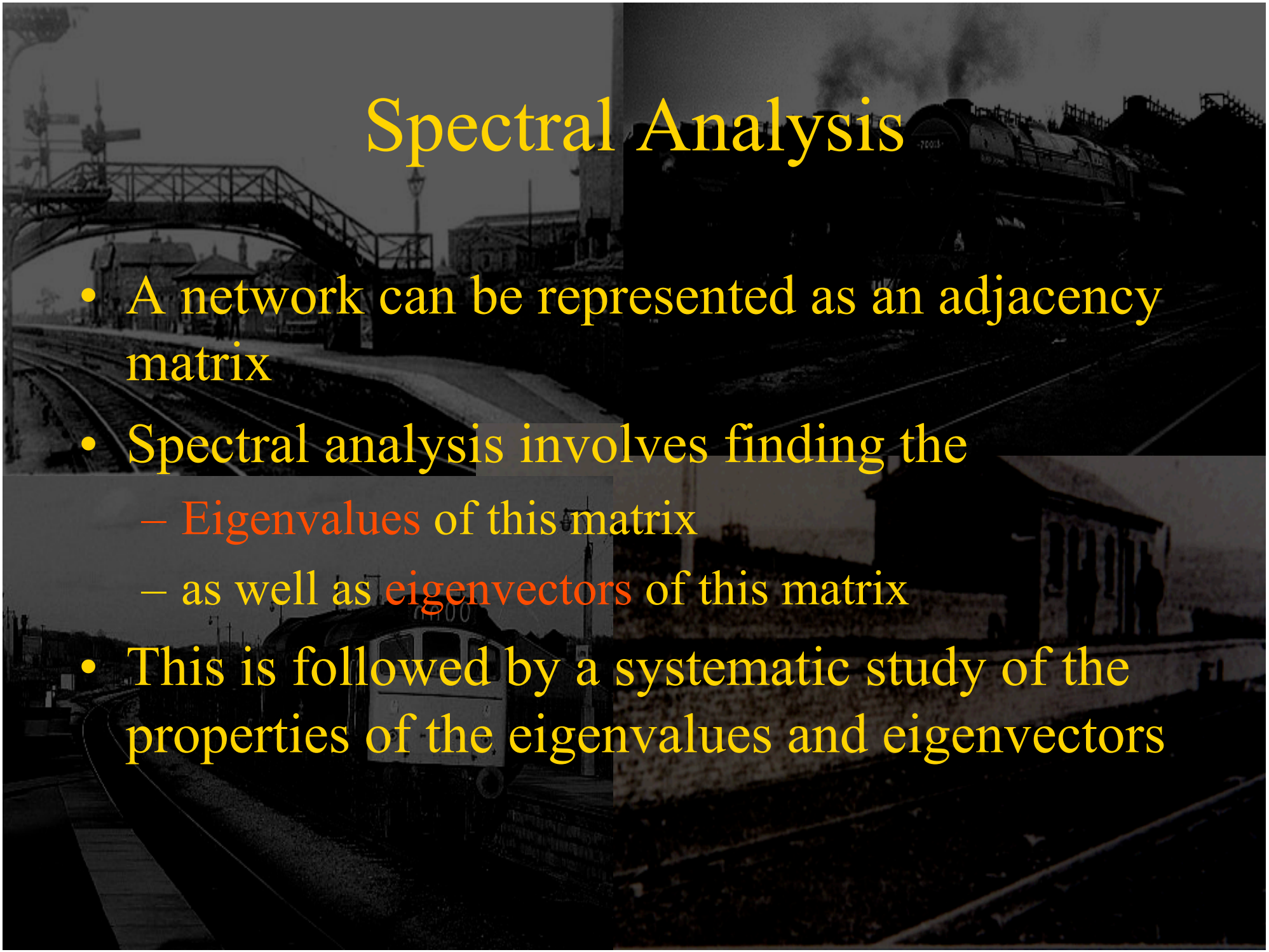
IRN – Communities on Map

Should have been a part of the blue circles → Not much train connectivity with this station though its a junction!!



GRN – Communities on Map

Should have been part of the brown circles



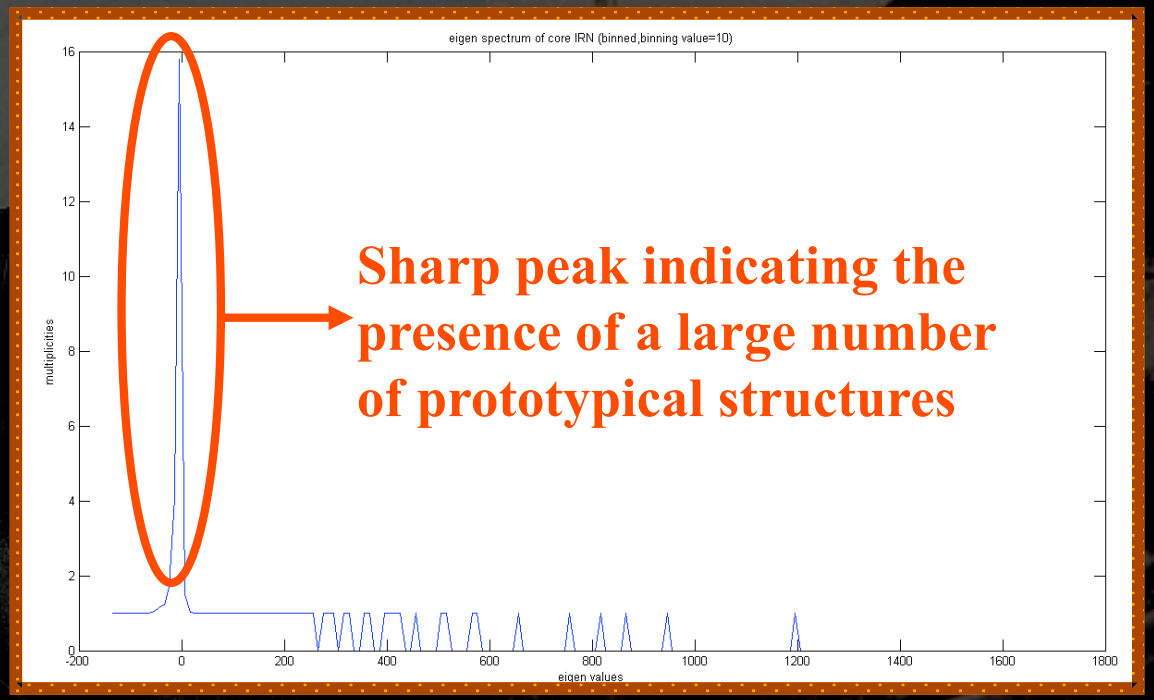
Spectral Analysis

- A network can be represented as an adjacency matrix
- Spectral analysis involves finding the
  - Eigenvalues of this matrix
  - as well as eigenvectors of this matrix
- This is followed by a systematic study of the properties of the eigenvalues and eigenvectors

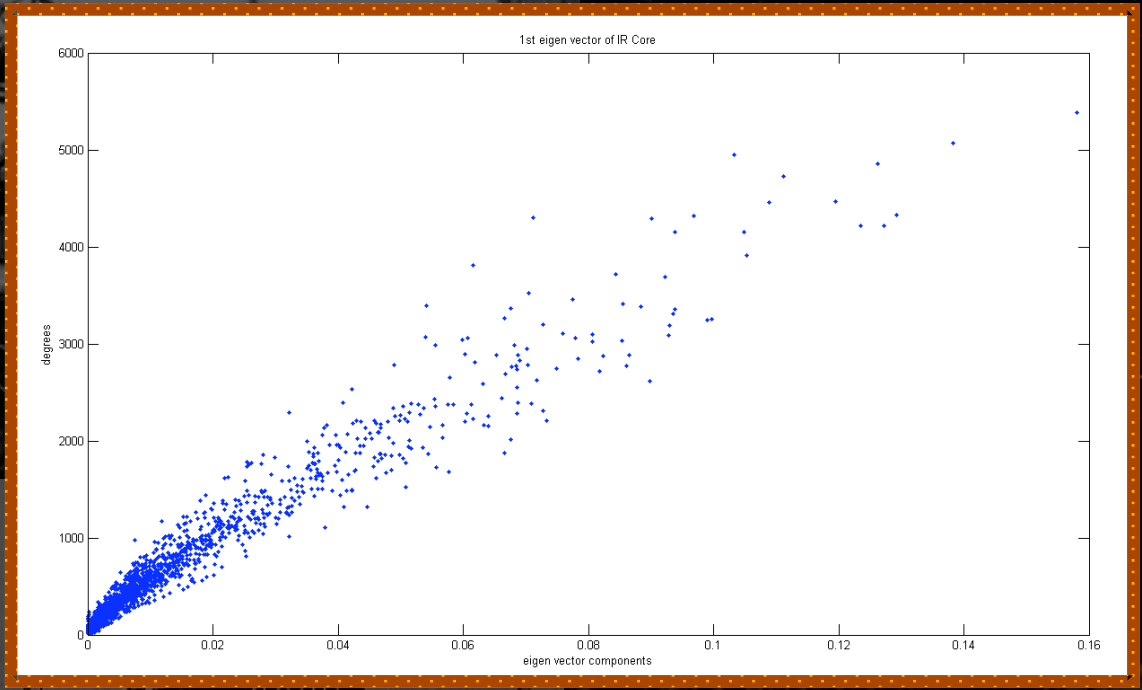
Spectral Analysis

- Systematic study of the eigenvalues and eigenvectors of the adjacency matrix for a network such as StaNet.

Eigenvalues



First Eigenvector (corresponding to the principal eigenvalue)



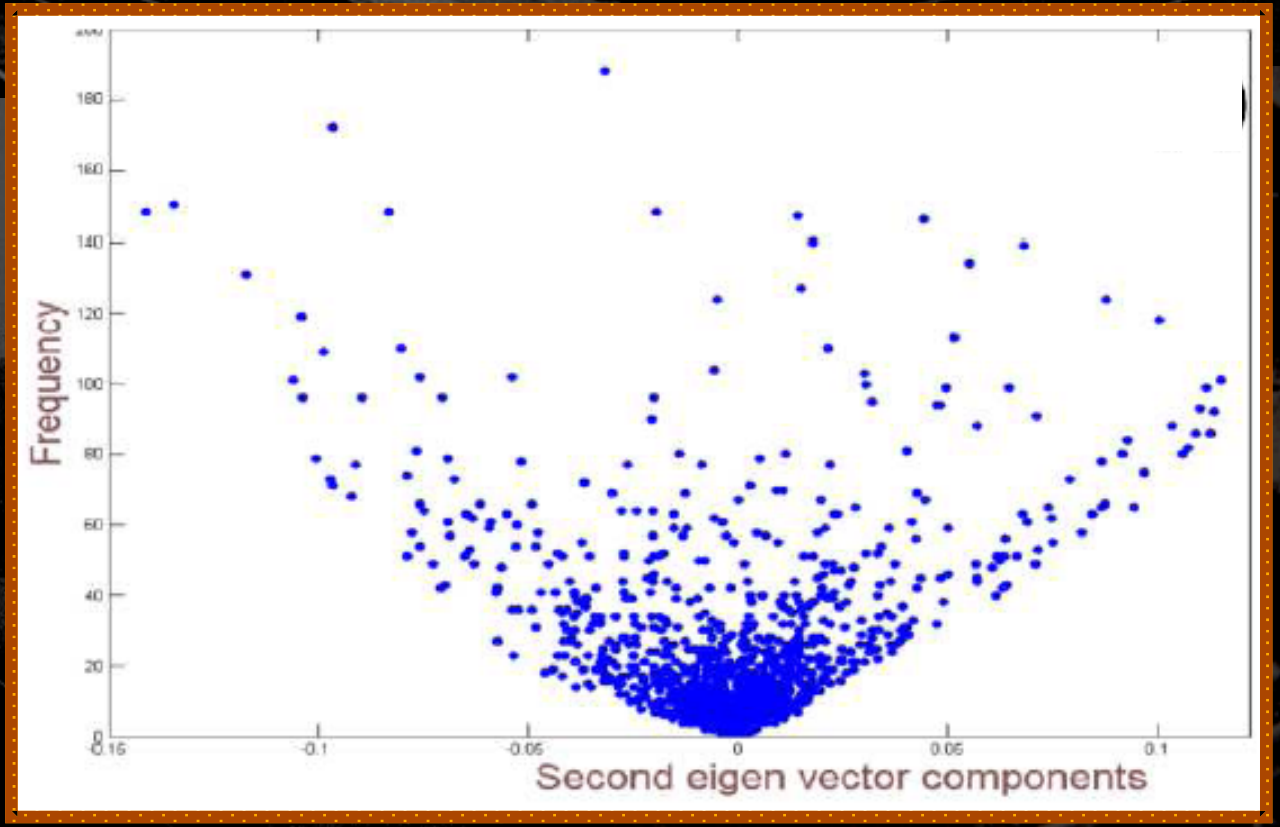
- Perfectly correlated to the degree of the station nodes in TrainSNet / StaNet. The degree of the station nodes in TrainSNet actually denotes the frequency of trains through that station.
- The above result is due to proportionate co-occurrence, i.e., two frequent station nodes also have a large number of trains halting at both of them

Spectral Clustering

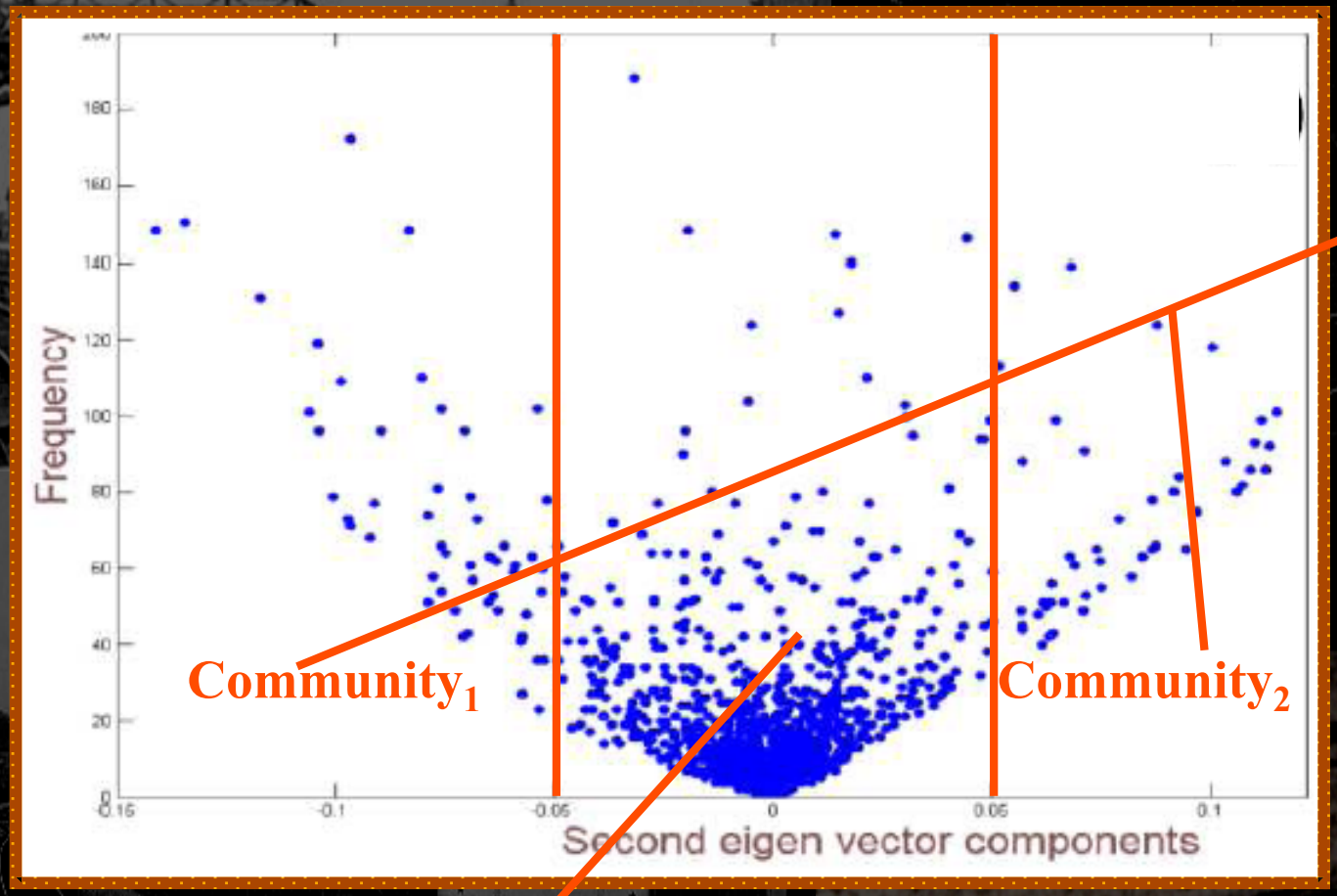
- Partitions nodes into two sets S_1 and S_2 based on the eigenvector corresponding to the second largest eigenvalue
- This partitioning may be done in various ways, such as by taking the median m of the components in v, and placing all points whose component in v is greater than m in S_1, and the rest in S_2. The algorithm can be used for hierarchical clustering by repeatedly partitioning the subsets in this fashion.

Spectral Clustering

- The second eigenvector of the adjacency matrix for a network such as StaNet is known to divide the it into two smaller sub-structures.



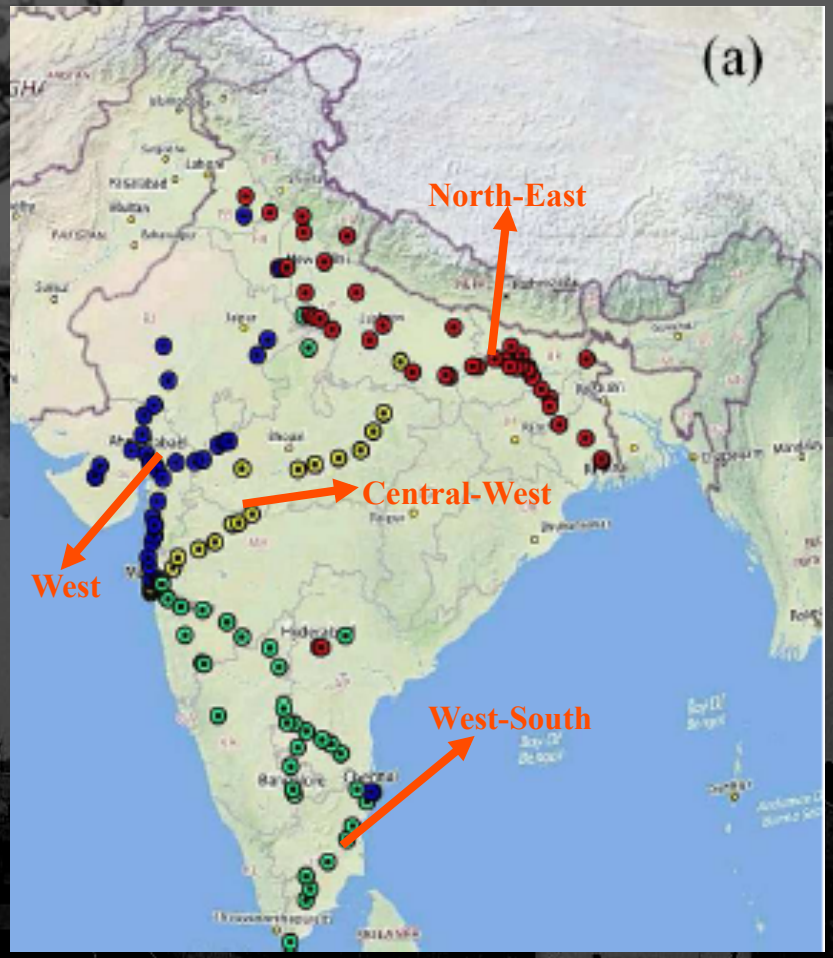
Iterative Spectral Clustering

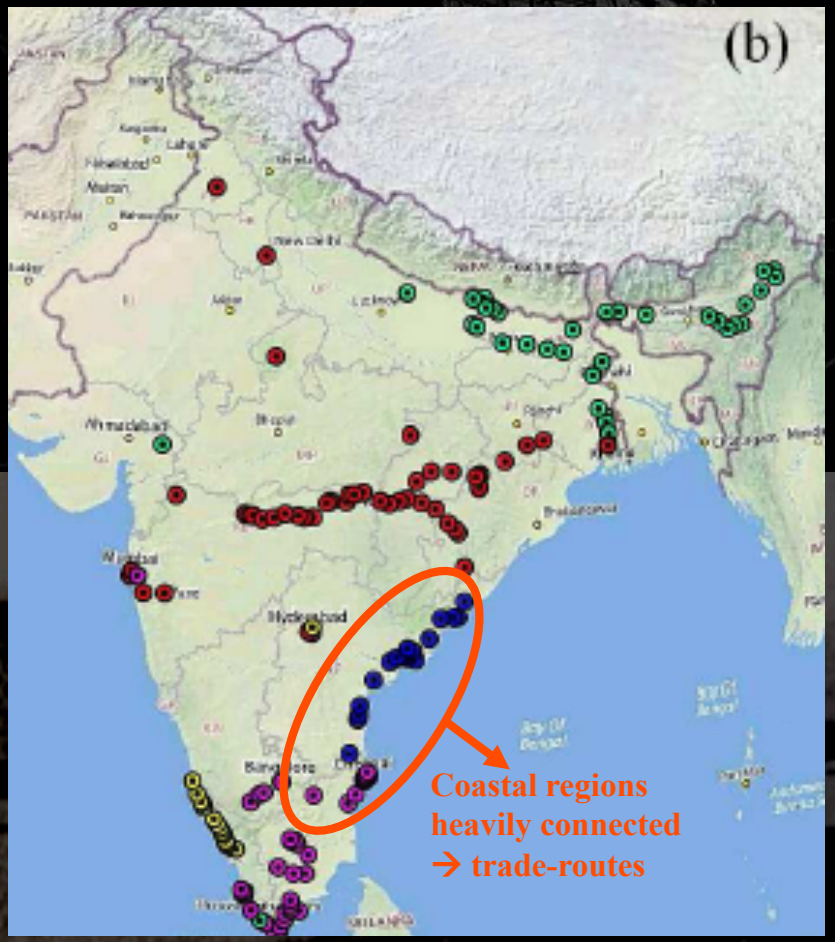


Forms the first level of hierarchy

Neutral middle limb which could not be clustered. Construct a smaller network with these residual nodes and repeat the second eigenvector analysis. Continue until there are no nodes here OR it is a complete scatter

Results on Map





Observations

- Community analysis shows that geographic proximity is the basis of the hierarchical organization of RNs for both the countries
- The geographically distant communities are connected among each other only through a set of hubs or junction stations.
- Can be useful while planning the distribution of new trains
  - India: Bharatpur not well-connected to the Farakka/ Maldah stations
  - Germany: Hanover not well-connected to Hamburg/ Bremen stations

Similarities across geography!

- How the two different nations with completely different political and social structures can have exactly the same pattern of organization of their transport system?
  - Transportation needs of humans are same across geography and culture
  - Short-distance travel for any individual is always more frequent than the long distance ones
  - On a daily basis, a much larger bulk of the population do short-distance travels while only a small fraction does long-distance travels


Thank You