RISE
Reconfigurable and Intelligent Systems Engineering
ILDS

Data Science for Smart Cities

Making Sense of a Connected World

B. Ravindran

Reconfigurable and Intelligent Systems Engineering (RISE) Group
Department of Computer Science and Engineering
and Interdisciplinary Laboratory for Data Sciences (ILDS)
Indian Institute of Technology Madras

Smart Cities

Smart Cities

- Key: Know the city better
- Instrument every aspect of the city
- Sensor Networks; Internet of Things
- Large volumes of data generated
- Make sense of these volumes

Smart Cities

- Key: Know the city better
- Instrument every aspect of the city
- Sensor Networks; Internet of Things
- Large volumes of data generated
- Make sense of these volumes

Data Science!

Data Science

- Explain observed phenomenon
  - Observe, model, predict, test
- Data Analytics does not exploit much of the understanding on the modeling front
  - Discover dependencies; Causal reasoning
  - Feature encoding; representation
  - Involve domain experts closely in the analytics workflow

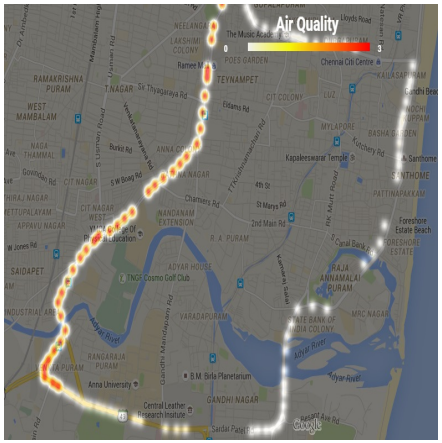
Observe

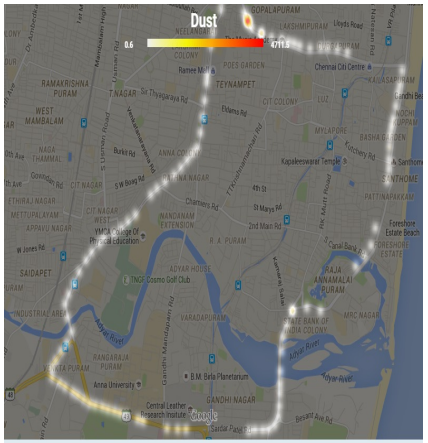
- Data Gathering
  - Sensing
  - Communication
- Data Handling
  - Storage
  - Indexing
  - Retrieval
- Some initiatives at IITM

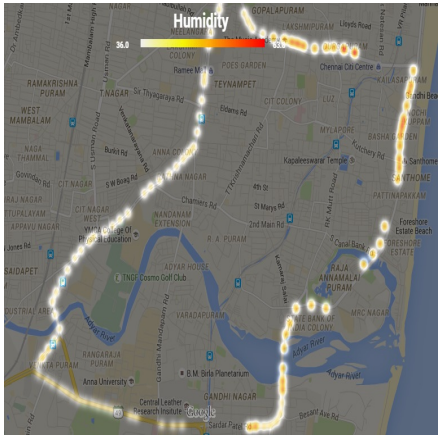
Intellimeter

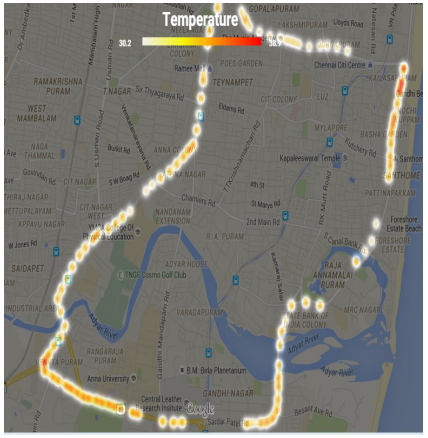
- BIG data through auto-meters as micro-agents for data collection
- Disparate sensor types
  - GPS data, Video, Audio
  - Noise, Pollution, Vibration data
- Direct impact on smart city planning
- In discussions with an auto rickshaw aggregator for deployment of these meters in Chennai autos
- A cloud solution for managing the data and analytics with this data
- In 3 months, this platform will be accessible from IIT Madras website

In Chennai: IIT-Nungambakkam-Santhome-IIT



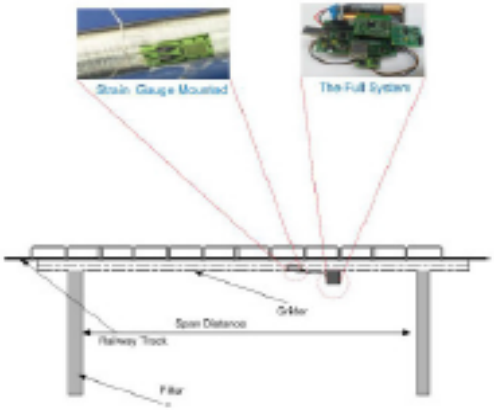






Structural Monitoring of Railway Bridges

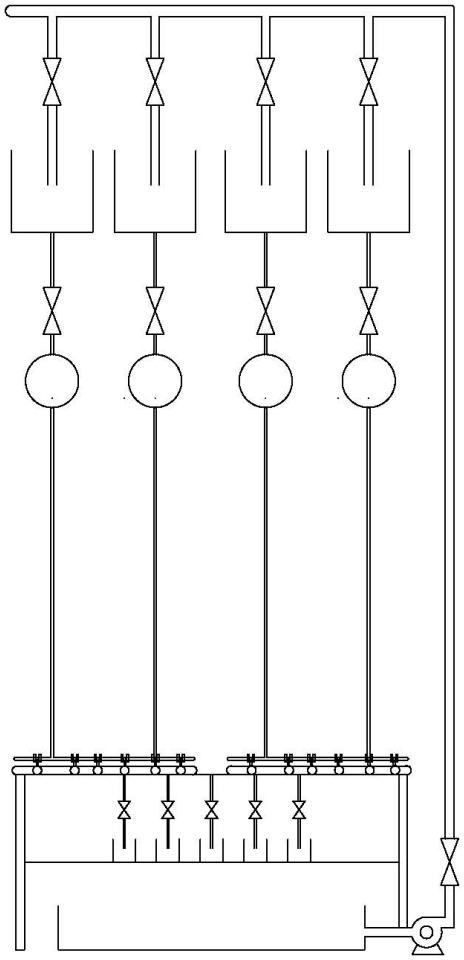
Deployment Scenario



- Minimum sampling rate is 100 readings per second.
- Continuous flow of strain output for 5 Minutes.
- Accelerometer reading for 5 Minutes
- Wake up on 0.063g
- Similar set up for monitoring of buildings

Experimental cyber-physical system for water distribution

- Complex:
  - 25 demand nodes
  - 4 balancing reservoirs
  - Gravity driven
  - Reconfigurable
- Small “footprint”
  - 2 m by 2m on ground
- Scaled down in space and time
  - 4-20 mm pipes
  - 1 minute sampling
- Sub-network
  - 5 demand nodes served by a balancing reservoir
  - Storage facilities at each demand node
  - Loops or tree type networks
- Sensors and actuators
  - Level sensors in all tanks/reservoirs
  - Limited pressure and flow
  - Limited continuous control
  - ON/OFF valves at demand points



*Supported by DST WTI and ILDS

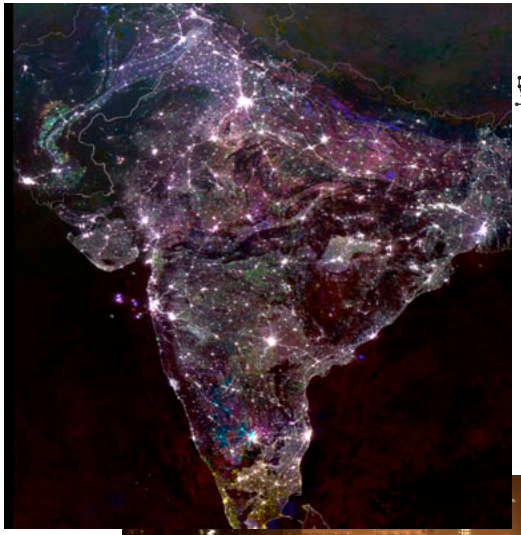
Model and Predict

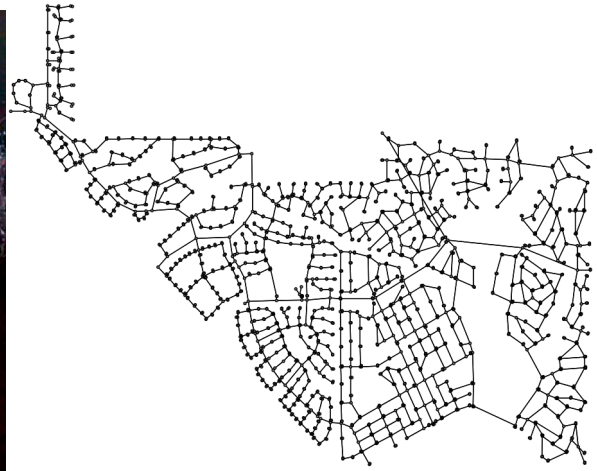
- Go hand-in-hand
- Novel Settings
  - Network data
  - Real-time constraints
  - Diverse data types
- Scaling up computation
  - Distributed computing
- Privacy and Security

Model and Predict

- Go hand-in-hand
- Novel Settings
  - Network data
  - Real-time constraints
  - Diverse data types
- Scaling up computation
  - Distributed computing
- Privacy and Security

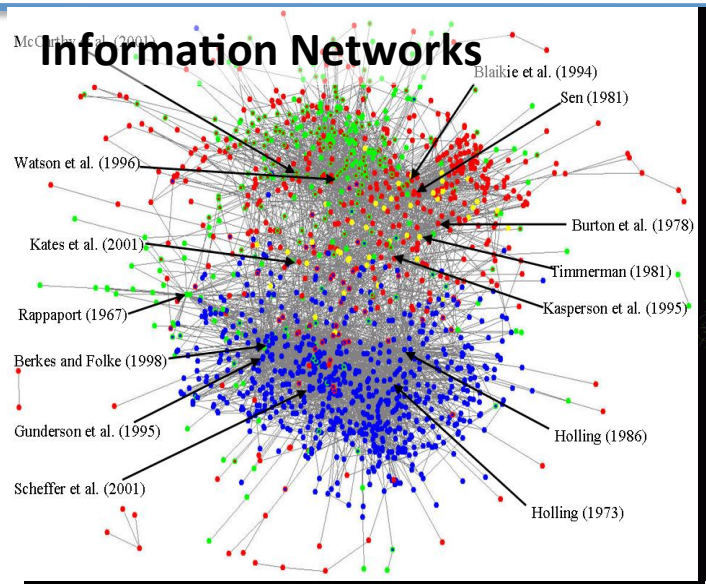
Networks are everywhere

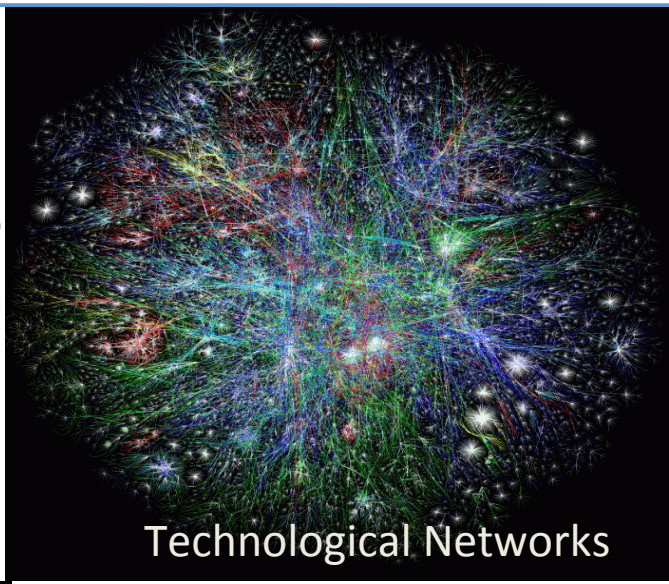


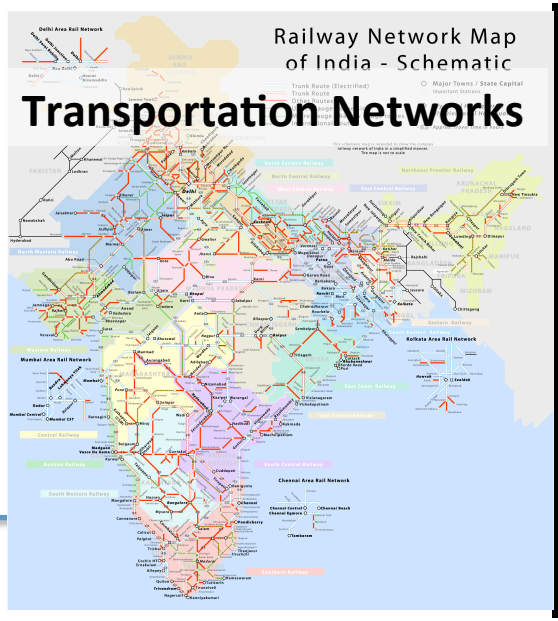


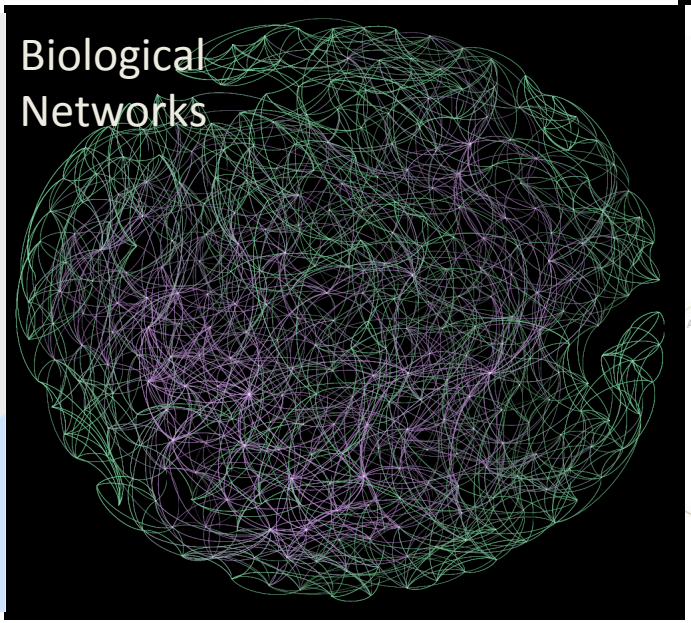
Networks are everywhere!

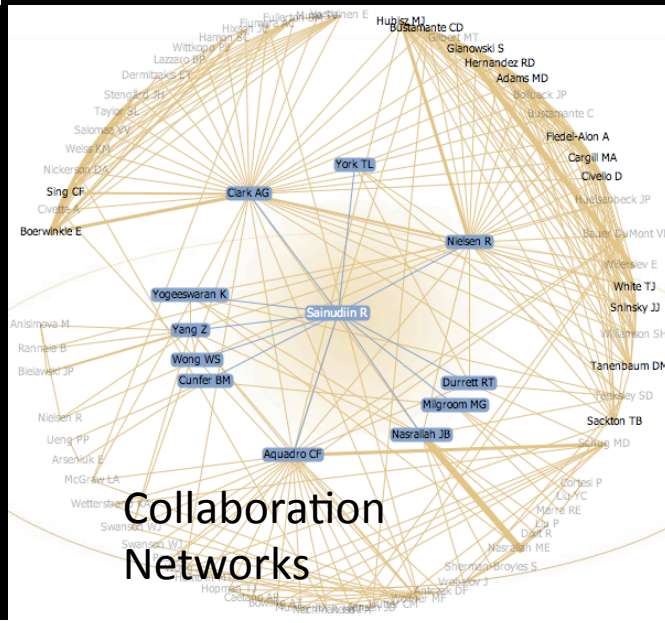
Social Networks











Why study networks

- Networks capture *connectedness*
- Structural Properties
  - What are the network properties? Small World, Scale Free
  - Who is connected to who? Communities, Homophily
  - Who influences who? Virality, Epidemics
- Behavioral
  - How do people communicate? Propagation
  - How do people make decisions? Cascades
  - How do connections form? Evolution

Network Models

- Can model many of the smart city problems on networks
  - Transportation Networks
  - Utility Networks
    - Power, water, etc.
  - Internet of Things (connected devices)
  - Telecom Networks
    - Call graphs, infrastructure, etc.

Network Analytics

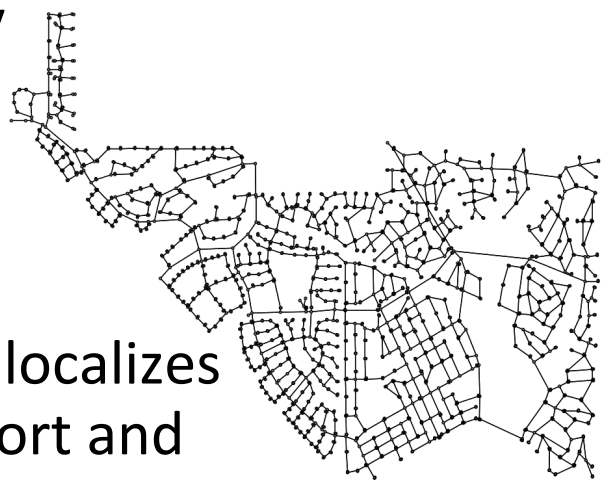
- Structural queries
  - Segmentation of zones
  - Predicting Importance
  - Identifying weak spots
- Behavioral queries
  - Travel time prediction
  - Traffic patterns
  - Usage patterns

Network Analytics

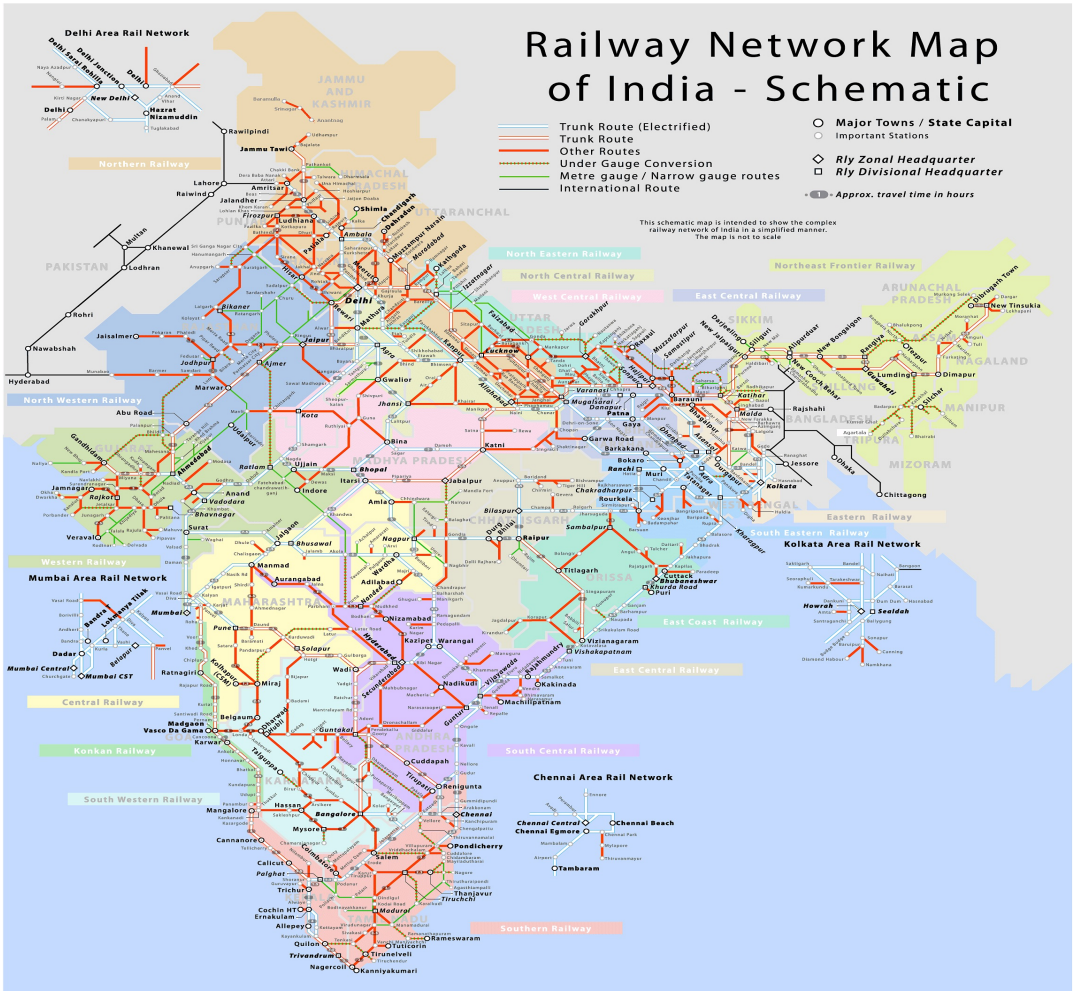
- Structural queries
  - Segmentation of zones — Clustering
  - Predicting Importance — Centrality
  - Identifying weak spots
- Behavioral queries
  - Travel time prediction
  - Traffic patterns
  - Usage patterns

Leak detection in large networks [5]

- Non-Revenue Water as high as 30-50%
  - Energy wastage, compromised water quality
  - Existing methods (acoustics, helium gas) are costly, disruptive, and highly localized.
- Requirement & Problem Statement:
  - Find scalable field campaign strategy which localizes leak to small enough size, with minimum effort and disruption.
- Our Solution:
  - Divide using graph partitioning
  - Conquer using water balance



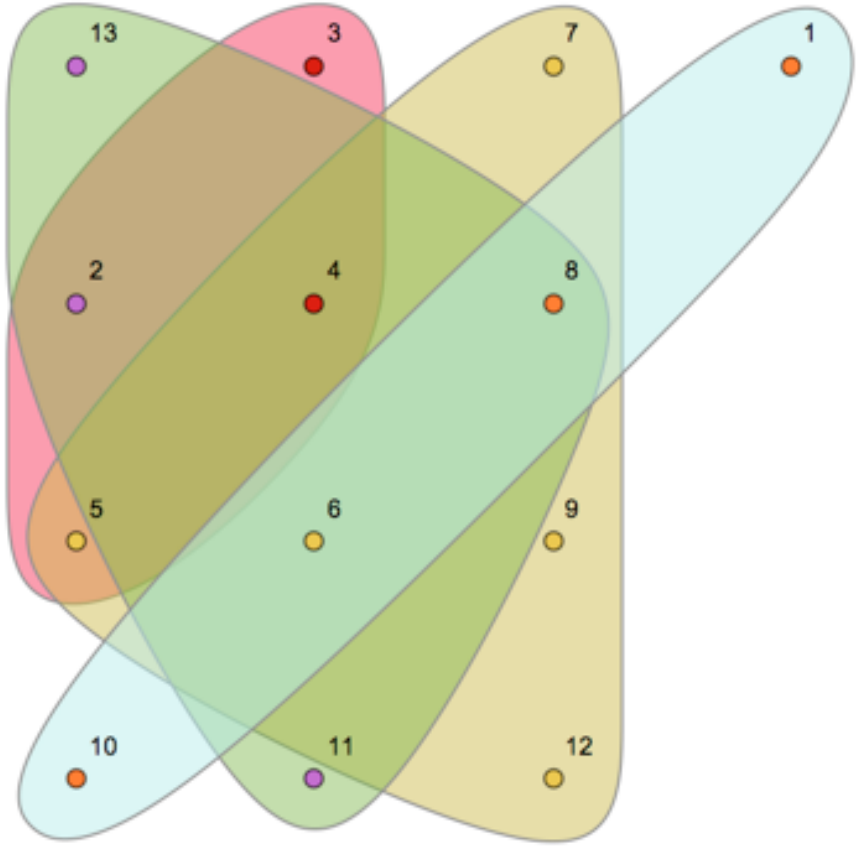
Indian Railway Network



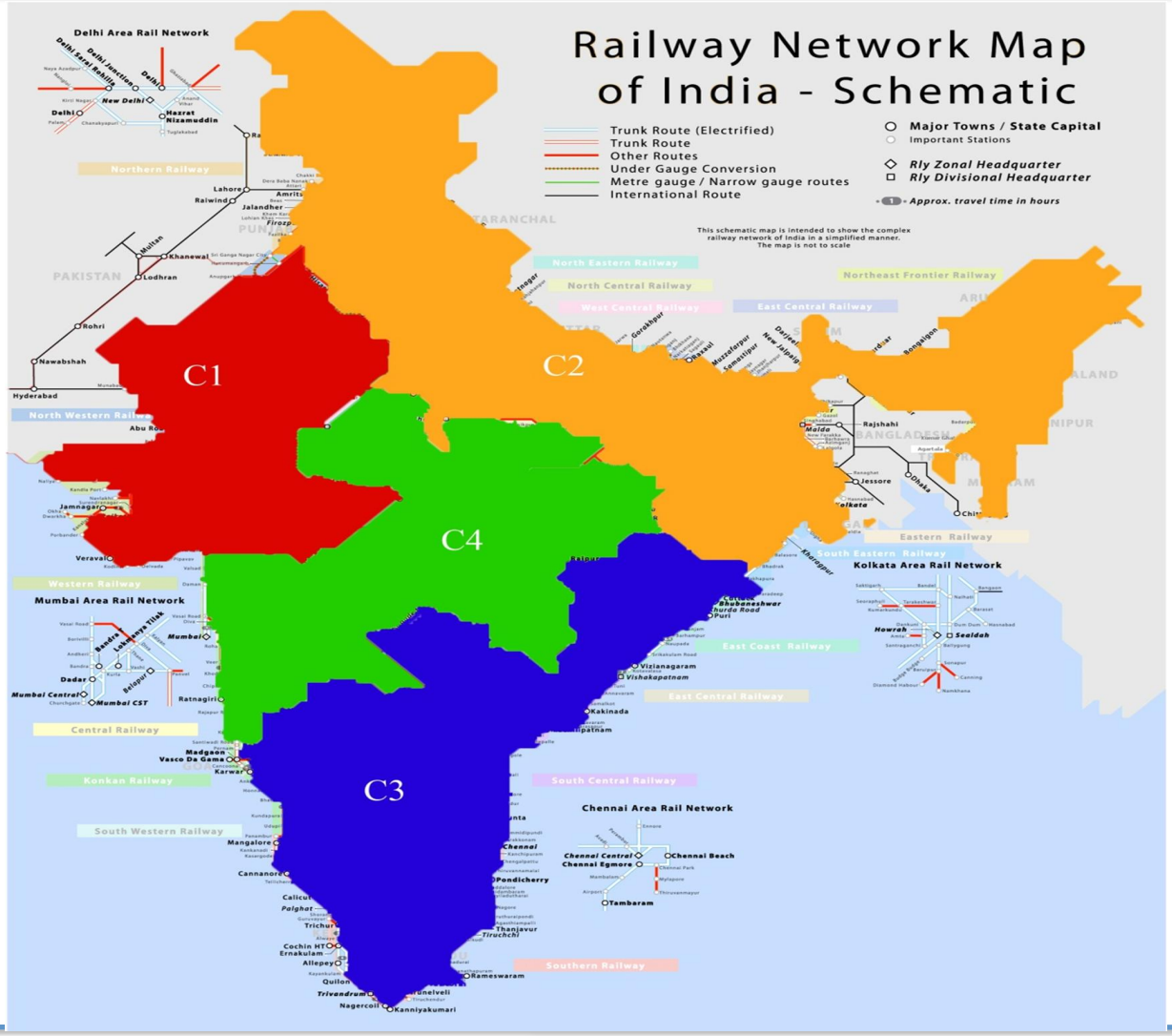
Community Detection

Railway Network [2]

- Hypergraphs – richer structure
- Stations are nodes
- Trains are hyperedges

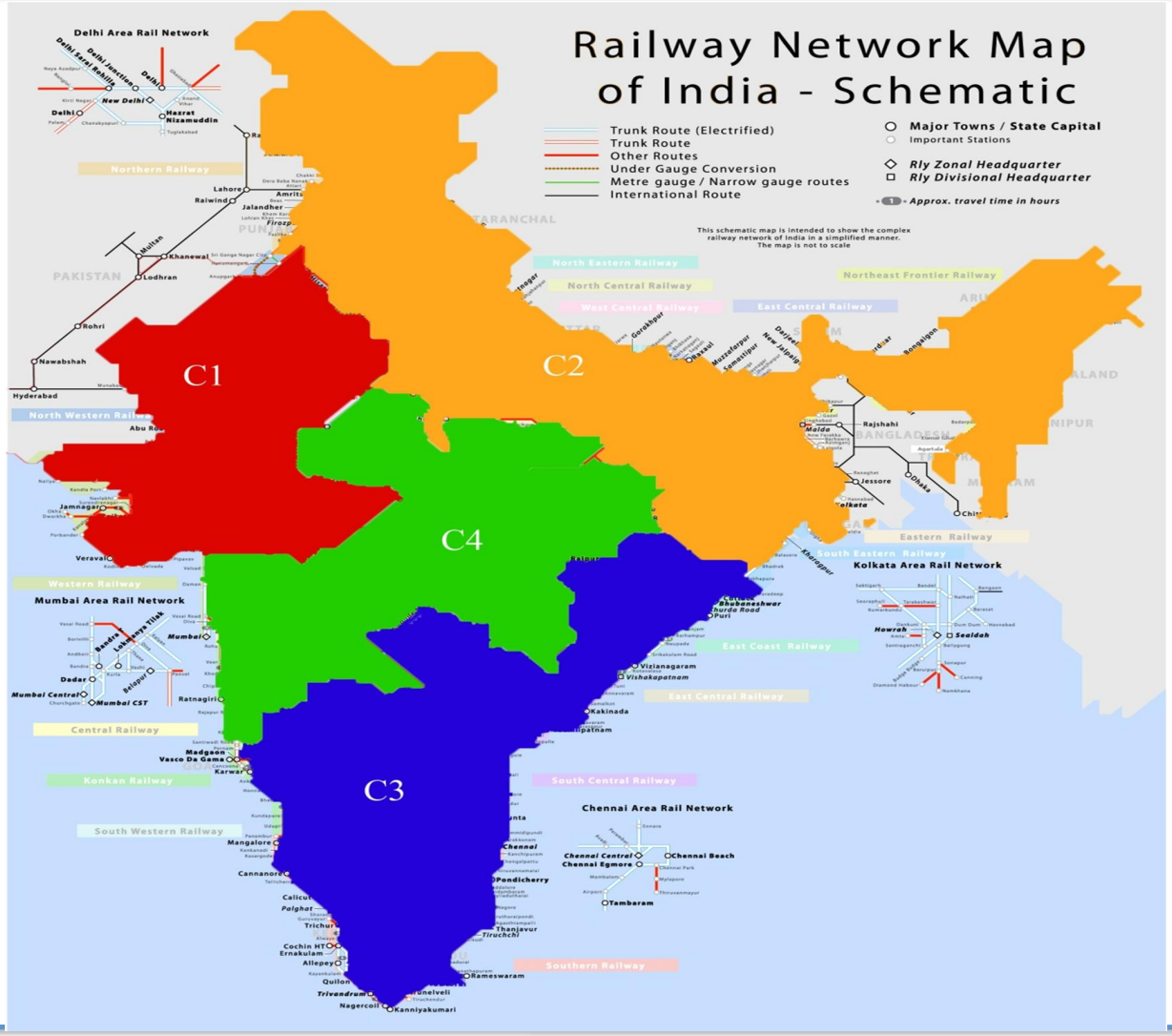


Communities and Zones



Communities and Zones

Partitioning on regular graph models produced 75 communities!



Scalable Community Detection [4]

Networks	# of nodes	# of edges	Louvain method	CEIL algorithm
Youtube	1,134,890	2,987,624	321.25s	395.25s
DBLP	317,080	1,049,866	134.39s	77.77s
Amazon	334,863	925,872	81.17s	80.68s

- CEIL score does not suffer from resolution limit
  – Can find small and large communities
- Correlates well with known community structures

Centrality

- The notion of centrality is a crucial one in network analysis
  - Traffic analysis
    - Congestion; volume of traffic
  - Marketing
    - Influential users
  - Impact Factor analysis
    - Citations
- Tuned to applications

Centrality Measures

- Degree centrality
  - More incident edges
  - Junctions
- Shortest Path Measures
  - Between-ness centrality
    - No. of shortest paths that pass through a node
    - Individually optimize paths!
  - Closeness centrality
    - Aggregate pair-wise shortest path length
    - More centrally located
- Clustering Coefficient
  - How densely connected is a neighbourhood
- Spectral notions of centrality
  - Link to important nodes
  - Page rank, HITS

Centrality Measures

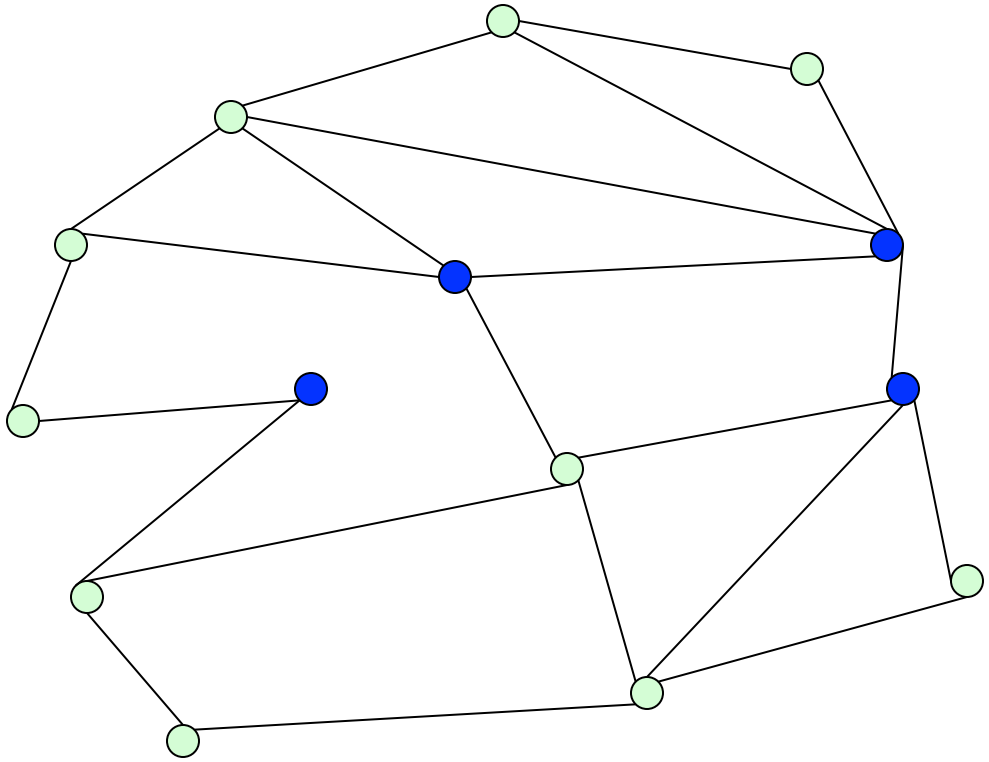
- Degree centrality
  - More incident edges
  - Junctions
- Shortest Path Measures
  - Between-ness centrality
    - No. of shortest paths that pass through a node
    - Indicates how to optimize paths!
  - Closeness centrality
    - Aggregate measure of shortest path length
    - Most centrally located
- Clustering Coefficient
  - How densely connected is a neighbourhood
- Spectral notions of centrality
  - Link to important nodes
  - Page rank, HITS

Hard to compute at large-scales

Game Theoretic Centrality Measures

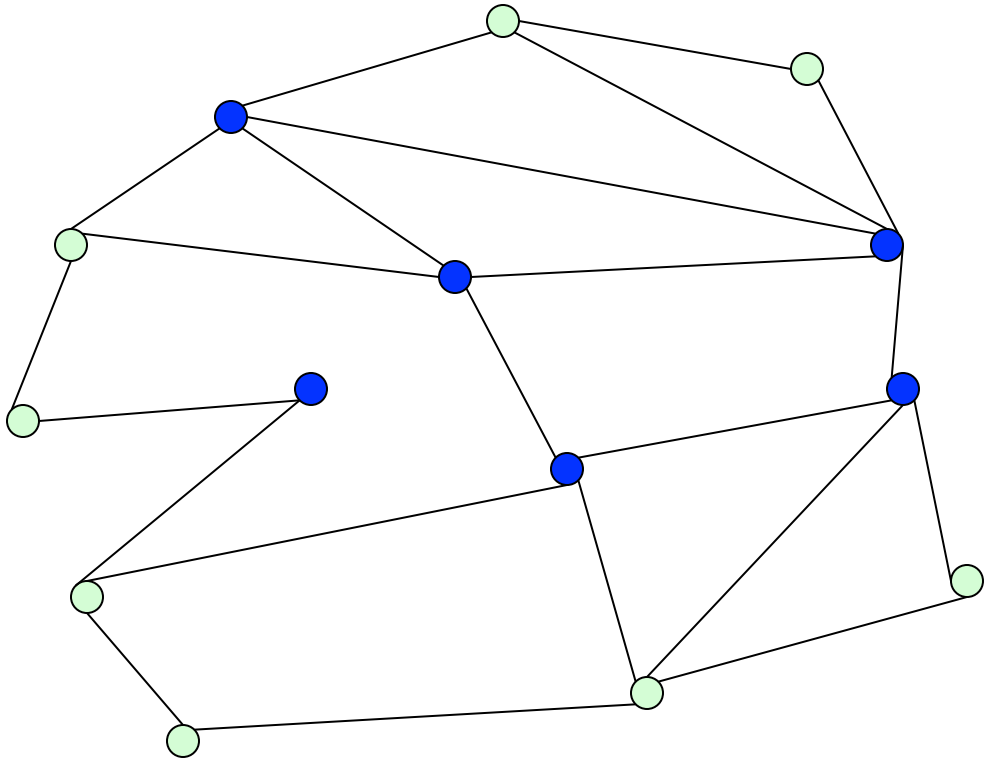
- Our focus: Diffusion
  - How effectively will events travel?

Diffusion of Influence



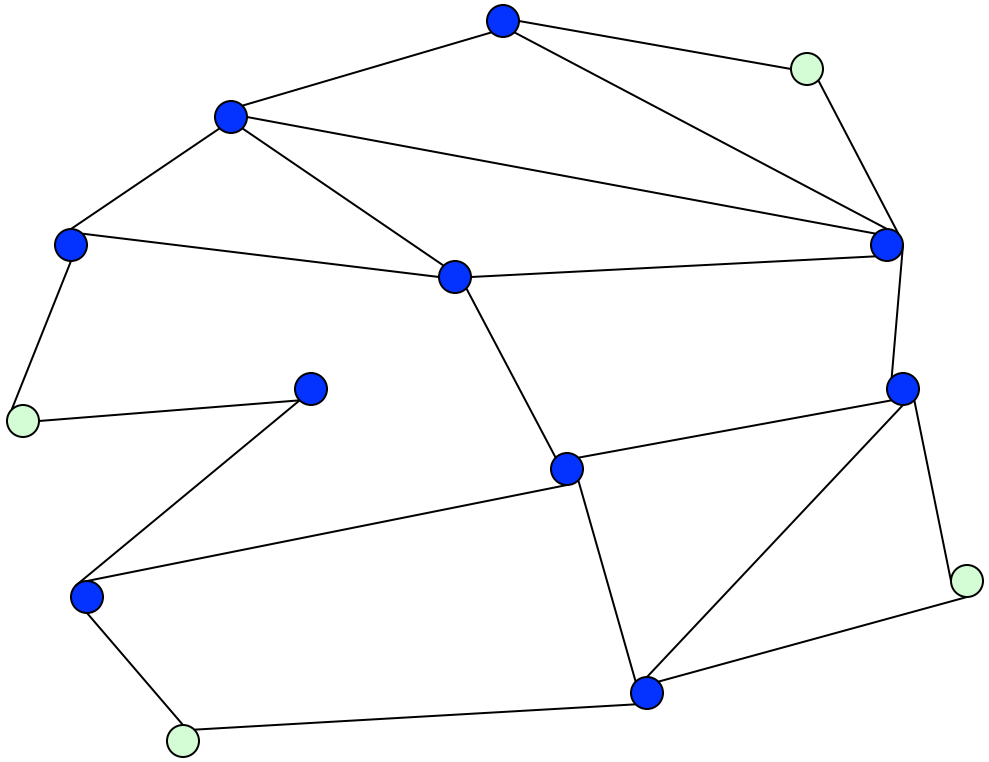
A node is influenced is at least two of its neighbours are influenced.

Diffusion of Influence



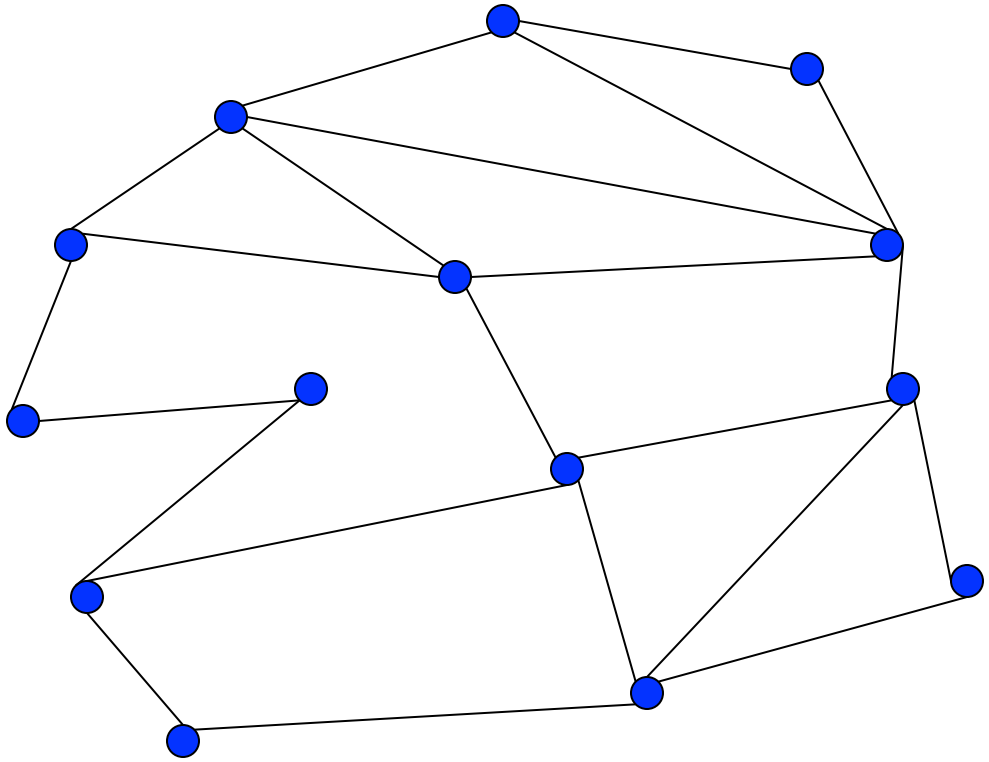
A node is influenced is at least two of its neighbours are influenced.

Diffusion of Influence



A node is influenced is at least two of its neighbours are influenced.

Diffusion of Influence



A node is influenced is at least two of its neighbours are influenced.

Game Theoretic Centrality Measures

- Our focus: Diffusion
  - How effectively will events travel?
- Model as cooperative games on networks
  - Marginal Contribution of Node to influence spread (Shapley Value)
  - Original idea: Ramasuri and Narahari.
  - Our Contribution: Efficient Computation [1,3]

Various Versions [1]

Model behavioral aspects as well

1. Top-k
   - Immediate influence
2. k-neighbour
   - Gossip spreading; peer pressure
3. k-hop; d-cutoff
   - infectious diseases
4. Influence is a function of distance
   - More suited for traffic modeling
5. Linear threshold

Various Versions [1]

Model behavioral aspects as well

1. Top-k
   - Immediate influence

$$SV_{g1}(v_i) = \sum_{v_j \in \{v_i\} \cup N_G(v_i)} \frac{1}{1 + \deg(v_j)}$$

2. k-neighbour
   - Gossip spreading; peer pressure
3. k-hop; d-cutoff
   - infectious diseases
4. Influence is a function of distance
   - More suited for traffic modeling
5. Linear threshold

Centrality Score – Results [3]

Table : Running time in seconds of all the game theoretic algorithms on Erdos Renyi graphs of density 0.0001

Network Size	Game 1	Game 2	Game 3	Game 4	Game 5
10^3	28	30	41	44	29
10^4	29	29	44	45	30
10^5	31	31	49	45	30
10^6	46	45	70	70	46
10^7	79	81	301	312	79

Hadoop cluster with 64 nodes

Model and Predict

- Go hand-in-hand
- Novel Settings
  - Network data
  - Real-time constraints
  - Diverse data types
- Scaling up computation
  - Distributed computing
- Privacy and Security

Computing Degree

- On large graphs computing even degree of a node is hard!
- Data arrives one edge at a time
  - *A* called *B* at time *t* and spoke for *m* minutes
  - Customer *C* bought *u* units of item *I*
  - User *X* posted on the wall of user *Y*
  - Protein *P* was observed to interact with protein *Q*
- How to efficiently aggregate the edge events into a graph?

Network Processing Stack?

- Most big data deployments are not optimized to process network data
- Some solutions available but mostly ad-hoc
- Need to rethink network data processing
  - Integrate with existing compute infrastructure
  - Storage models
  - Another line: Re-do from scratch

Going Forward

- Big role for data sciences in smart cities
  - Smart mobility, smart houses, health care, smart infrastructure, etc.
- Interdisciplinary Laboratory for Data Sciences
  - Analytics for smart cities
  - Other aspects of data sciences as well
    - Systemic as well as algorithmic
  - 6 different departments
    - BT, CE, CHE, CSE, EE, DoMS

ILDS

- Strong background in data-driven modeling and model based prediction/reasoning
  - Smart Cities
    - Traffic engineering
    - Water and electricity distribution
  - Systems biology
  - Financial data (insurance, risk analytics, etc.)
- Developing compute infrastructure suited for specific use cases
- Bring together people with strong sciences background and computational skills

http://cmsrv.iitm.ac.in/ilds/home/

References

1. Aadithya, K. V., Ravindran, B., Michalak, T., and Jennings, N. R. (2010) "Efficient Computation of the Shapley Value for Centrality in Networks". WINE 2010
2. Jain, S. K., Satchidanand, S. N., Maurya, A. K., and Ravindran, B. (2014) "Studying Indian Railways Network using Hypergraphs". COMSNETS workshop on social networking, 2014
3. M. Vishnu Sankar and Balaraman Ravindran, “Parallelization of Game Theoretic Centrality Algorithms,” Sadhana, Academy proceedings in Engineering Sciences, Special issue on machine learning for big data, 2015
4. M. Vishnu Sankar, Balaraman Ravindran and S. Shivashankar, “Fast Method for Finding Resolution Limit Free Communities in Large Networks,” IJCAI 2015
5. Aravind Rajeswaran, Sridharakumar Narasimhan, and Shankar Narasimhan, “A graph partitioning approach for leak detection in distribution systems.” SDM workshop on Network Science, 2015.

http://www.cse.iitm.ac.in/~ravi