RISE
Reconfigurable and Intelligent Systems Engineering

TCS Eminence Lecture

Machine Learning in a Connected World

B. Ravindran
Reconfigurable and Intelligent Systems Engineering (RISE) Group
Department of Computer Science and Engineering
Indian Institute of Technology Madras

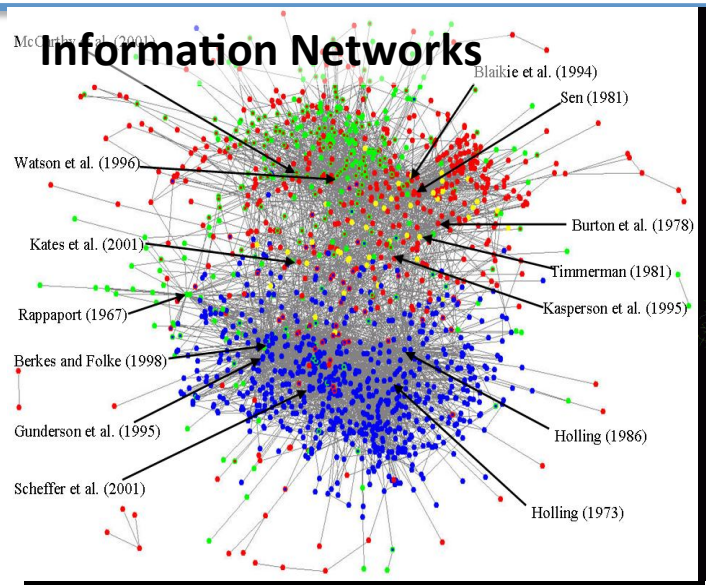
Connected World

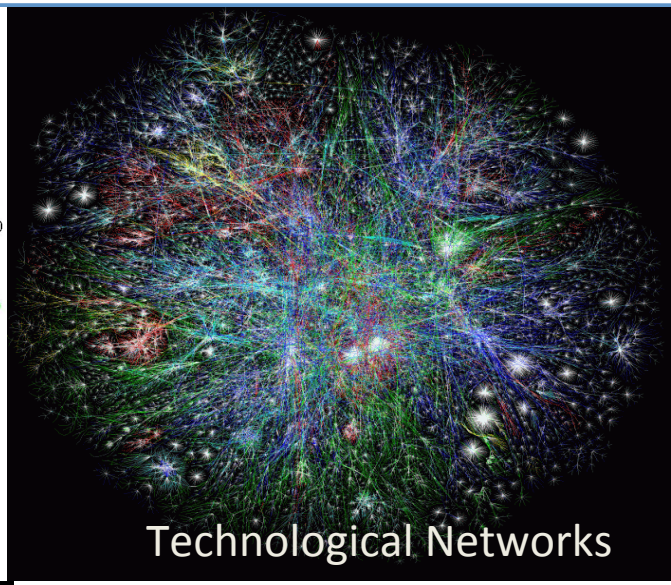
Connecting the World

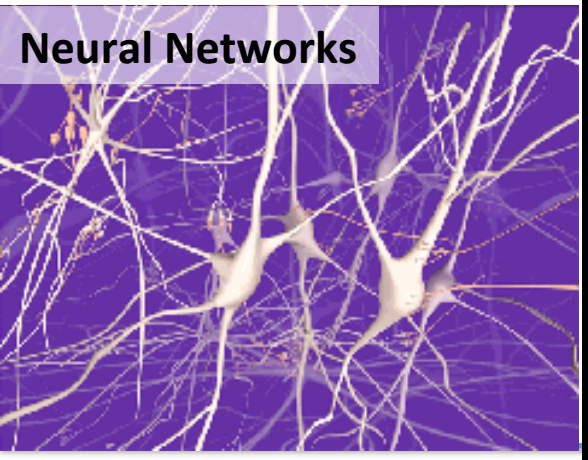
Networks are everywhere!

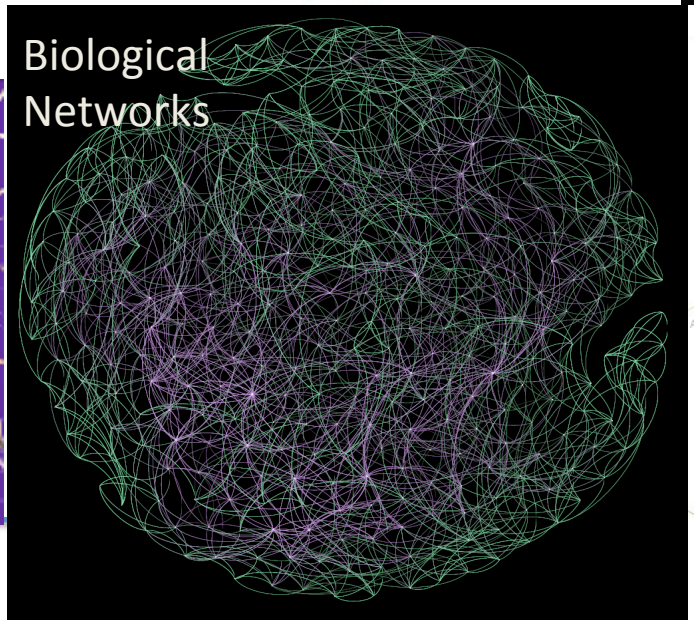
Social Networks

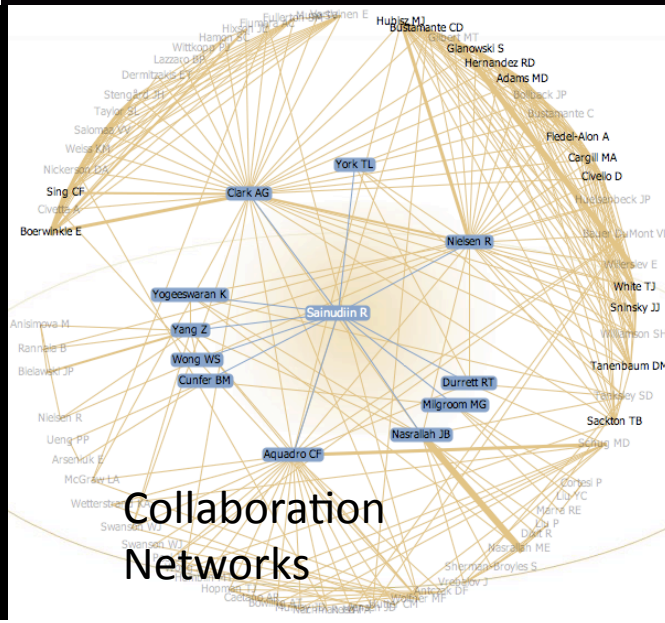
twitter











Two Themes

- Learning about Networks
  - Structure of networks
  - Network formation
  - Dynamic behavior on networks
- Learning on Networks
  - Attributed nodes
  - Collective learning
  - More powerful predictors

RISE
Reconfigurable and Intelligent Systems Engineering

Learning about Networks

Tracking two Identified terrorists

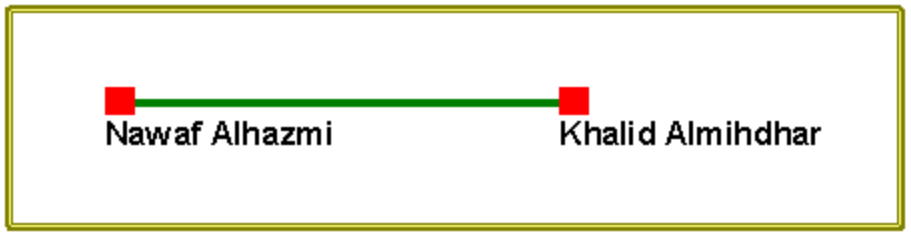


Figure 1 - Two known suspects in January 2000

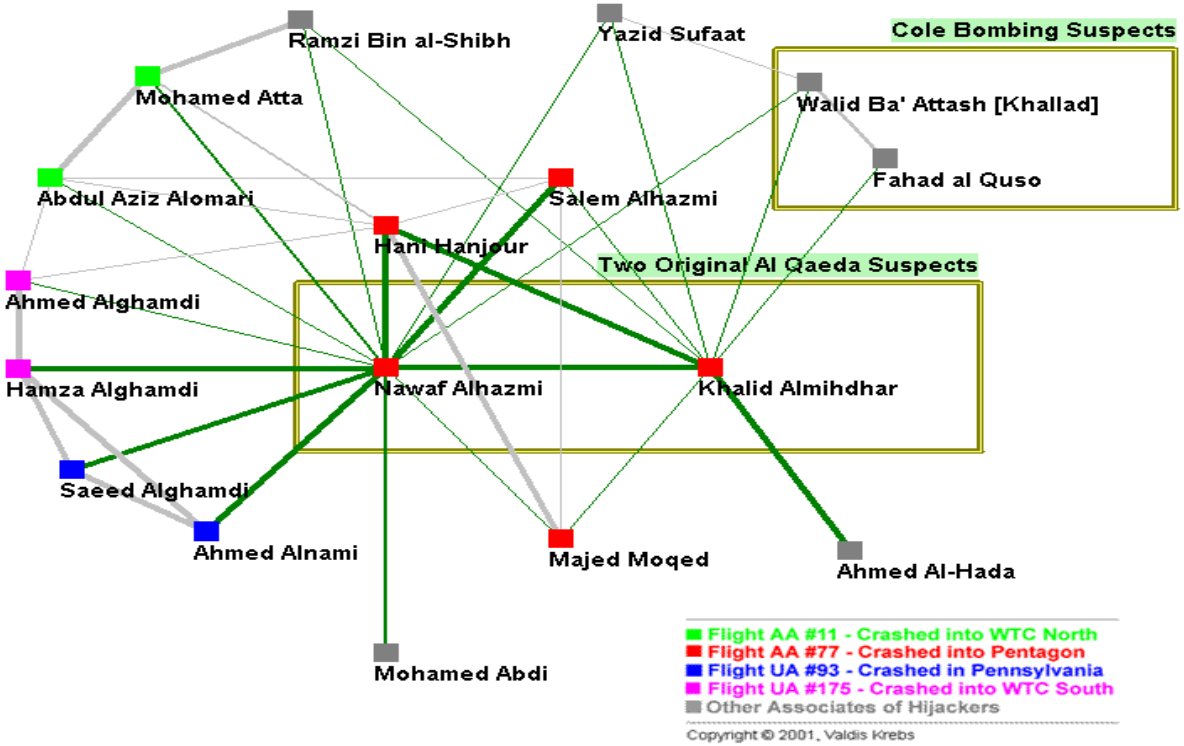


Figure 2 - All nodes within 1 step (direct link) of original suspects

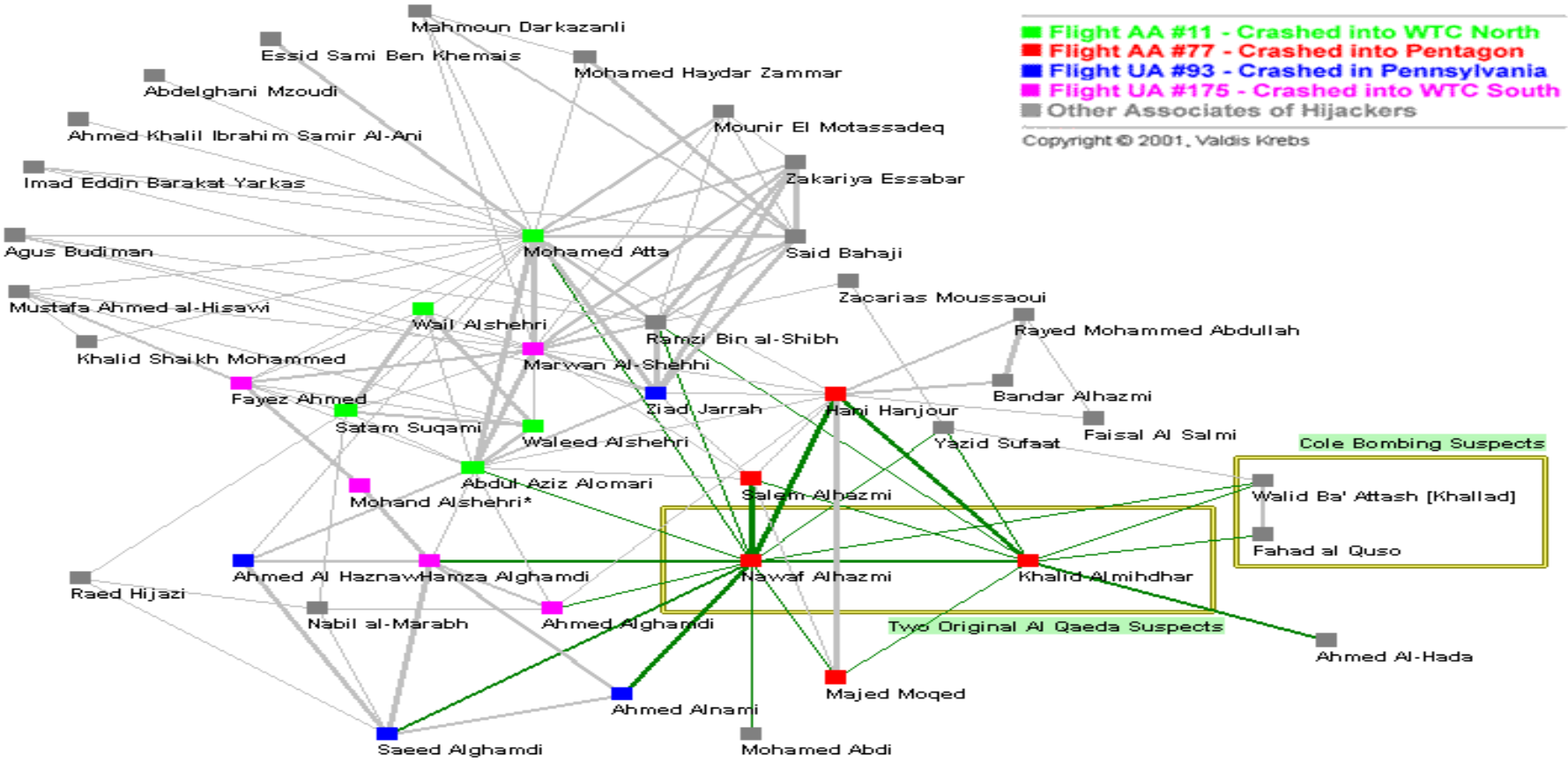


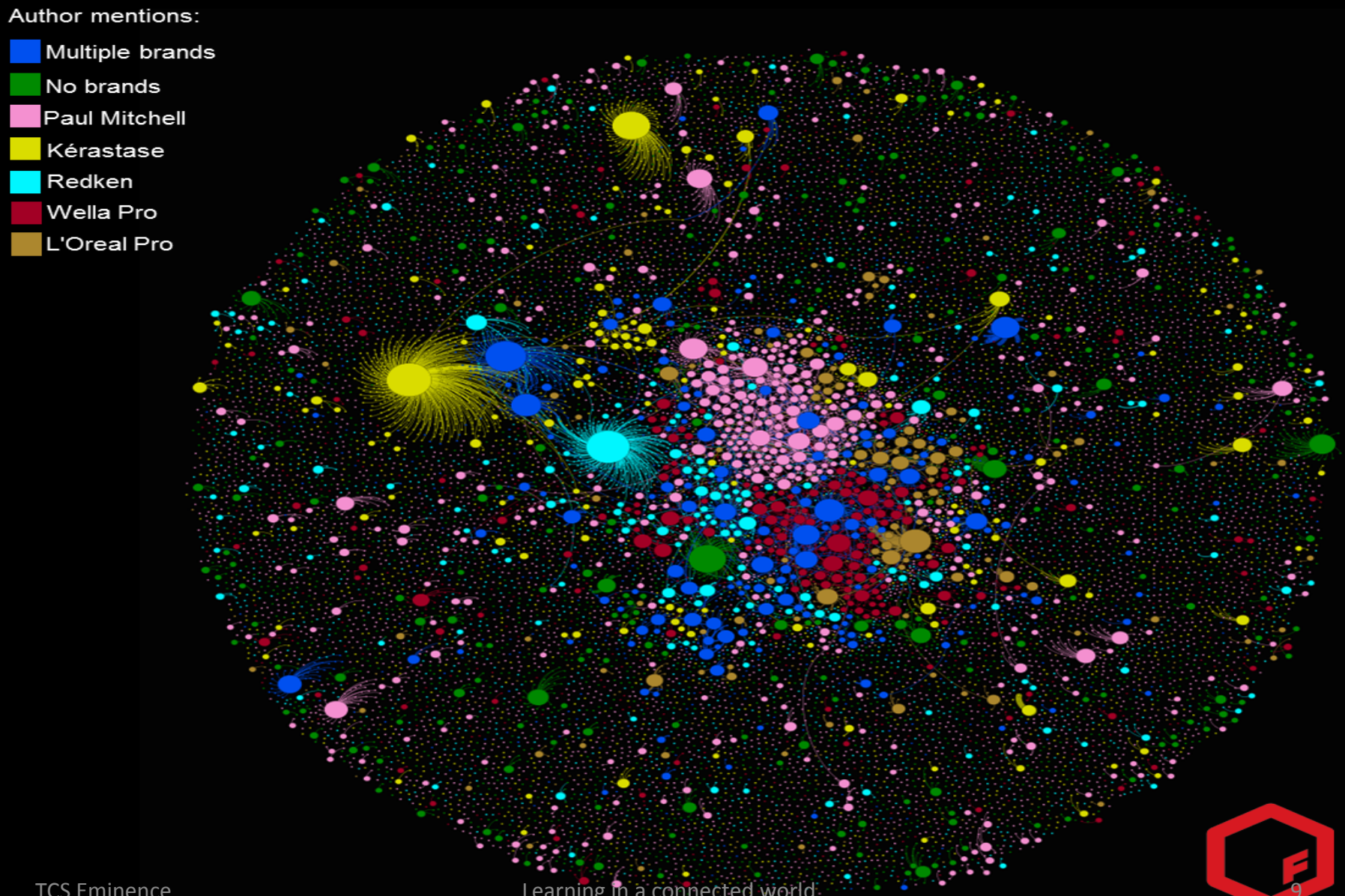
Figure 3 - All Nodes within 2 steps / degrees of original suspects

Insights:

1> All 19 hijackers were within 2 steps of the two original suspects uncovered in 2000!

2> Social network metrics reveal Mohammed Atta *emerging* as the local leader

Viral Marketing



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 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

- Many application specific measures of centrality
- Our focus: Information diffusion
  - How quickly will news travel?
  - Basic model proposed by Kempe, Kleinberg, and Tardos.
- Model as cooperative games on networks
  - Marginal Contribution of Node to influence spread (Shapley Value)
  - Original idea: Ramasuri and Narahari.
  - Our Contribution: Efficient Computation [1,4]

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

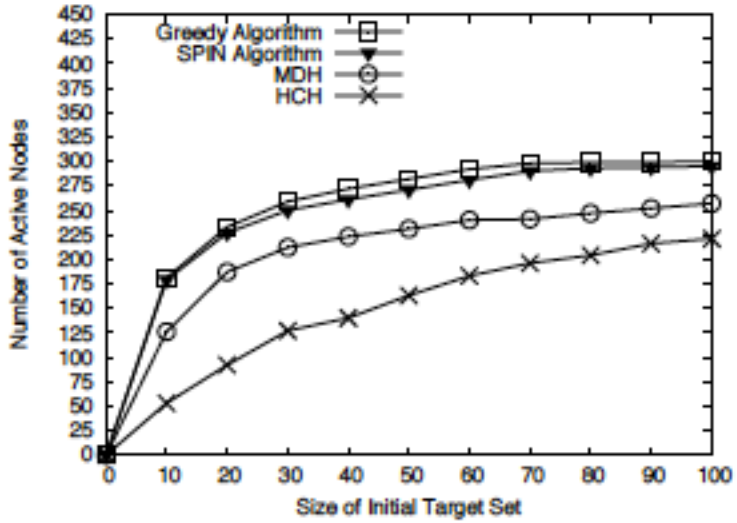
1. Top-k
   - Immediate influence

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

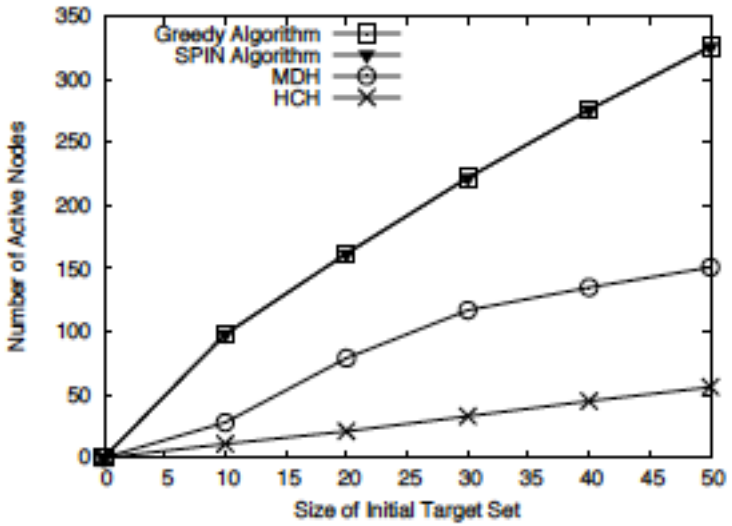
2. k-neighbour
   - Gossip spreading; peer pressure
3. k-hop; d-cutoff
   - infectious diseases
4. Influence is a function of distance
   - More suited for traffic modeling
5. Linear threshold

Results RN, 2010

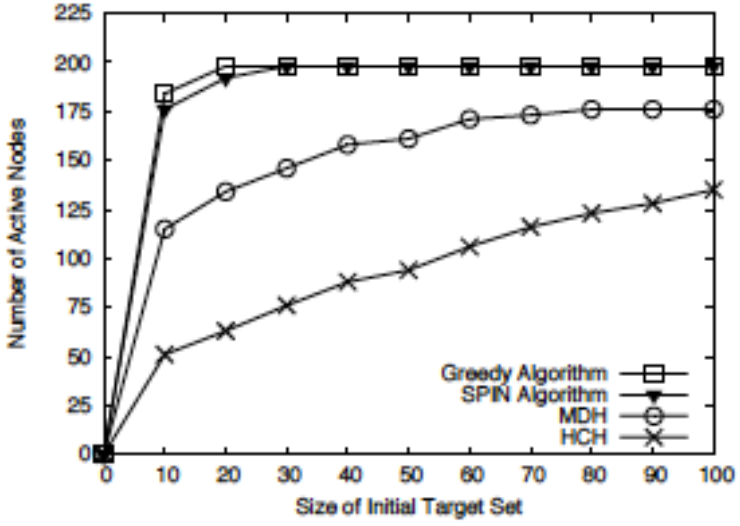
Celegans



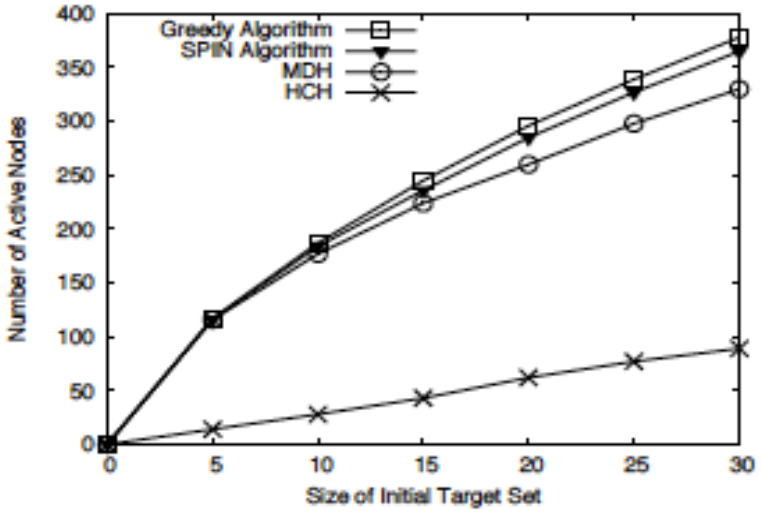
Netscience



Jazz Musicians



NIPS



Centrality Score – Results [4]

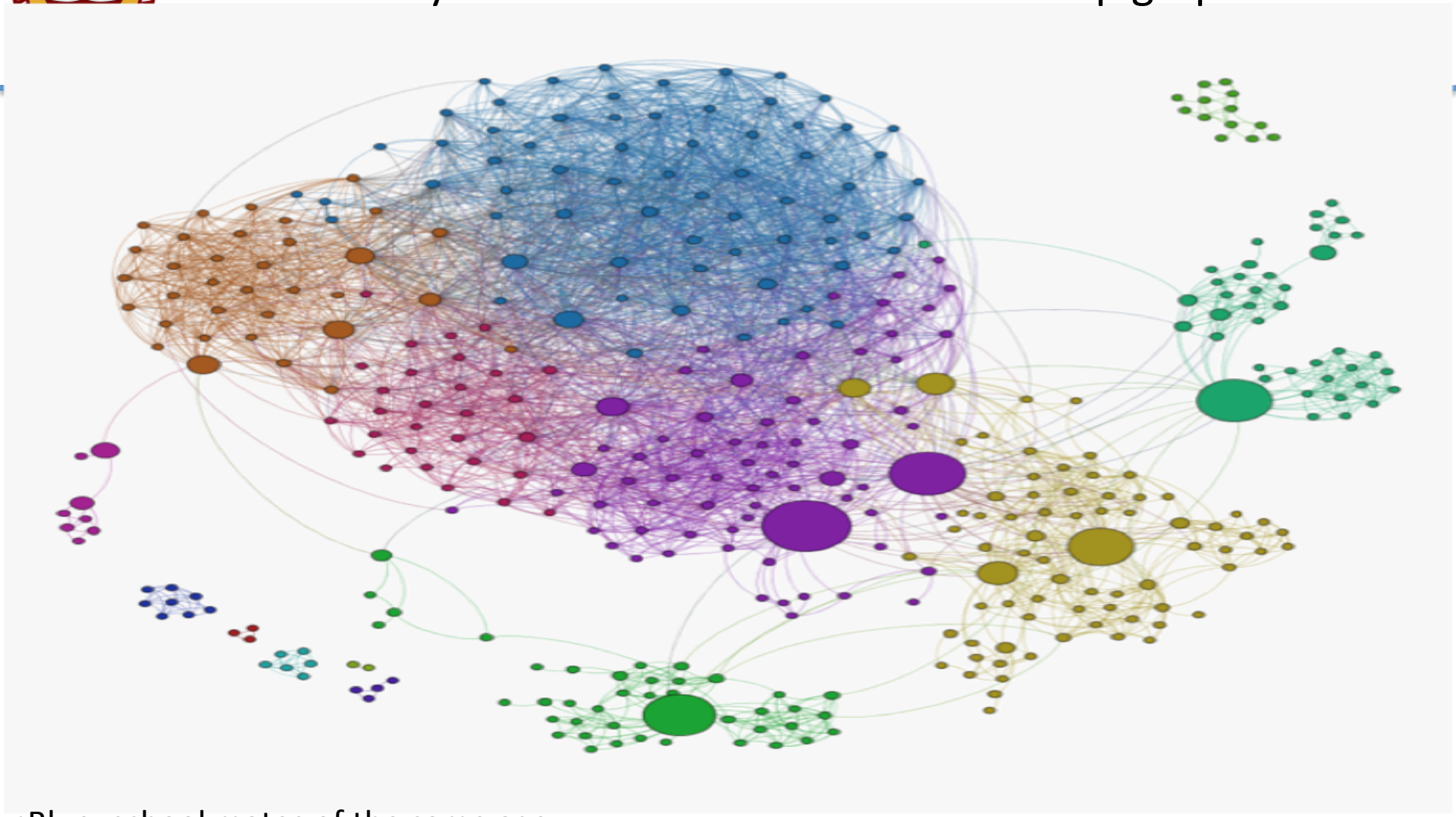
Table : Running time in seconds of all the game theoretic algorithms on Erdos Renyi graphs of density 0.0001

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

Structural Questions

- Community Detection
  - Identify groups of closely related nodes
  - Special interest groups
  - Research areas
  - Understand structure of network better
- Positional Analysis
  - Significant structural signatures in the network
  - Temporal evolution

Community detection on Facebook friendship graph



- Blue: school mates of the same age
- Brown: school mates from the same locality of residence
- Maroon: seniors from school
- Purple: college mates

Scalable Community Detection [5]

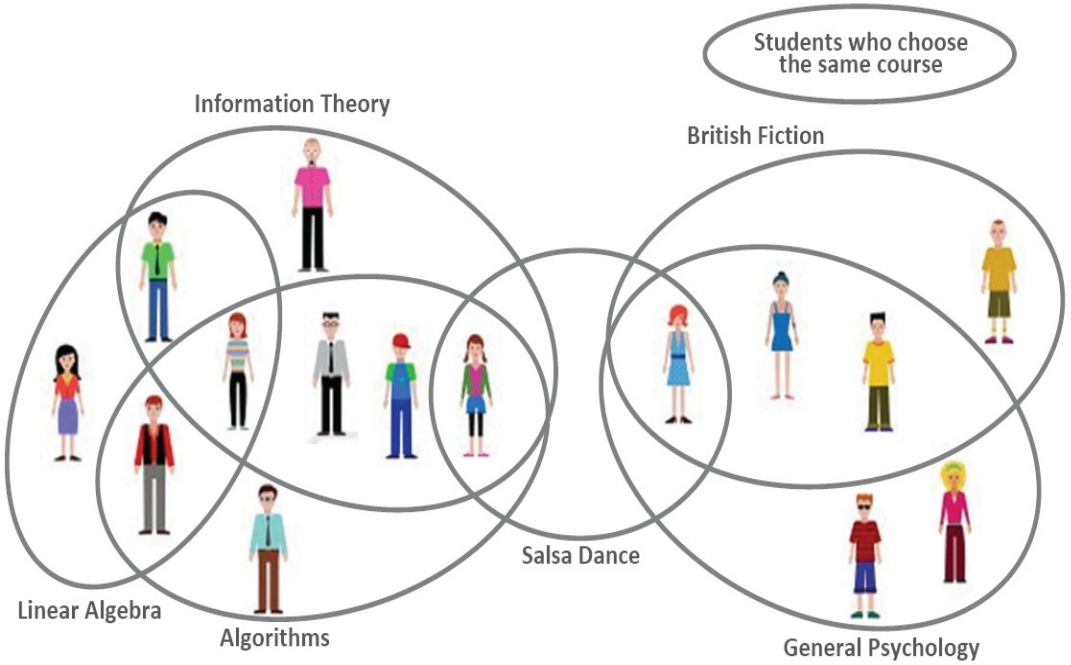
Networks	# of nodes	# of edges	Louvain method	CEIL algorithm
Youtube	1,134,890	2,987,624	321.25s	395.25s
DBLP	317,080	1,049,866	134.39s	77.77s
Amazon	334,863	925,872	81.17s	80.68s

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

Community Detection

Railway Network [2]

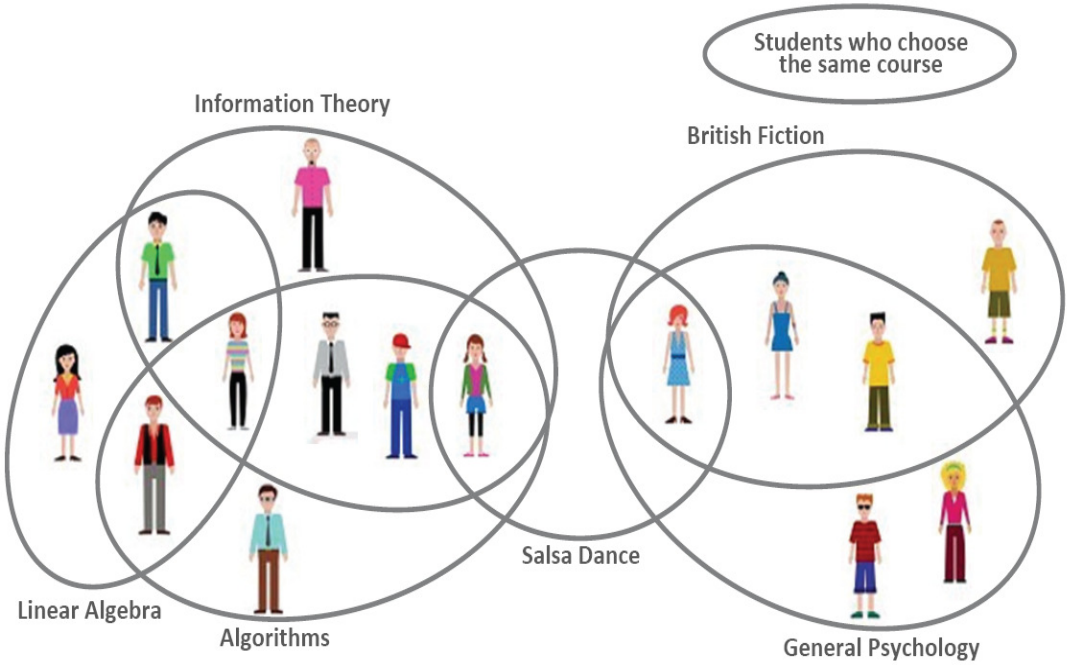
- Hypergraphs – richer structure



Community Detection

Railway Network [2]

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

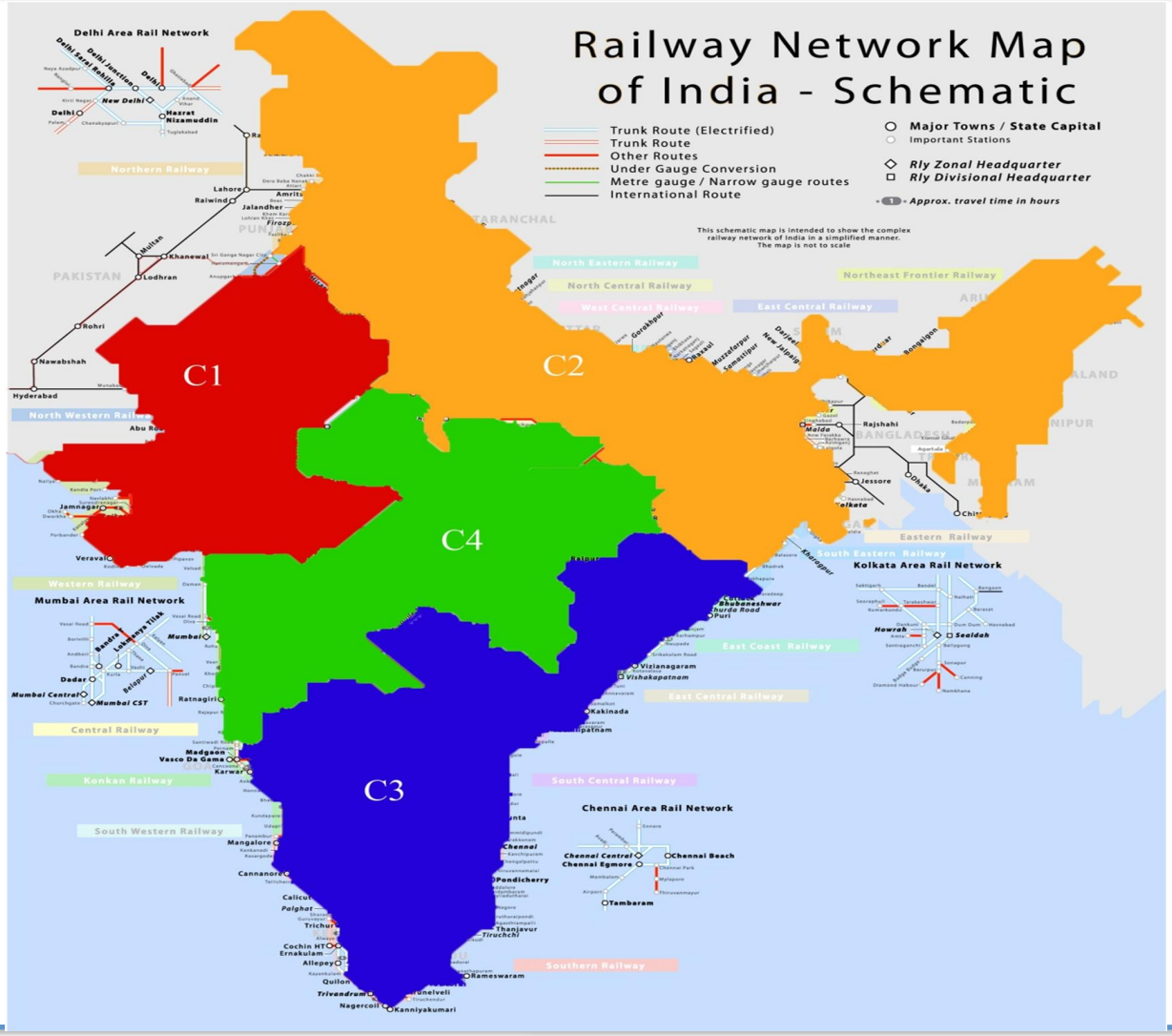


Community Detection

Railway Network [2]

- Hypergraphs – richer structure
- Stations are nodes
- Trains are hyperedges
- Find communities on the line graph
  - *Line graph:* Trains are nodes and if two trains stop at the same station they have an edge between them
- Convert to communities of nodes

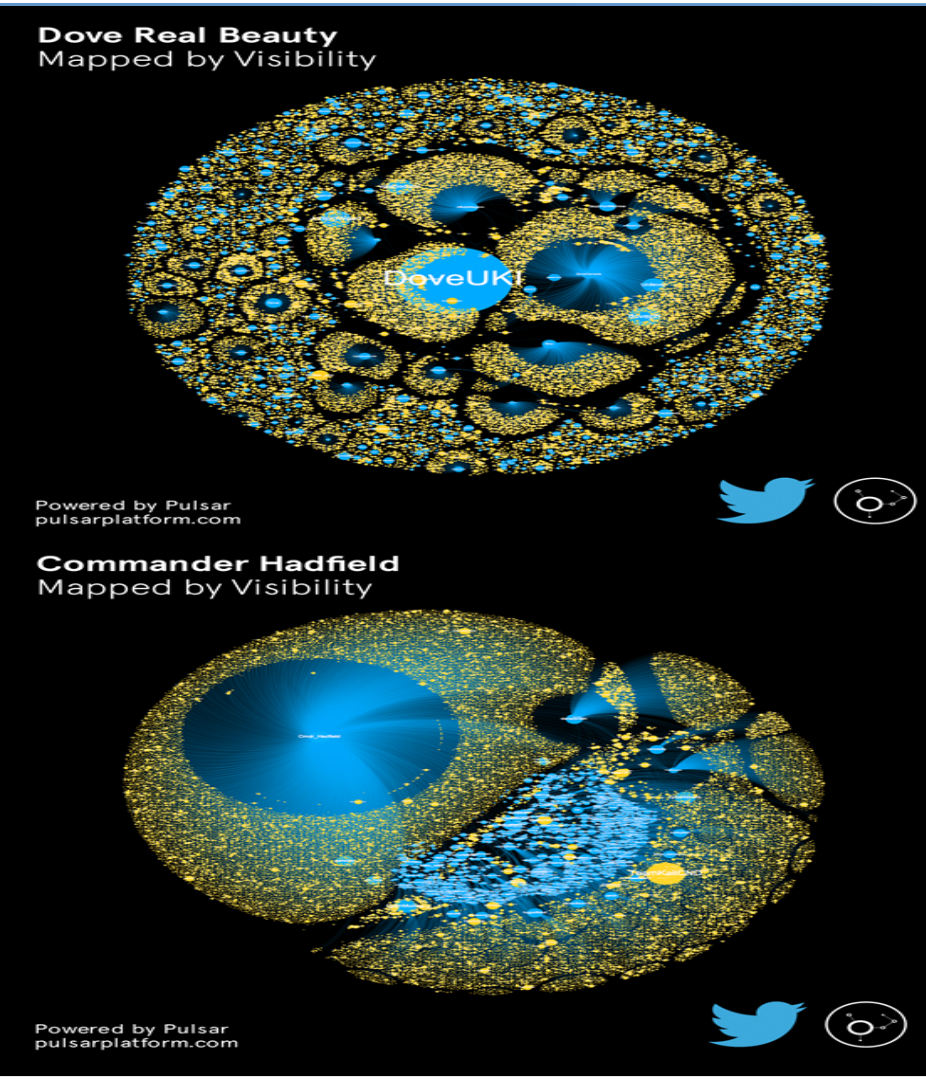
Communities and Zones

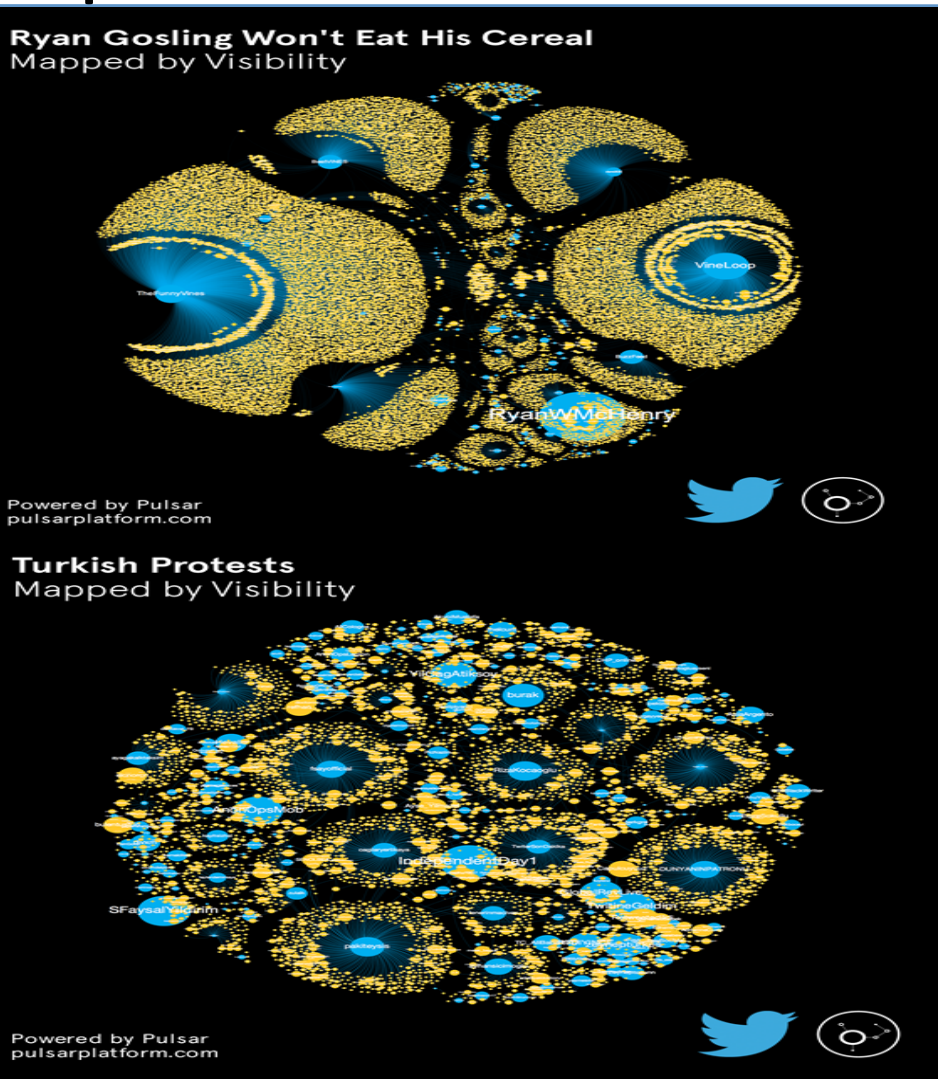


Positional Analysis [3]

- Certain nodes play important roles in a network
  - Can be a *bridge* node
    - Connect different localities
  - Can be a *broker* node
    - Connect different modalities
  - Can indicate lack of connectivity
- RIDεR: Find at scale [6]
  - Our algorithm works on networks with 10 million plus edges
  - Cannot achieve real-time constraints yet

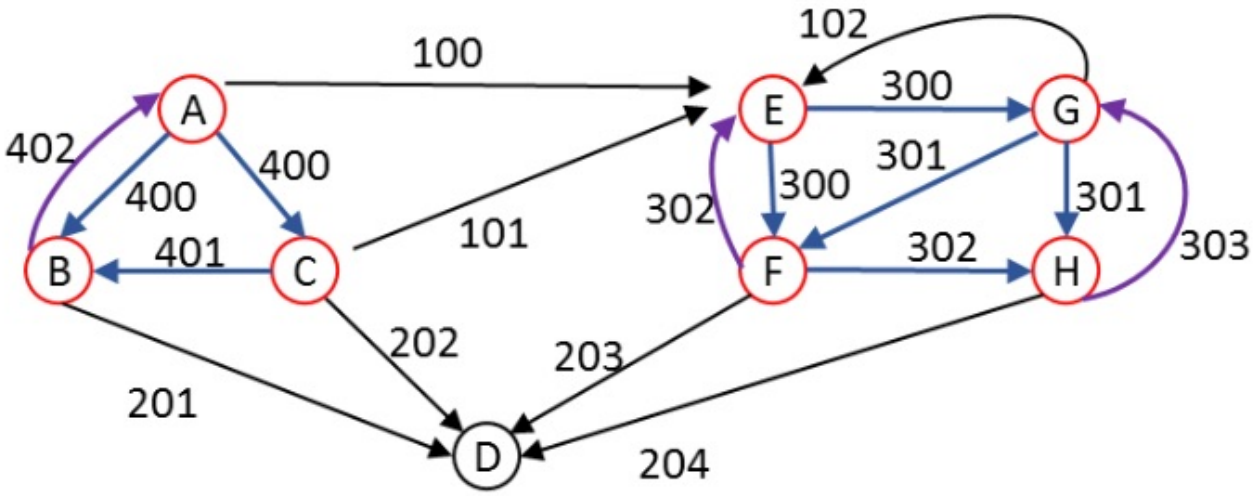
How information goes viral on twitter - Influencers and spread?





Blue nodes = tweeters. Yellow nodes = retweeters

Communication Motifs



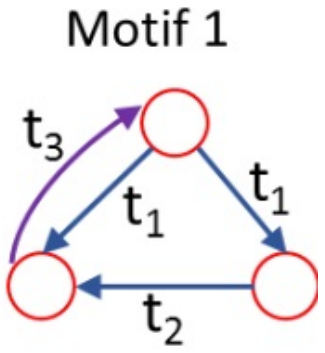


Figure 4 : A dynamic network. Motif 1 :$\{A, B, C\}$, $\{E, F, G\}$, and $\{G, H, F\}$.

Challenges

- Incorporating temporal information.
- Unlabelled Nodes - Search Space becomes Huge.

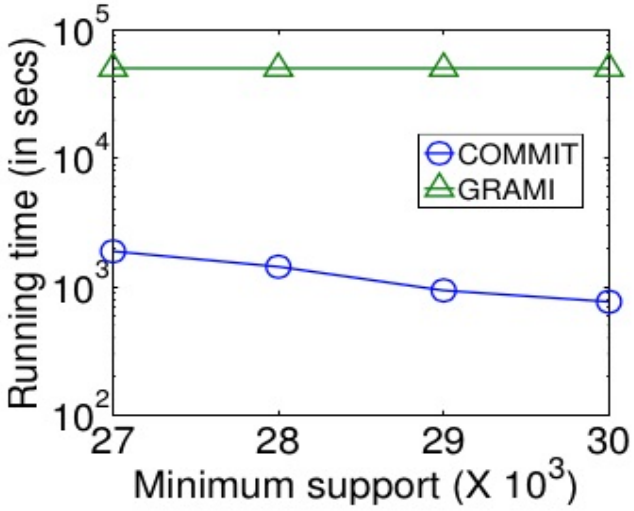


Figure : Running time comparison of GRAMI and COMMIT against the support threshold on the Facebook dataset

M. Elseidy, E. Abdelhamid, S. Skiadopoulos, and P. Kalnis. Grami: Frequent subgraph and pattern mining in a single large graph. Proceedings of the VLDB Endowment, 7(7), 2014.

Proposed Approach [7]

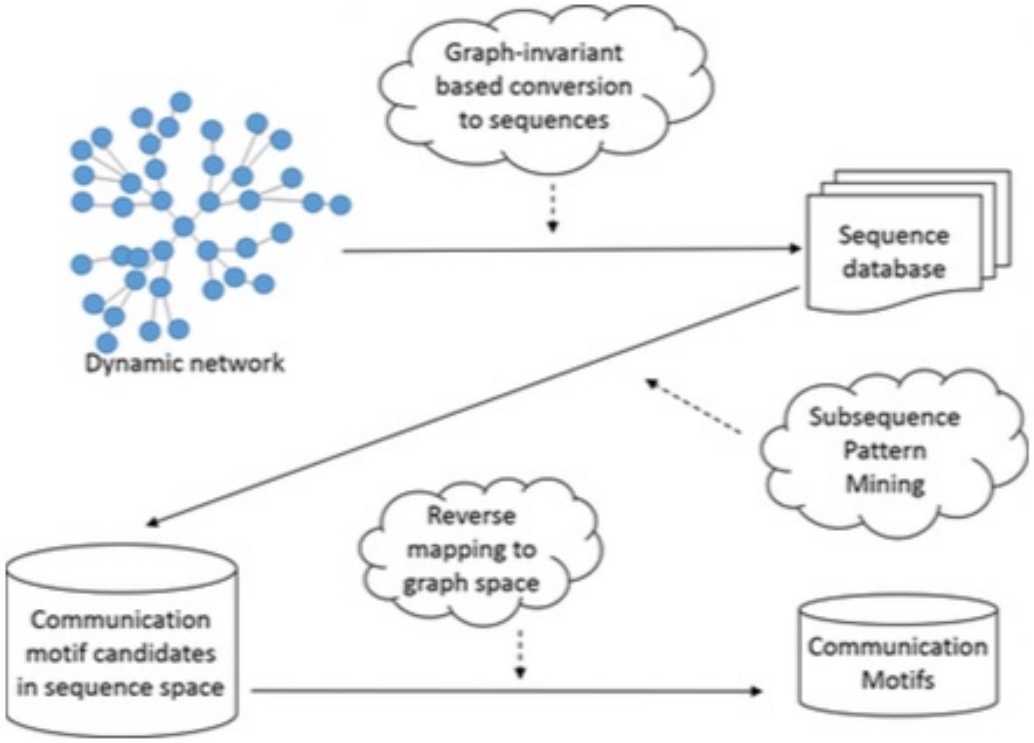


Figure : Pipeline of the COMMIT algorithm.

Mined Motifs

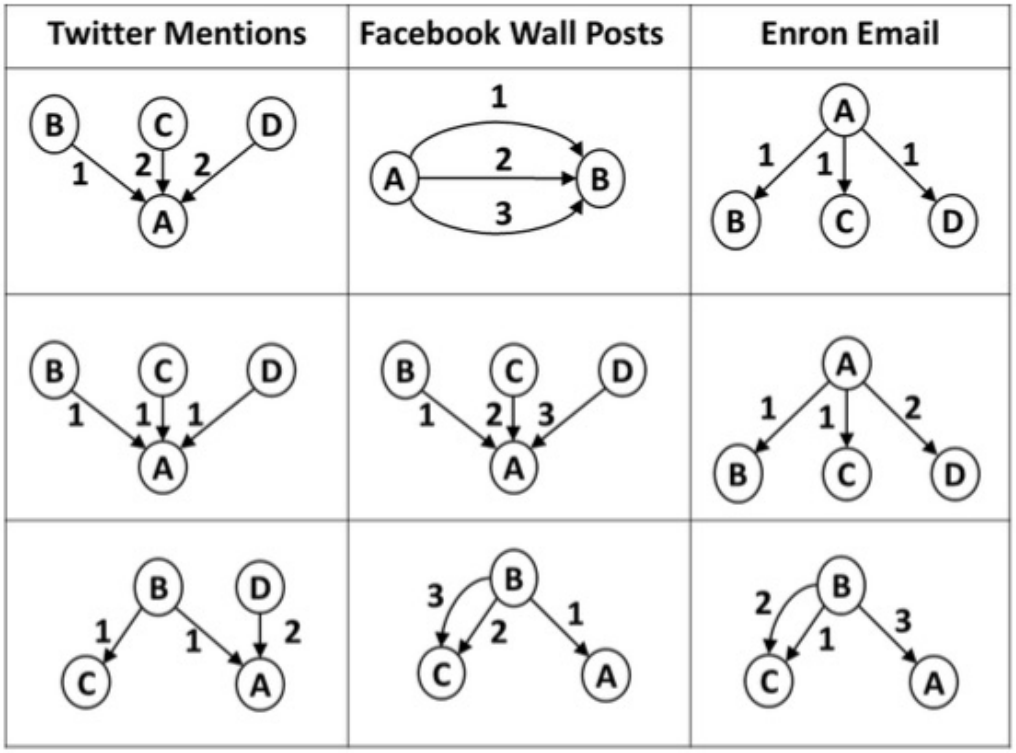


Figure : Communication motifs in Twitter, Facebook and Enron.

Further

- Network formation
  - Link prediction
  - Friend Recommendation
  - Provisioning
- Behavioral studies
  - Purchase behavior
  - Targeted marketing [12]
  - Epidemiology
- Technology enables the formation of large networks as well as tracking of activities on the networks
  - Exciting challenges as well as opportunities

RISE
Reconfigurable and Intelligent Systems Engineering

Learning on Networks

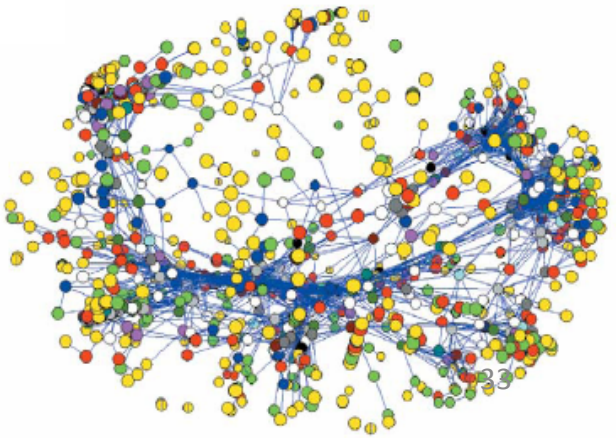
Collective Approach

- Traditional Machine Learning
  - Use attributes of data

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	?
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Collective Approach

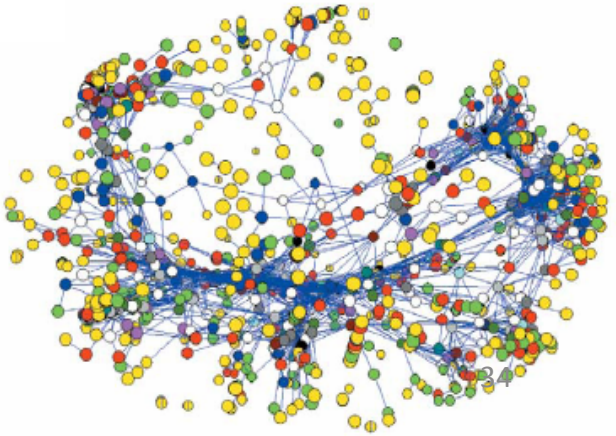
- Traditional Machine Learning
  - Use node attributes (content) alone
- Collective approaches
  - Use content and link information
- Label of a node depends on labels of neighbors on the link structure (graph)
  - *How many friends bought computers?*



Collective Approach

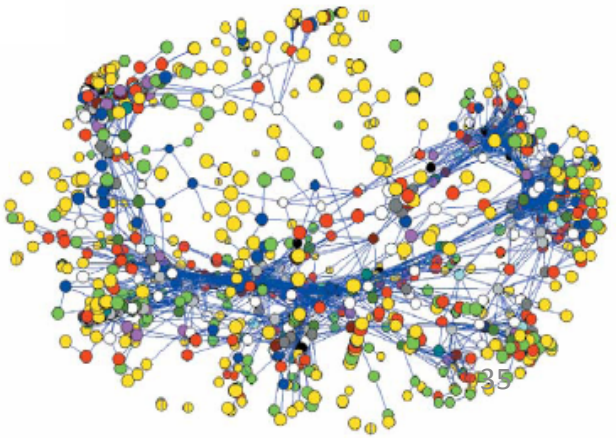
- Traditiona
  - Use node
- Collective
  - Use cont
- Label of a neighbors
  - *How many friends bought computers?*

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no



Collective Approach

- Traditional Machine Learning
  - Use node attributes (content) alone
- Collective approaches
  - Use content and link information
- Label of a node depends on labels of neighbors on the link structure (graph)
  - *How many friends bought computers?*
  - Web page text and hyperlinks
  - Content of papers and citations
  - Tweet text and followers



Two Applications

- Multi-grain sentiment analysis (MGSA)
  - Unweighted graph based Collective classification
- Protein Functional Site prediction
  - Signed graph based Collective Classification

Opinion Mining

- Sentiment expressions on some objects/ entities, e.g., products, events, topics, individuals, organizations, etc
  - E.g., "Maruti SX4 is well designed and feels like a thoroughly engineered car."
  - Has positive opinion on the "design of Maruti SX4".

Coarse grained opinion

CNET editors' review

CNET editors' rating:

Very good

Detailed editors' rating

Reviewed by:
David Katzmaier
Reviewed on: 02/26/2010
Updated on: 03/01/2010

***Editors' note, March 1, 2010**: Updated review to correct features and overall ratings.*

Photo gallery: Sony KDL-NX800 series

As the first official 2010 HDTV reviewed by CNET, and the first mainstream edge-lit LED-based LCD produced by Sony we've tested, the KDL-NX800 series arrives with plenty of anticipation. However, before you equate "LED" with "awesome picture quality," it's worth reiterating that the backlight technology comes in a bunch of varieties--and not all are created equal. The Sony NX800 performs on a par with other like-equipped LCDs that we've tested, such as the UNB7000 series from Samsung, so people seeking a premium home theater picture might be disappointed.

In other areas, the NX800 shines. Sony completely redesigned the exterior of its higher-end 2010 models in what it calls a monolithic style--and this TV would be at home near the Tycho crater or orbiting Jupiter. Sony also kept the superb selection of Internet services found on 2009 models, but adds built-in Wi-Fi to make them easier to use. All told, this svelte Sony feels more thoughtfully put-together than any TV we've tested in awhile, and it will easily find a niche in design-conscious living rooms.

Series note: We performed a hands-on evaluation of the 52-inch Sony KDL-52NX800, but this review also applies to the other sizes in the series, the 46-inch KDL-46NX800 and the 60-inch KDL-60NX800. The three sizes share identical specifications and should exhibit similar picture quality.

Fine grained opinion

Product summary

+ Add to my list

THE GOOD: Excellent design with stylish monolithic exterior; ergonomic remote control; snappy menu system; relatively accurate color; built-in Wi-Fi; solid Internet services including Netflix, Amazon Video on Demand, niche video services, and Yahoo Widgets; energy-efficient.

THE BAD: Relatively expensive; reproduces lighter black levels; darker areas tinged blue; cannot adjust dejudder processing much; less-even screen uniformity; glossy screen reflects ambient light; Netflix image quality worse than on other streaming devices.

THE BOTTOM LINE: Despite a picture that won't wow sticklers, Sony's edge-lit LED-based NX800 sets a high bar for its beautiful design and well-executed features.

Motivation

- Feature level and product level sentiments are dependent
  - Majority of features are positive -> product has positive sentiment
  - Product has a very poor review (say 1/10), then most of its features should have negative opinion
- Feature level dependencies : "The manual gear shifter is rubbery.." Can be disambiguated using "..has an unpleasant driving experience.."
- Neighborhood structure from domain knowledge base can be used for imposing relationship between features to generate a graph

Proposed Approach [9]

- Multi-view collective classification with label propagation
  - To the best of our knowledge there is no approach which combines collective classification + Domain Information
- Applied on automotive domain sentiment analysis dataset

Results

- MV used Naïve Bayes with a product rule for combining the votes of content only and link only classifiers
- Collection of 300 reviews from CNET, Epinions and Edmunds

sentiment	precision	recall	f1-score
positive	0.38	0.83	0.52
negative	0.61	0.16	0.25
average	0.5	0.5	0.38

ICA-MGSA

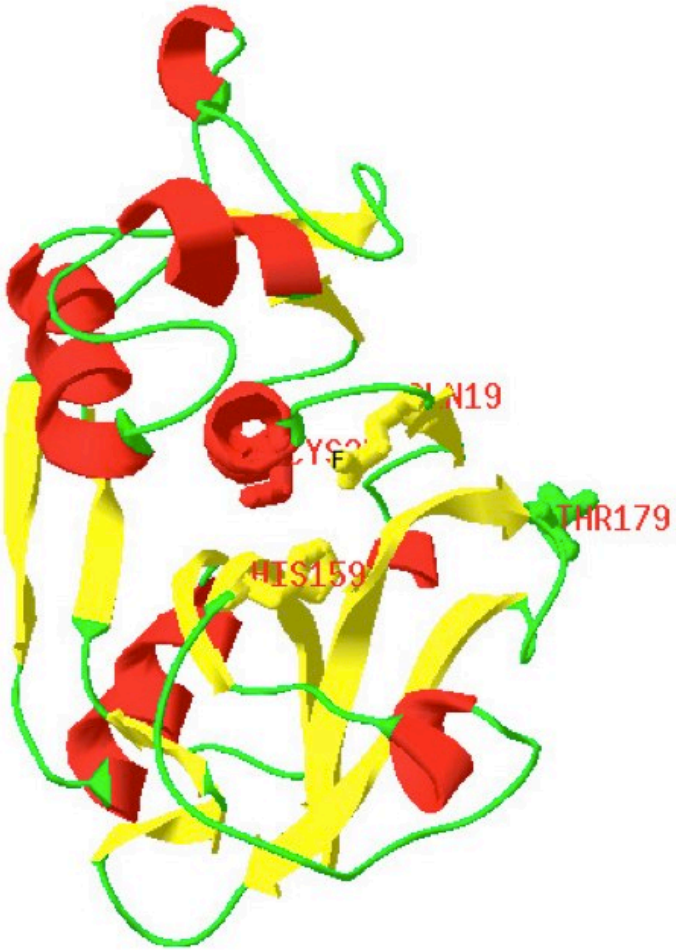
sentiment	precision	recall	f1-score
positive	0.77	1	0.87
negative	0.97	0.14	0.25
average	0.82	0.78	0.71

Bootstrapping

sentiment	precision	recall	f1-score
positive	0.96	0.77	0.86
negative	0.69	0.94	0.8
average	0.87	0.83	0.84

MV-MGSA

Functional Site Prediction [10]



- Given a protein structure with n amino acid residues, find its functional residues
- Experimental determination of each choice takes about one to two months.
- Computational methods that provide a list of highly likely functional residues are extremely useful for experimental biologists.

Features of Functional Residues

- Usually occur at the surface of proteins
- Conserved in evolution
- Shows preference for specific amino acids, secondary structure
  - Local/Attribute information
- Spatially proximal or in contact in structure; may not be close in sequence
  - Link information
- Approximately 0.7% to 1% of residues are functional residues.

Results

- CATRES-FAM dataset containing 140 enzymes
- 471 out of 49180 residues are functional
- Precision at 69% Recall and Recall at 18% Precision 69% recall at 18% precision was state-of-the-art

Method	$Precision_{69}$ (%)	$Recall_{18}$(%)
Collective Inference	**23.06**	**87.78**
Discern	18	69
CRF	18	69
Local classifier	14.38	58

Future

- Attributed graph data is ubiquitous
- Develop more efficient scalable and robust algorithms
  - Multi-relational data
  - Missing attributes/links
  - Time evolving links
- Connect the two strands of works
  - Use network properties as node attributes
  - Explain/predict performance based on network characteristics
- Wide applicability and interesting theoretical challenges as well!

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 2014+
3. Gupte, P. V. and Ravindran, B. (2014) "Scalable Positional Analysis for Studying Evolution of Nodes in Networks". SDM 2014+
4. 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 (Under Review)
5. M. Vishnu Sankar, Balaraman Ravindran and S. Shivashankar, "Fast Method for Finding Resolution Limit Free Communities in Large Networks," IJCAI 2015 (Under Review)
6. Gupte, P. V. and Ravindran, B. "RIDeR: Scalable Role Discovery in social networks". IJCAI 2015 (Under Review)

References

7. Gurukar, S., Ranu, S., and Ravindran, B. (2015) “COMMIT: A Scalable Approach to Mining Communication Motifs from Dynamic Networks”. SIGMOD 2015
8. Roy, S., and Ravindran, B. (2015) “Measuring Network Centrality Using Hypergraphs”. CoDS 2015 (Best Student Paper).
9. Shivashankar, S., and Ravindran, B. (2010) "Multi Grain Sentiment Analysis using Collective Classification". ECAI 2010.
10. Kar, S., Deepak, V., Ravindran, B., and Tendulkar, A. V. (2012) "Functional Site Prediction by Exploiting Correlations between Labels of Interacting Residues". ACM BCB 2012.
11. Vijayan, P., Shivashankar, S., and Ravindran, B. (2014) "Multi-label Collective Classification in Multi-attribute Multi-relational Network Data". ASONAM 2014.
12. Pasumarthi, R. K., Narayanam, R., and Ravindran, B. (2015) "Near Optimal Strategies for Targeted Marketing on Social Networks". AAMAS 2015.

Thank You!

Questions?

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

Acknowledgements:

ERICSSON

GM