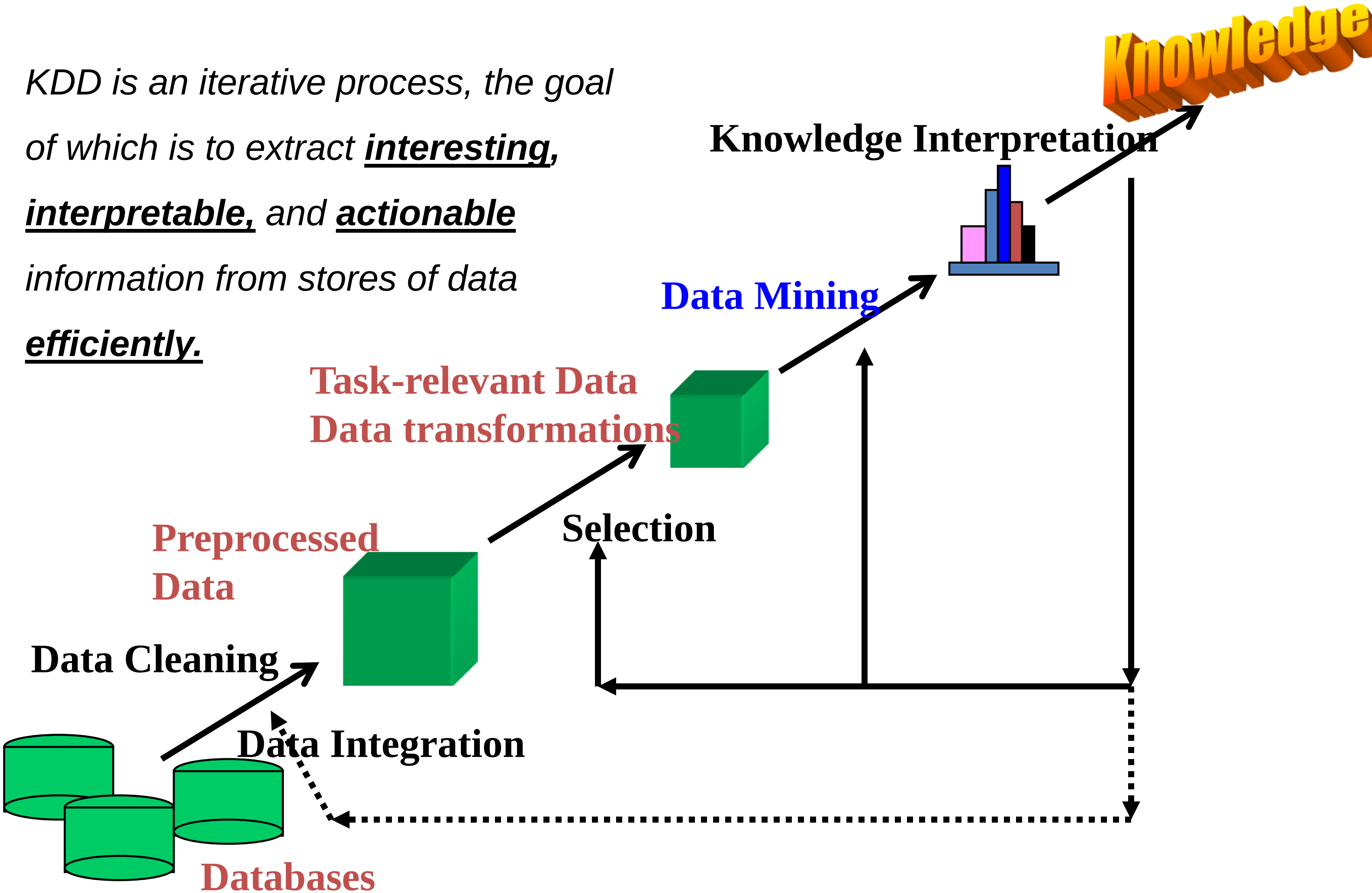
Toward Visual Knowledge Discovery and Analytics

Srinivasan Parthasarathy
The Ohio State University

srini@cse.ohio-state.edu

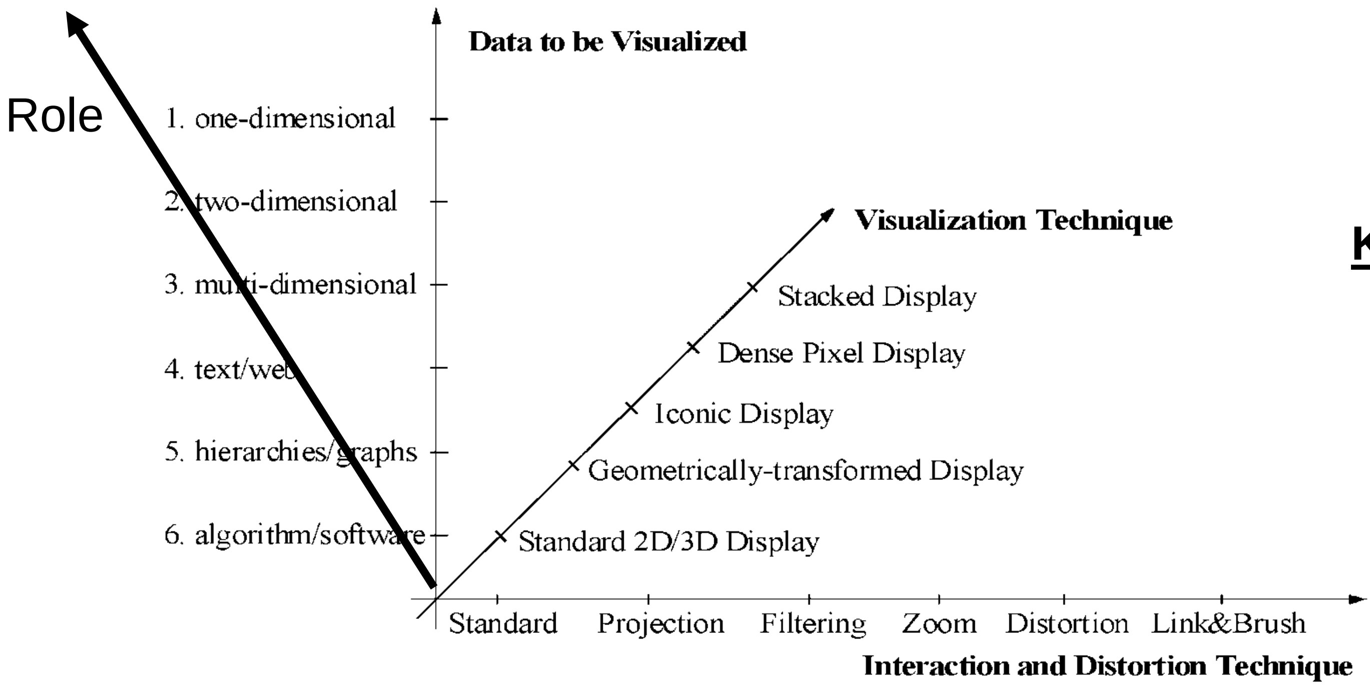
Knowledge Discovery Process

KDD is an iterative process, the goal of which is to extract ***interesting, interpretable,*** *and* ***actionable*** *information from stores of data* ***efficiently.***



Information Visualization and KDD

- Why? [Fayyad et al 2000, Sneidermann 2008]
  - Human in the loop
  - Efficient and effective knowledge discovery



Keim 2002

Roles for Visualization in KDD

1. As a basic method to visualize data and information
   - This has been the focus of much of the work to-date
2. As an approach to lend transparency to the knowledge discovery process
3. As a mechanism to validate patterns unearthed by discovery process
4. As a method to tightly integrate with the discovery process to enable visual-exploration

2, 3, and 4 will be discussed next

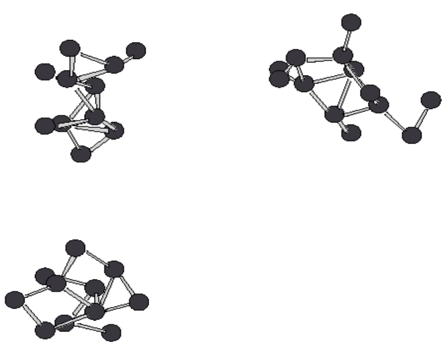
Case Study I
Analyzing Scientific Simulation Data

Visualization Role: Pattern Validation and Verification

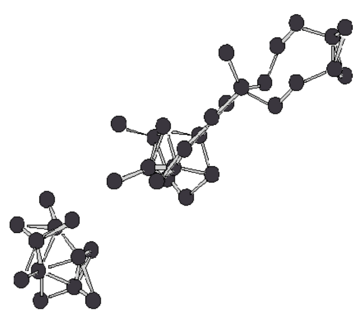
Acknowledgements: **S. Mehta, H. Yang**, R. Machiraju (Viz), and J. Wilkins (Phy)

Motivation

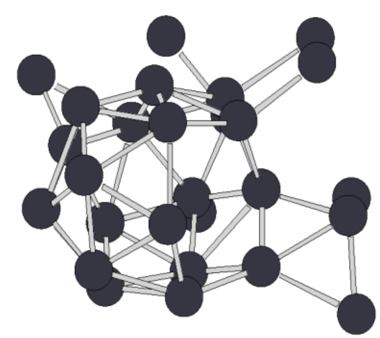
- Defect structures affect properties/performance of materials
  - Silicon chips, Titanium Alloys etc.
- Understanding the evolution of defect structures is important
  - Formation of elongated defects, cracks etc.
- Analyze from large scale Molecular Dynamics Simulations
  - Used for many other problems (e.g. protein folding)



1st time frame



19,000th time frame



130,000th time frame

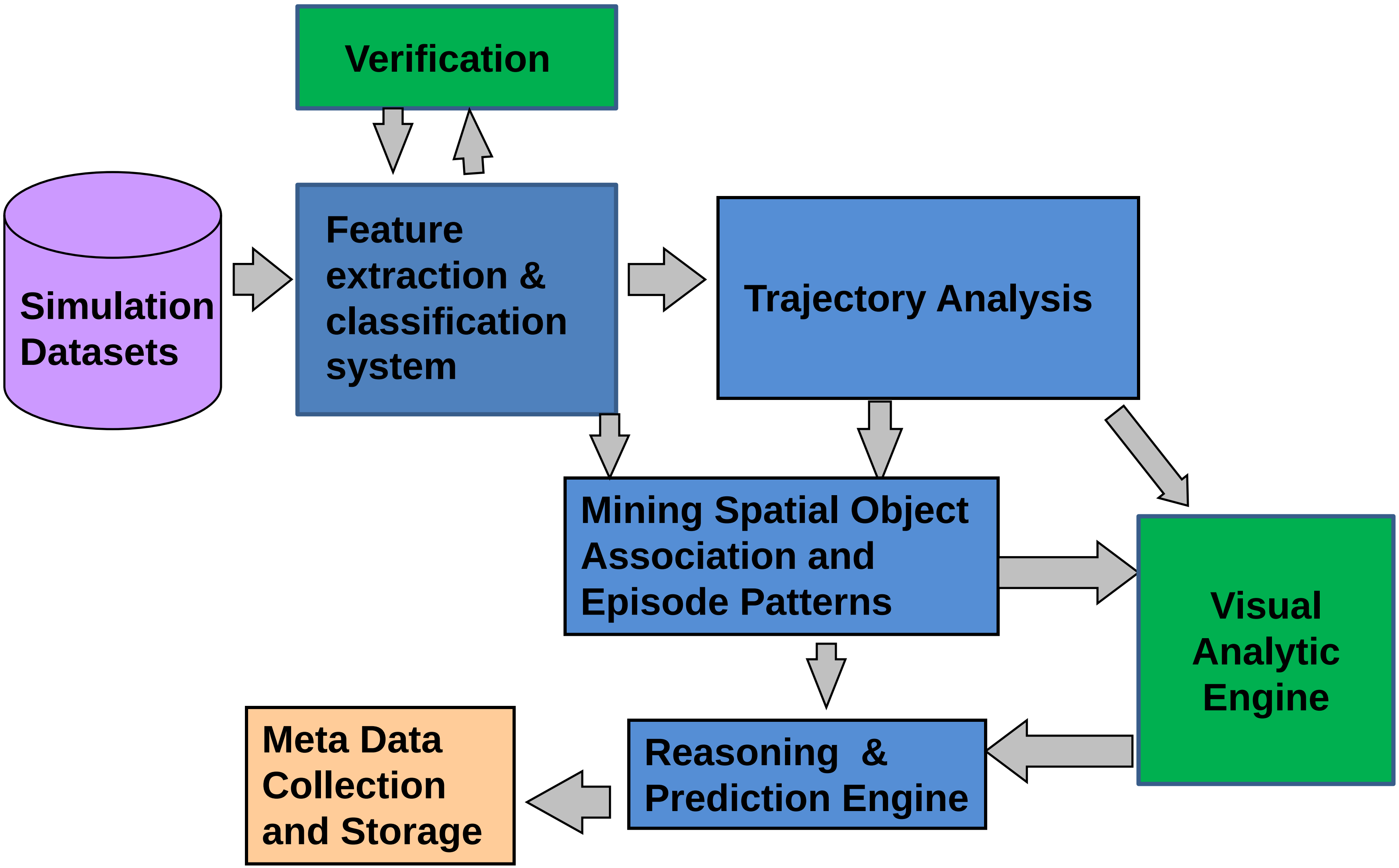
Challenges and Objectives

- Challenges
  - Large data (GB/TB range)
  - Dynamic time-varying data
  - Noisy data (thermal noise)

- Objectives
  - Characterization of Defects (detection, classification)
  - Characterization of Interactions and Evolution (spatio-temporal patterns)
  - Need to enable real-time steering and verification

- Role for Visualization
  - Verification of defect structures and class labels
  - Visualization of spatio-temporal interactions

Framework Details



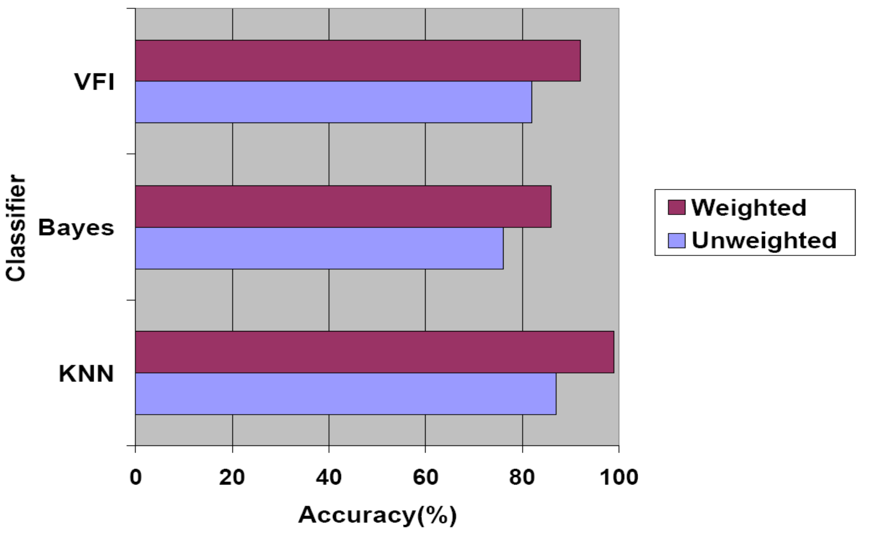
Verification Objectives

- Goal is to help validate results
  - Need to limit number of defects presented to user
    - Cannot possibly show all
    - Need to limit corridor of uncertainty
      → more effective classification
  - Need to efficiently identify best way to visualize data
  - Need to support multiple views

Limiting the Corridor of Uncertainty

2 stage classifier – first stage narrows down candidate classes second stage performs and exact match

Build accurate classifier -- use biased sampling to display mostly defects one is uncertain about (e.g. new defects).

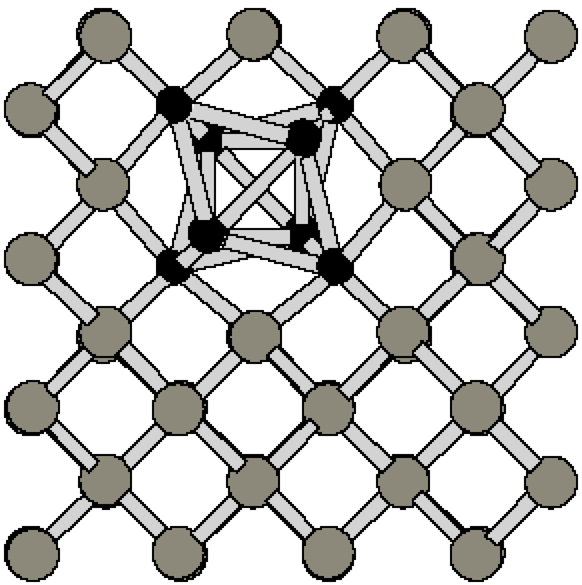


Data	#Frame	Sz (GB)	#Atoms	#Def	#Unique
Two I	512,000	4	128	350,000	2841
Three I	200,200	6	512	320,000	1,543
Four I	297,000	11	1,024	410,000	3,261

Verification: Basic Strategies

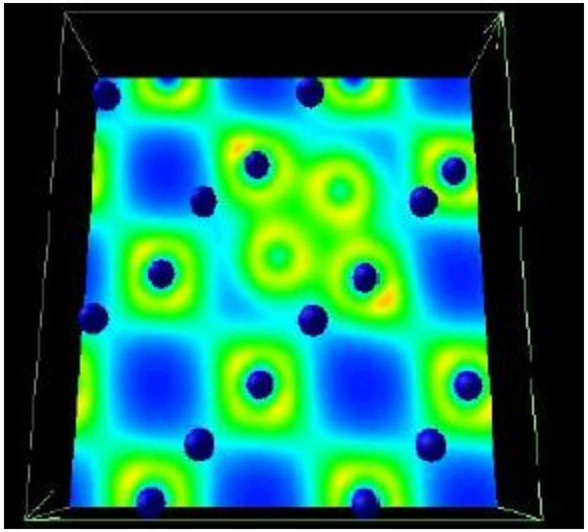
Ball and Stick Model

- Pros: Efficient, simple
- Cons: Hard to visualize in large lattices, does not model uncertainty



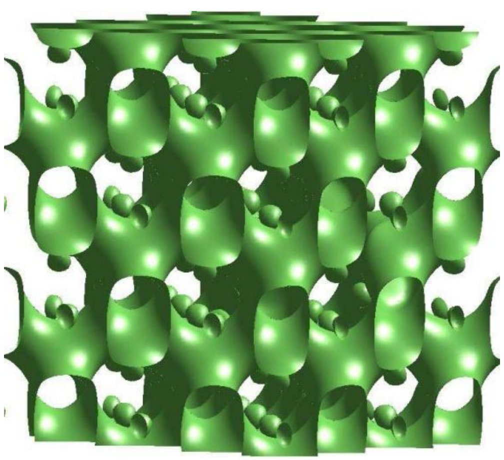
Electron Density Maps

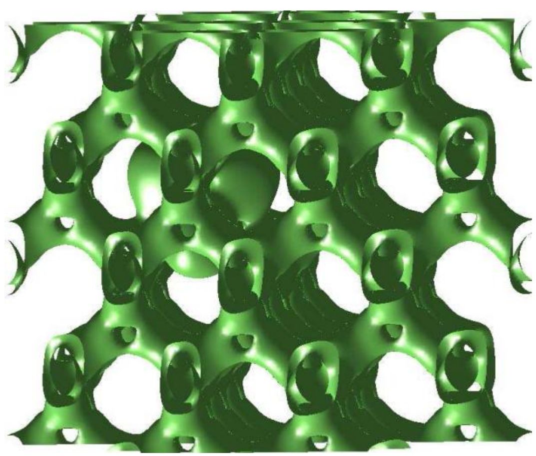
- Pros: Efficient, simple, models uncertainty
- Cons: Requires viewing by slicing, interactivity constraints

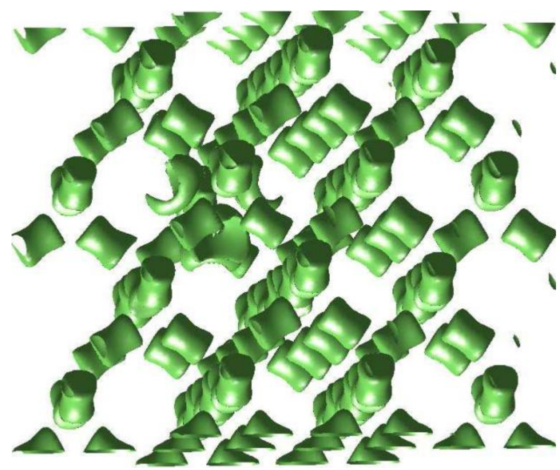


Verification: Isosurface Modeling

- Pros: Enables viewing through layers.
- Cons: requires finding the right iso-value, higher complexity.
- Computing iso-value
  - Should cleanly show and differentiate defect and base atoms
  - Relied on domain (electron density scalar field)
  - Found isovalue ~ 450 electron density to be the best (middle)

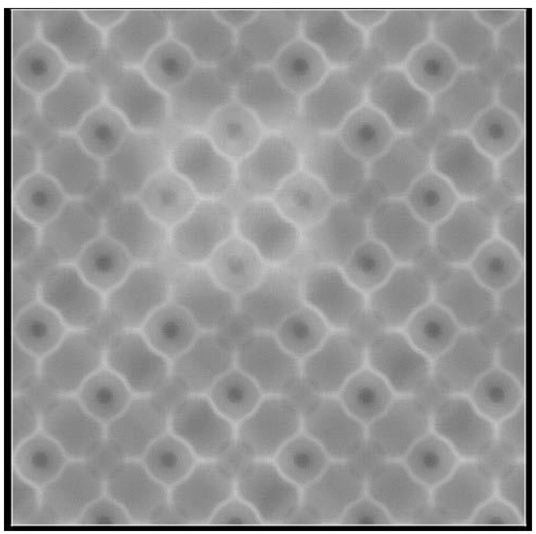


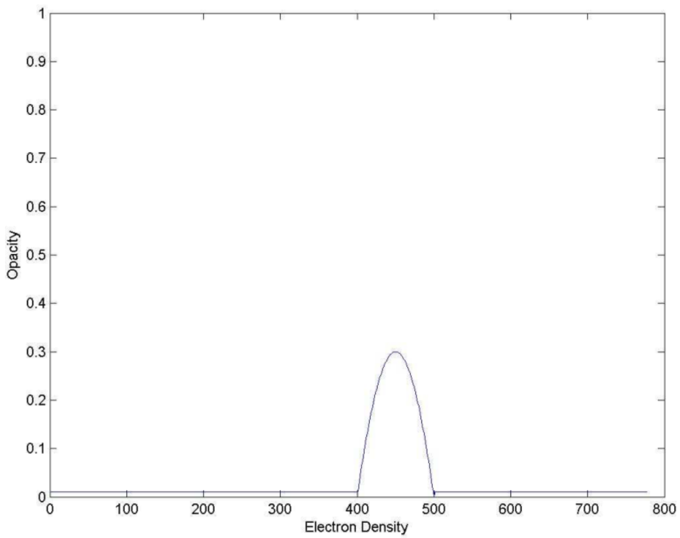


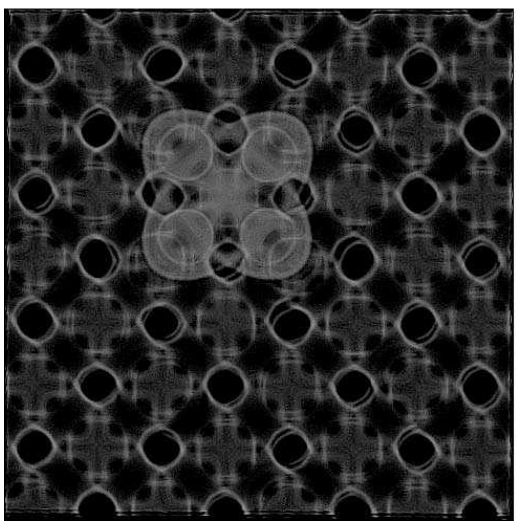


Verification: Volume Rendering with Transfer Functions

- Pros: Enables viewing through material, models uncertainty.
- Cons: Complexity, constructing transfer function
- Transfer function with a small Gaussian near 450







Take Home Message

- Visualization can help validate patterns extracted and promote computational steering
- Can also help visual analytics
  - Spatio-temporal visual analysis (not discussed)
- Generalizations
  - Feature Mining and Visualization for Fluid Flow Simulations
    - Aircraft Wing Modeling
    - Respiratory Systems (e.g. to study impact of Anthrax)
- Impact: New scientific discoveries, better understanding of underlying phenomenon.

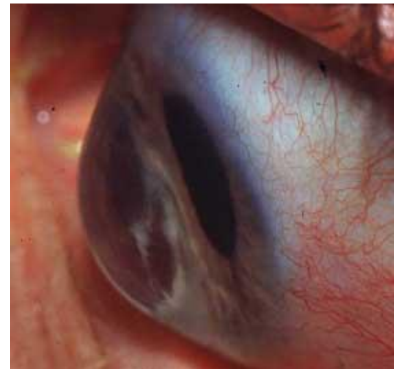
Case Study II
Clinical Diagnosis of Keratoconus

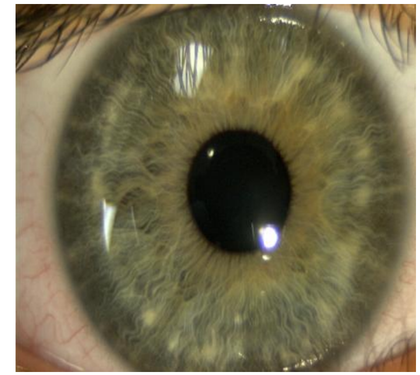
Visualization Role: Transparent Knowledge Discovery

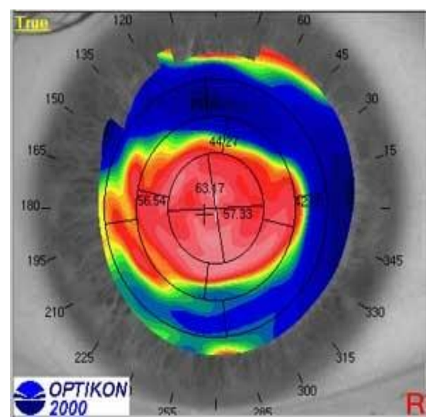
Acknowledgements: **M. Twa, K. Marsolo**, M. Bullimore (Opt)

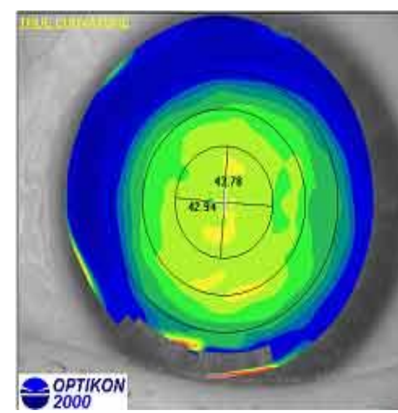
Case Study: Keratoconus

- Progressive, degenerative, non-inflammatory disease.
  - A leading cause of blindness and corneal transplant.
- Early detection is difficult & important
  - Has implications for eye surgery and control-of-disease
  - Initial Symptoms: Minor fluctuations in corneal shape
- Diagnosis procedure
  - Video-keratography exam
  - Manual analysis of results by clinician
- Challenges to detection
  - Voluminous data
    - one image is 1000s of data points representing corneal surface
    - spatial and temporal (longitudinal)
  - Features of interest small in scale to mean shape
  - Leads to variance in prognosis







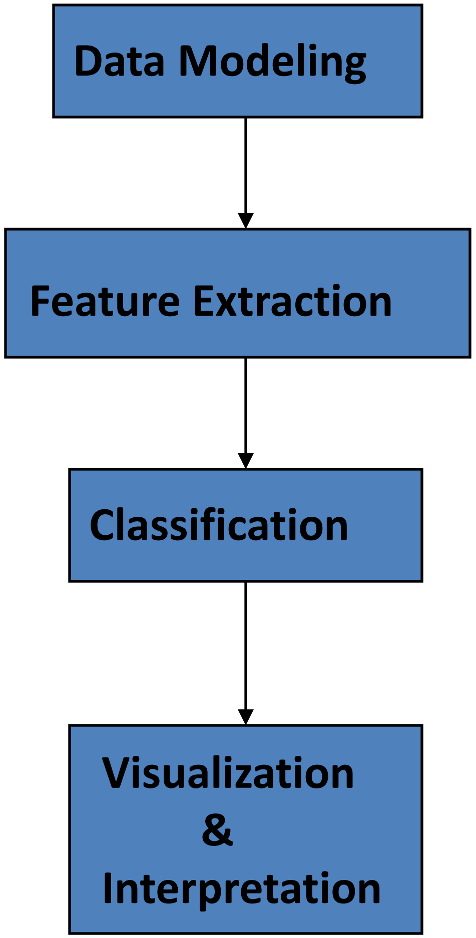


Late stage Keratoconus

Normal (clinically ideal)

Desiderata for Clinical Diagnosis

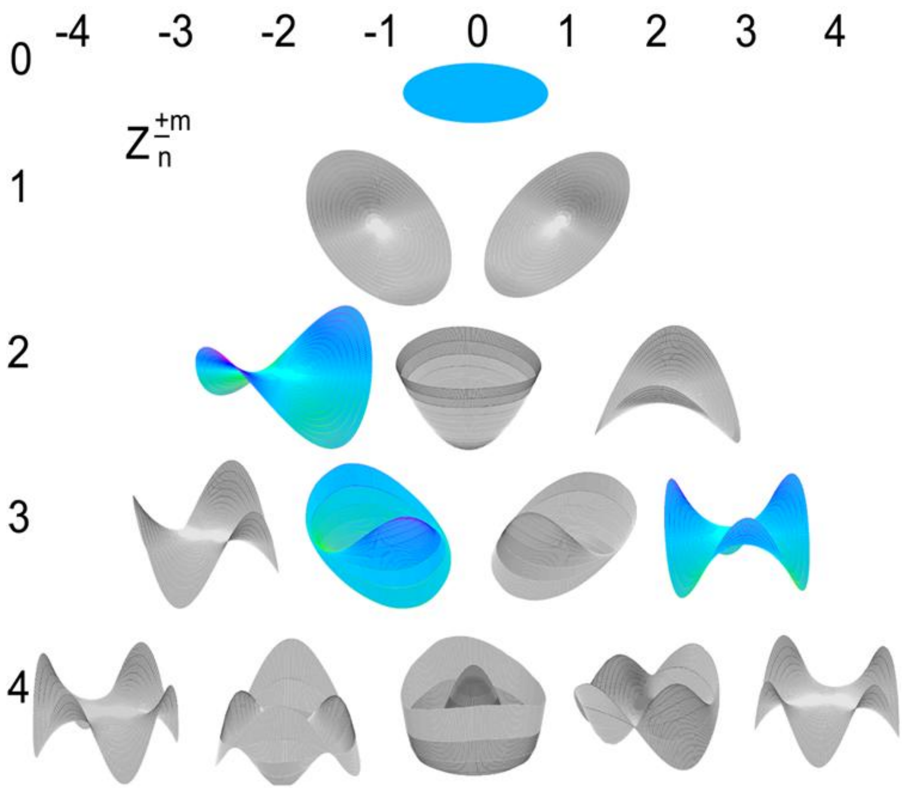
- Should be **accurate** and ideally interoperable
  - Can we use mathematical modeling?
- **Should be interpretable**
  - Can we visualize the decision making process effectively?
  - To a clinician very important
  - They do not like black box models!
- Should be **responsive**
  - Modeling step and discovery process can potentially be expensive



Synopsis of Approach

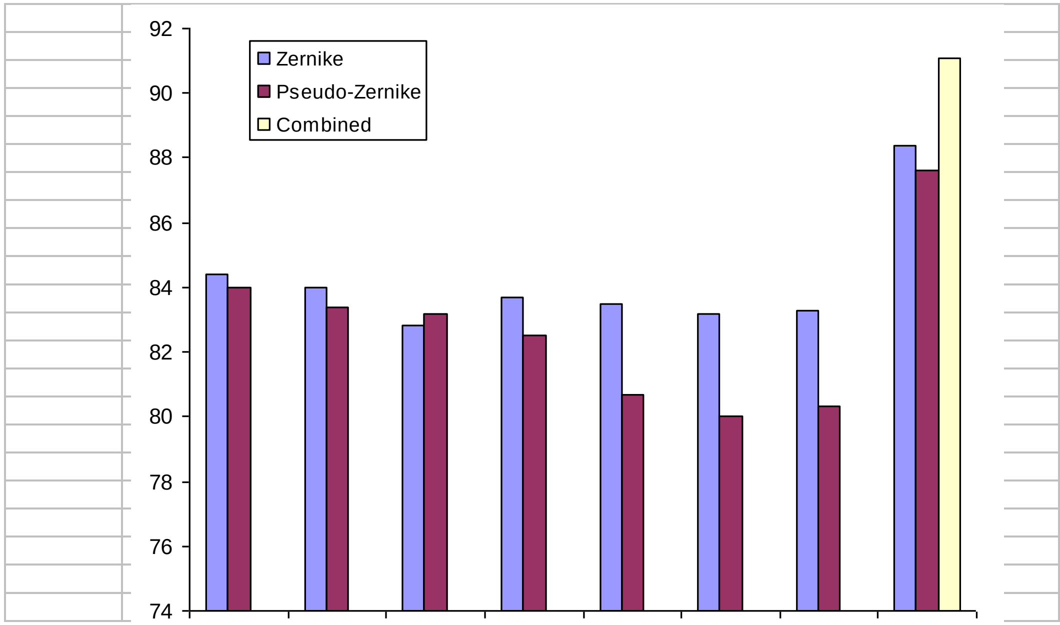
Modeling Corneal Shape with Zernike

- Hyper-geometric radial basis functions
  - Each term (mode) in the series represents a 3D geometric surface.
  - Orthogonal building blocks
  - Lower order →basic shape
  - Higher order → local harmonics
  - Compact representation
  - **Anatomic correspondence to clinical concepts**



Key Ideas

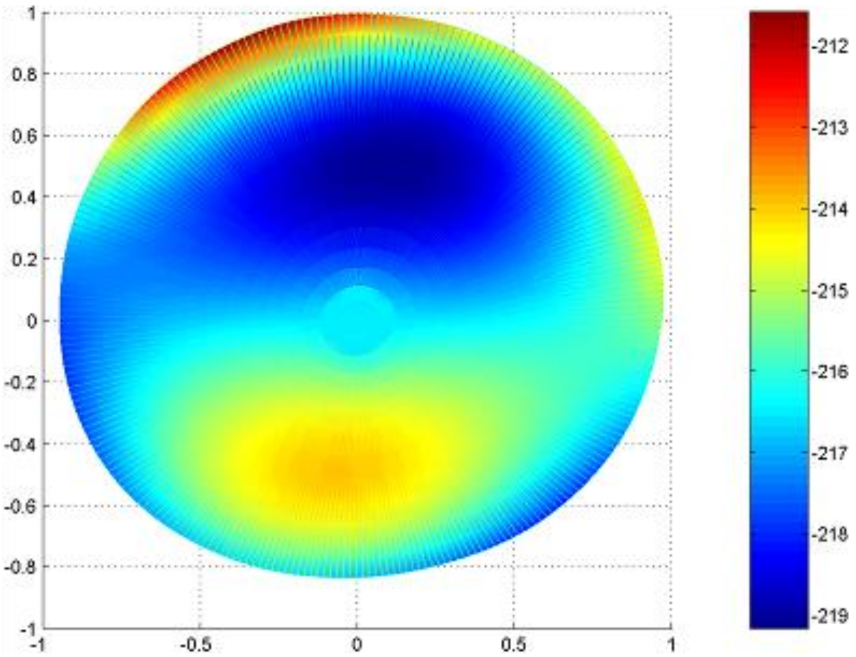
- Model data using Zernike and variant (Pseudo Zernike)
- Use coefficients derived as features
- Train classifier
  - Decision Trees work great
- Data
  - 254 Patient Records
  - Normal (119)
  - Diseased (99)
  - Post-LASIK (36)



- Accuracy **> 91% (with more information >95%)**
- **Decision trees are relatively easy to understand but can we do better in terms of lending transparency to the process?**

Visualization of Results

- Task: Visualize results to provide decision support for clinicians.
  - Give intuition as to why a group of patients are classified the way they are.
  - Contrast an individual patient with others in the same group
- How?
  - Modes of Zernike/Pseudo-Zernike polynomial correspond to specific features of the cornea.
  - Can use as building blocks.

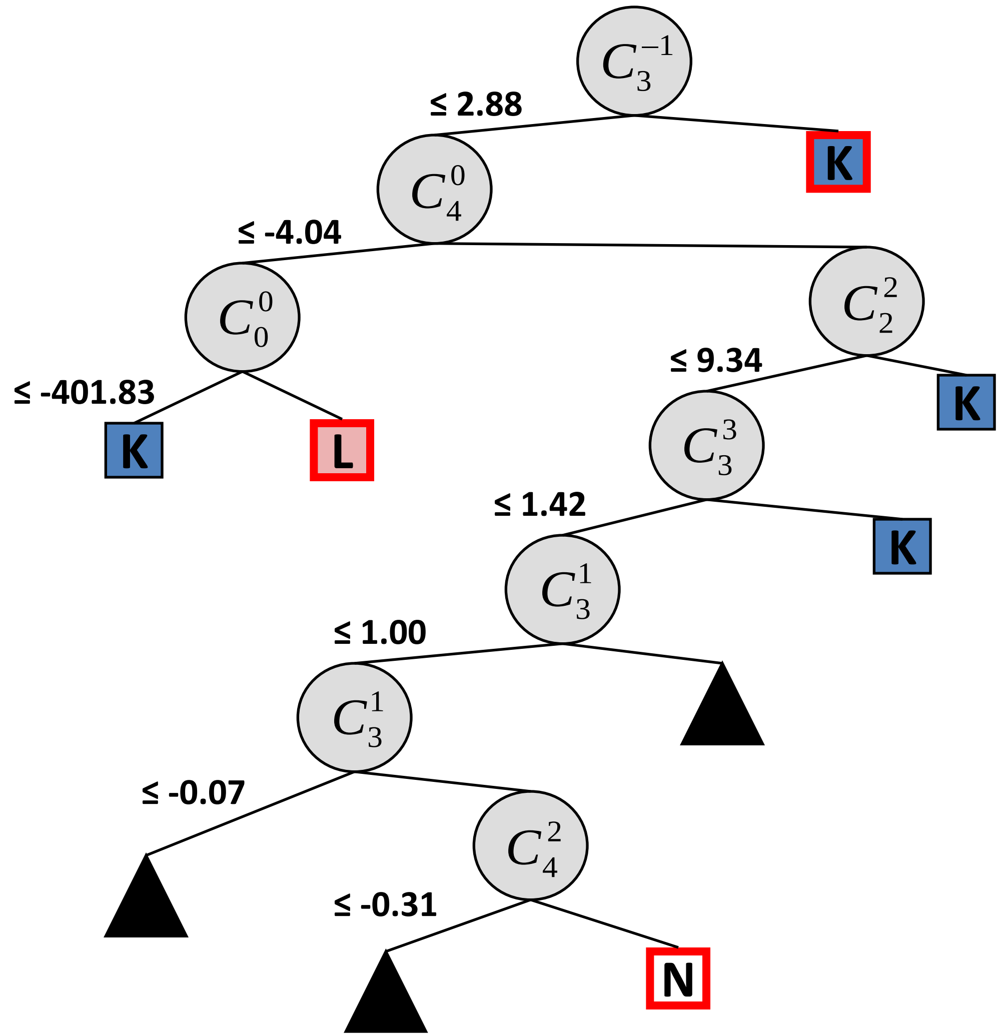


Patient-Specific Decision Surface

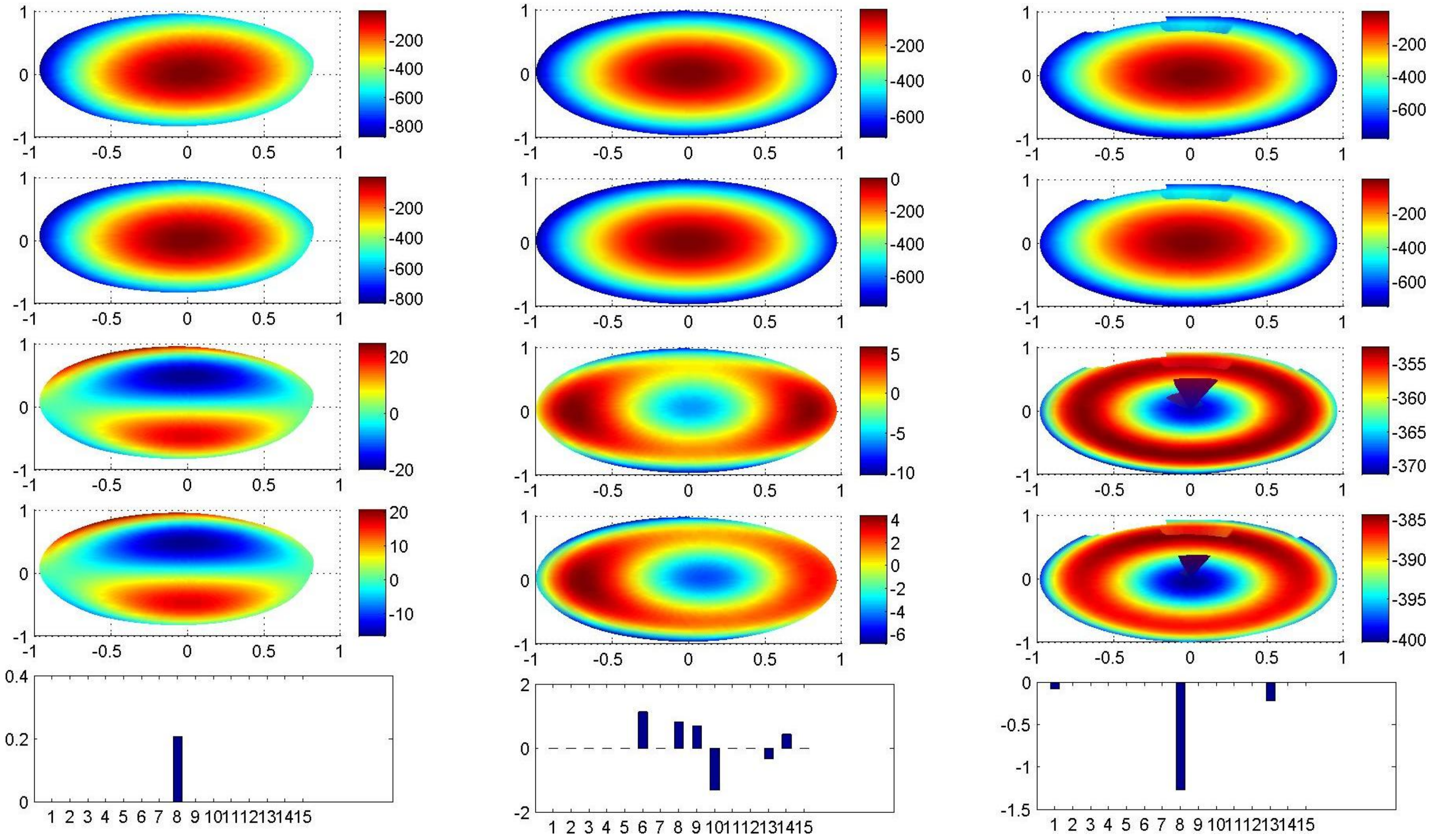
1. Treat each path through the decision tree as a 'rule.'
2. Cluster training data by rule.

For each patient:

1. Compute patient surface
2. Compute cluster surface → average coefficient values for all patients in cluster.
3. Compute patient "rule surface" → keep the 'rule coefficients', set others to zero.
4. Compute cluster "rule surface"
5. Compute deviation bar chart → relative error from rule mean coefficients



Visualization: Strongest Rules



Rule 1 - Keratoconus   Rule 8 - Normal   Rule 4 - LASIK

Take Home Message

- Visualization as a mechanism that lends transparency to the discovery process.
- Generalizations
  - The idea of rule-surfaces can be exploited for other problems where features are extracted from orthogonal generative models
    - E.g. Wavelet, FFT features etc.
- Impact: Clinical trials – new treatment protocols – improving quality of life

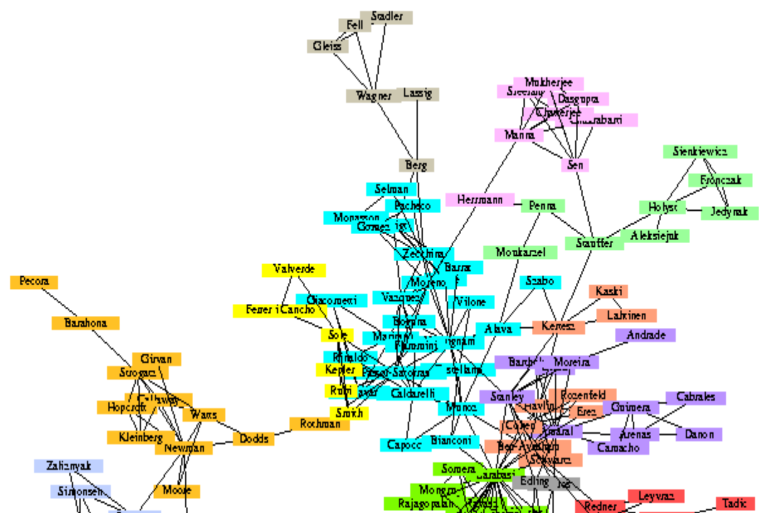
Case Study III: Analyzing Interaction Networks

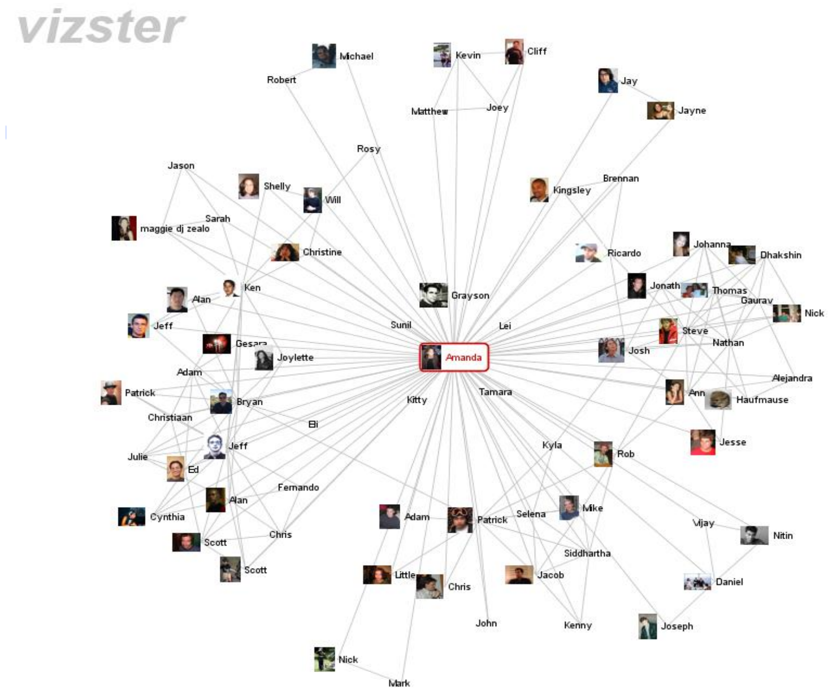
Visualization Role: Exploratory data analysis

Acknowledgements: **S. Asur, D. Ucar, V. Satuluri, X. Yang, N. Wang** (NUS),
S. Mehta (IBM) K. Tan (NUS), A. Tung (NUS)

Problem Domain(s)

- Interaction Networks
    - Nodes represent entities
    - Edges represent interactions among entities
  - Examples Abound:
    - Biological Networks
    - Collaboration/Friendship networks
  - Challenges
    - Community Discovery
    - Scale
    - Dynamic Nature
    - Visualization





Questions & Challenges

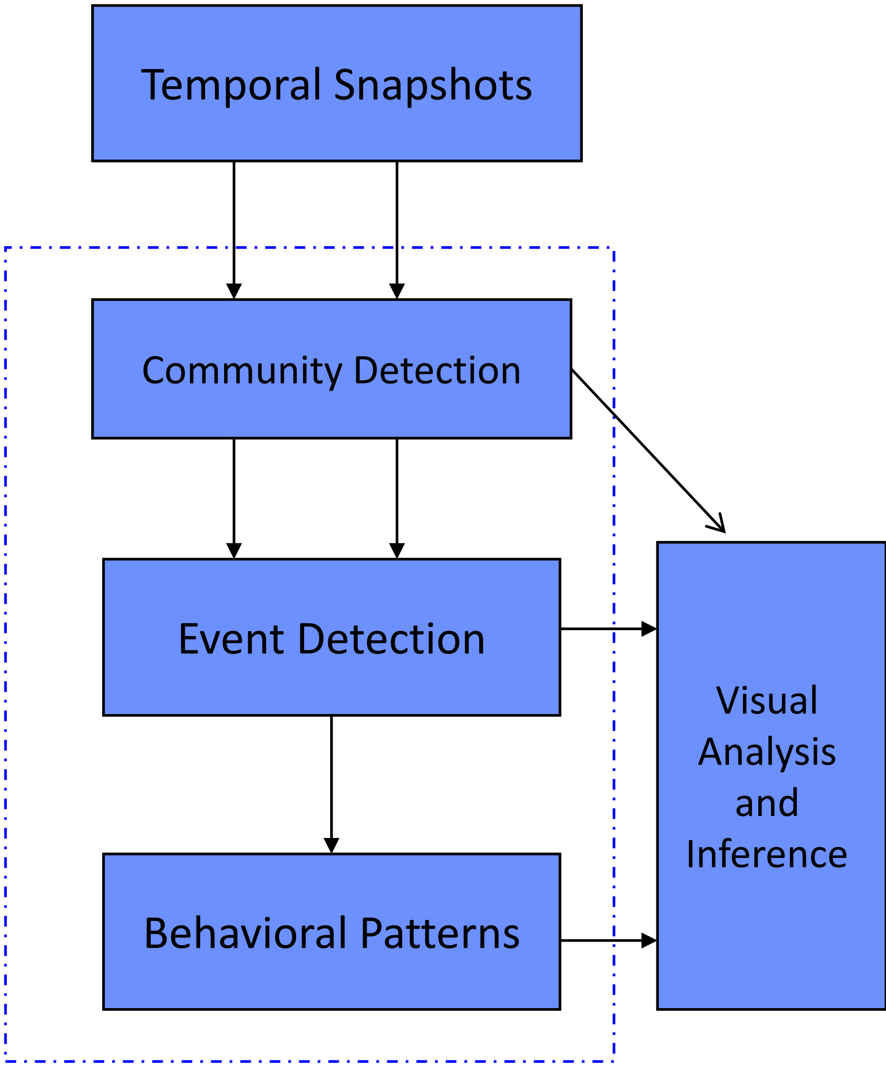
- How to extract modular structure?
  - common functional proteins, stable collaboratories etc.
- What characterizes stability of groups over time?
- What are the behavioral characteristics of nodes and communites:
  - Which nodes are influential, which are bridging, which are sociable, which are followers?
- What are the inter-relationships among communities?
- Challenges:
  - How to visualize?
  - Scalability (time, display)

Dynamic Analysis Framework

- Community Detection
  - MLR-MCL (KDD'09)
  - Viewpoints (KDD'09)
  - Graph Partitioning (Metis)
  - CSV (SIGMOD'08)

- Event detection (KDD'07, TKDD'09)
  - Entity Driven Events
  - Community Driven Events
  - Composing Behavioral Measures
    - Stability, Sociability, Influence

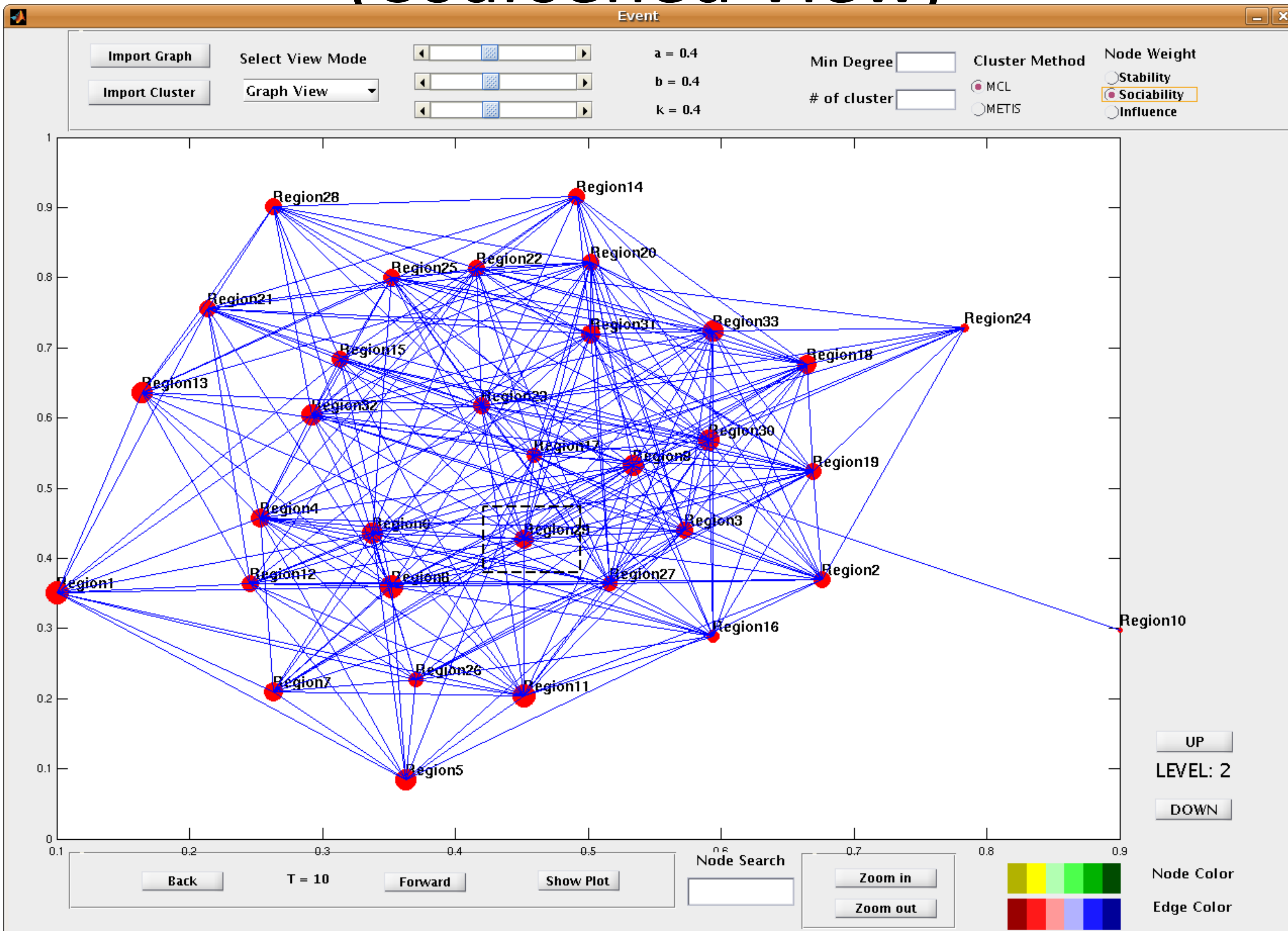
- Visual Analysis and Inference



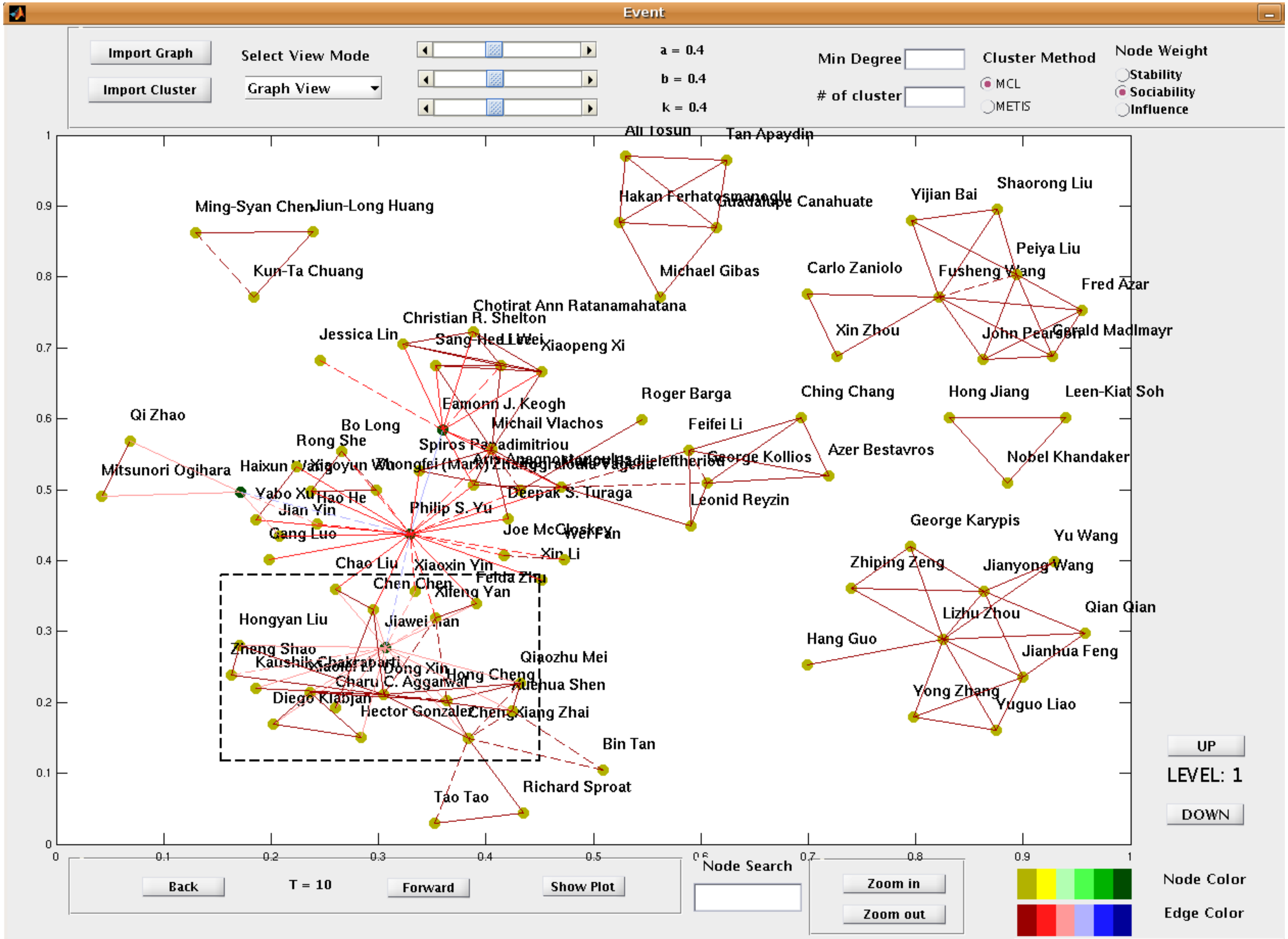
Visualization Challenges

- What to show?
  - Raw network, Coarsened view, Exploratory nugget (e.g. density plots), Event-driven view
- How to show it?
  - Layout and interface challenges
- How do we handle dynamism?
  - Efficiency
  - Mental Map/Cognitive Correspondence

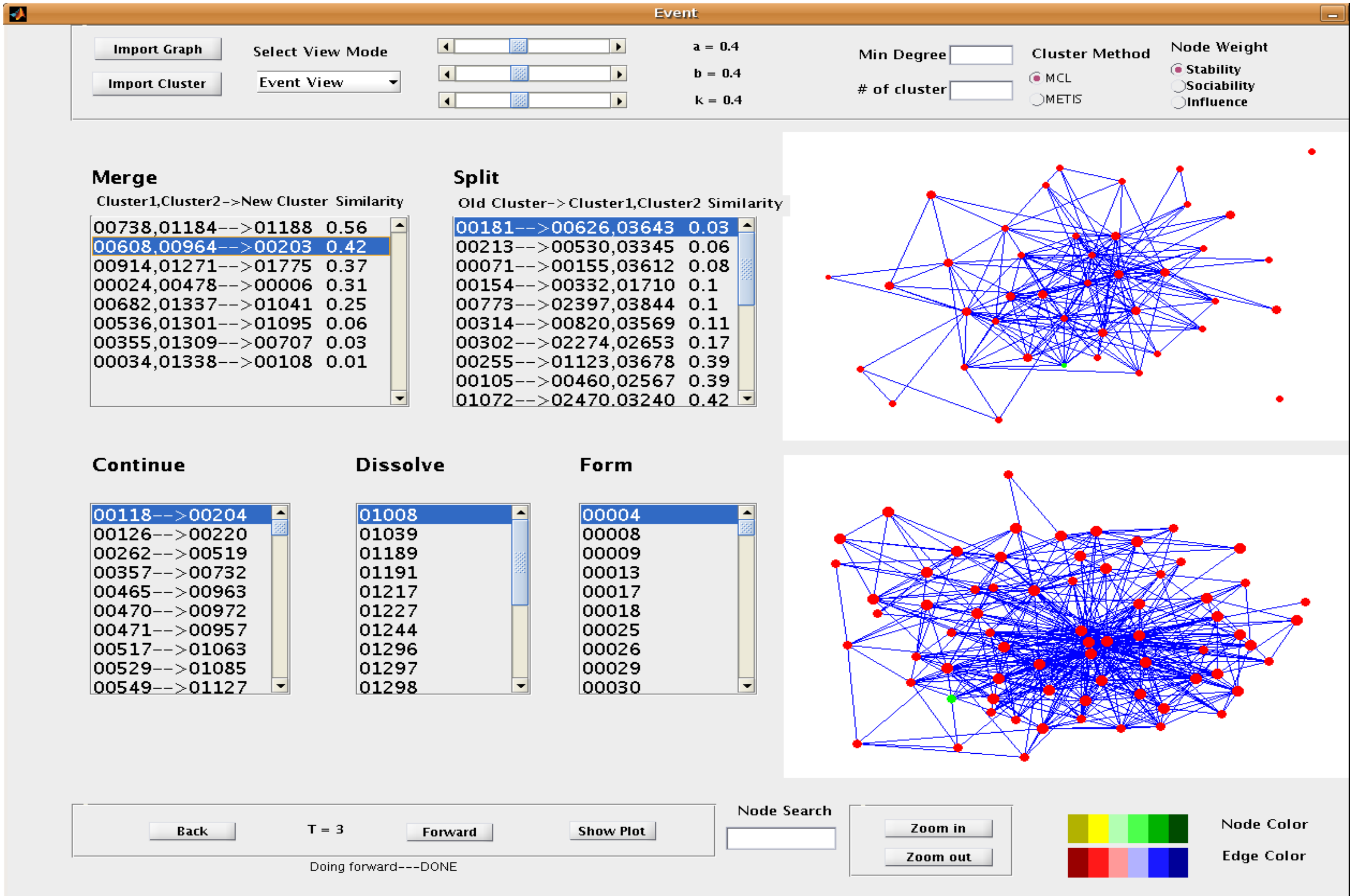
Visualization: Overview First (Coarsened View)



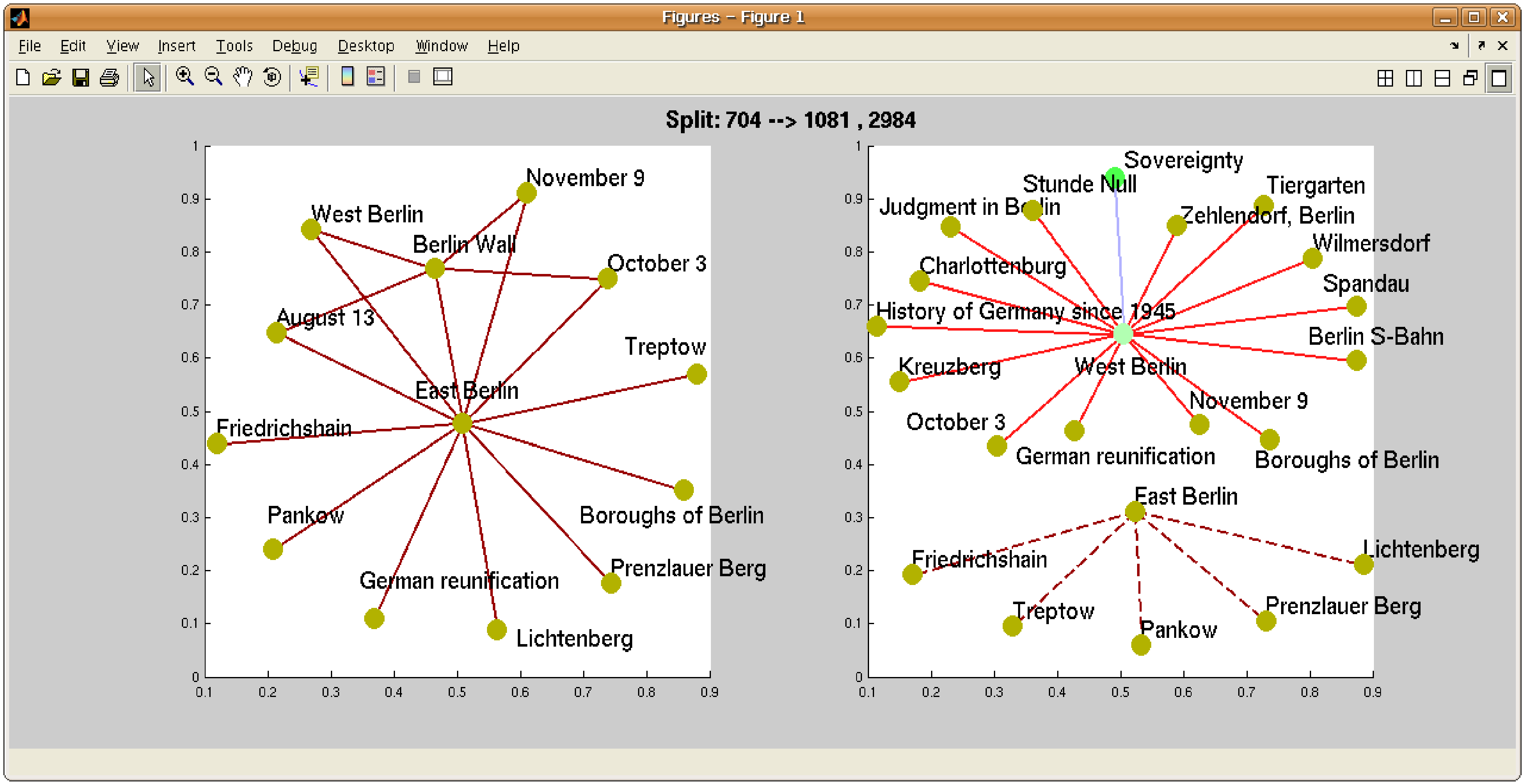
Zoom and Filter



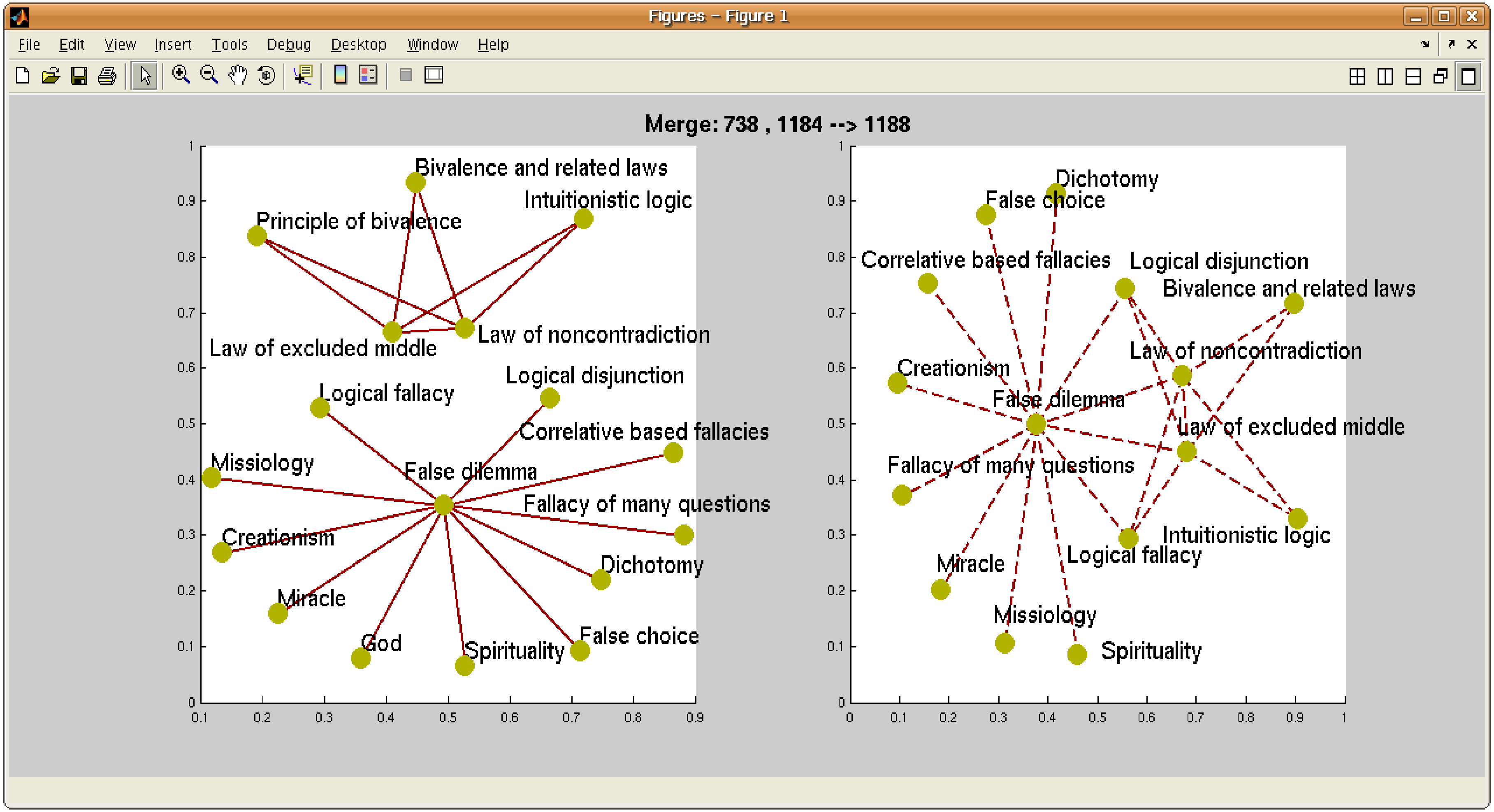
Event View (Importance of Ranking)



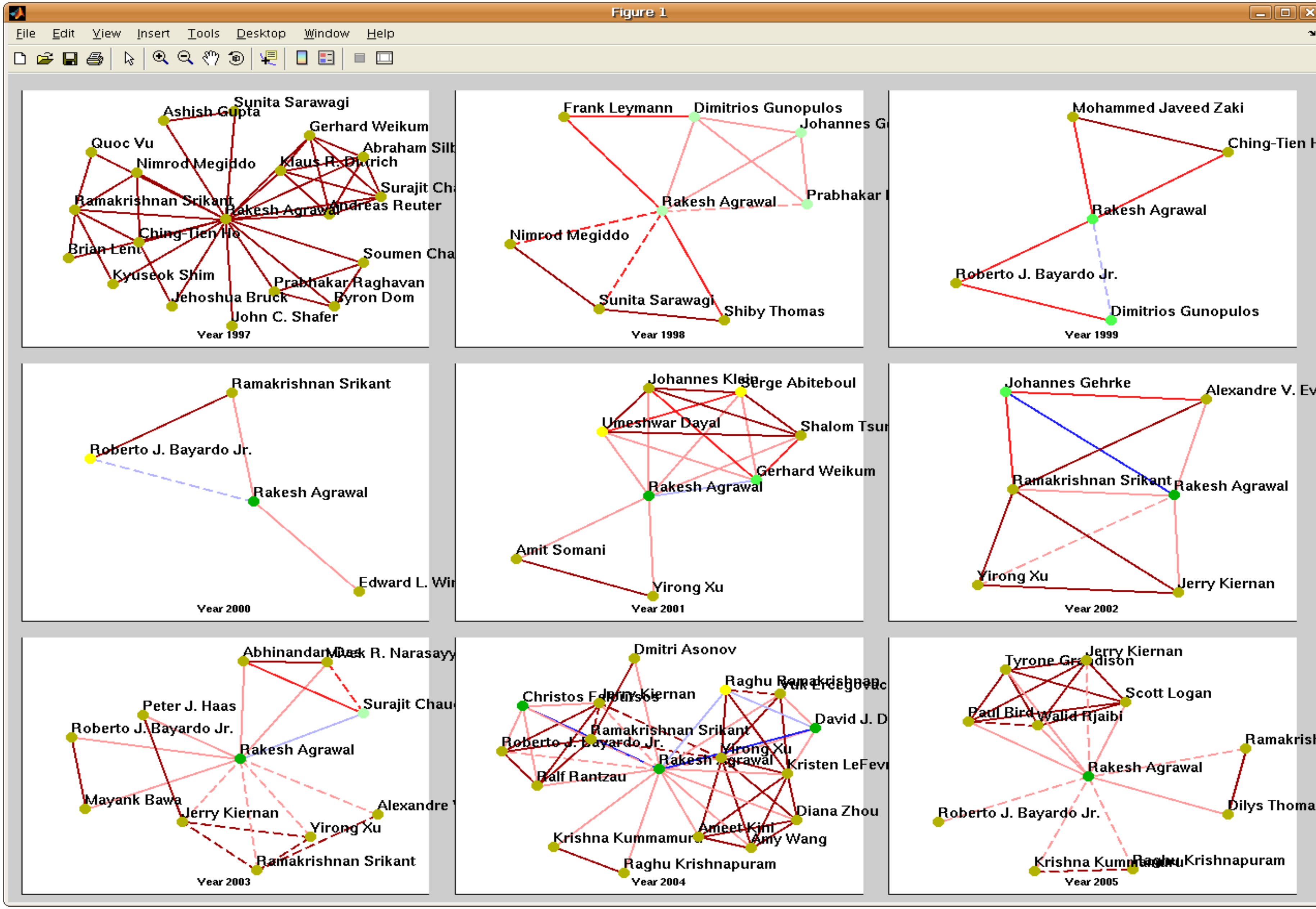
Split: Details on Demand (ironic example ☺)



Merge (Philosophy + Logic)

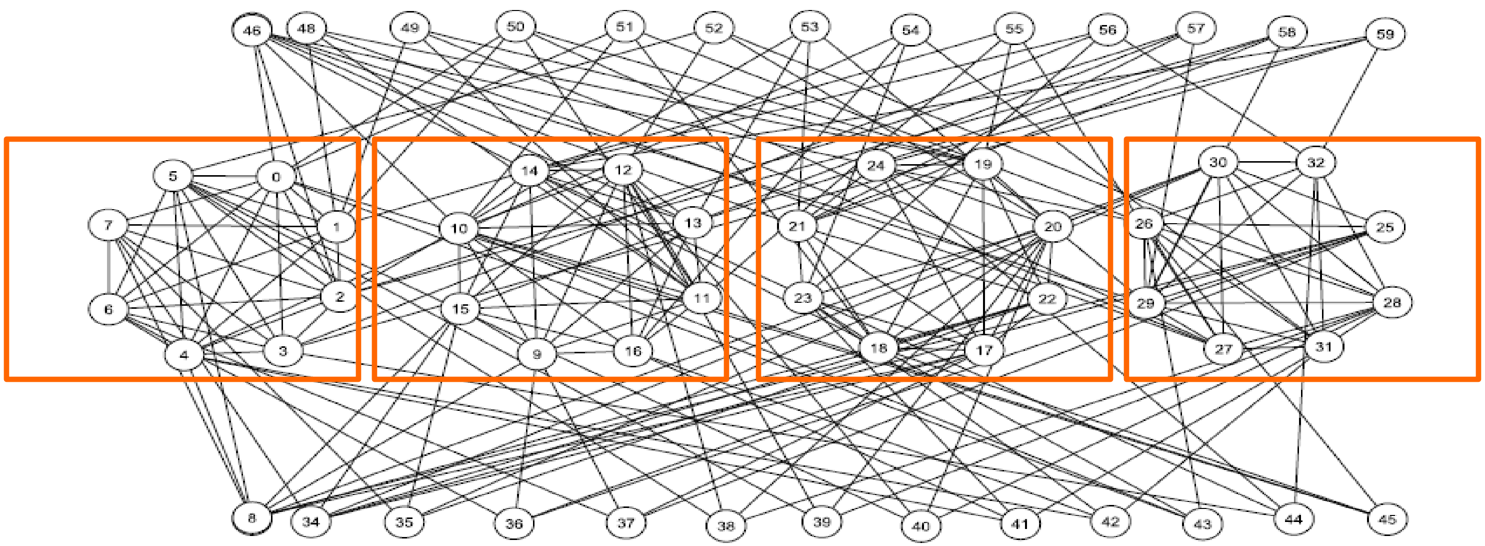


Dynamic Details (Sociability+ Influence)



Density (CSV) Plots

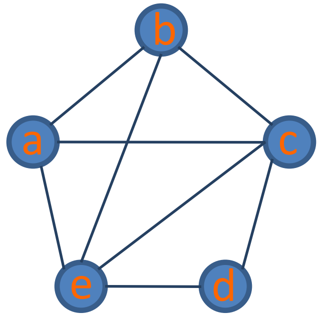
- Computing density plots efficiently was identified by SIGMOD keynote on Extreme Visualization as an important grand challenge problem
- Density Plots
  - Can help quickly localize dense subgraphs hidden within a large graph
  - The challenge is to compute them efficiently



Connectivity measurement

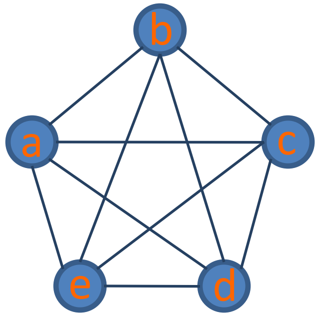
Connectivity measurement is closely related to clique (fully connected sub-graph) size.

The connectivity between two vertices in a graph (η_{max}) is defined to be the biggest clique in the graph such that both are members of the clique



η_{max}(a, d) = 0
η_{max}(a, c) = 4

The "connectivity" of a vertex (ζ_{max}) is similarly defined as the biggest clique it can participate.

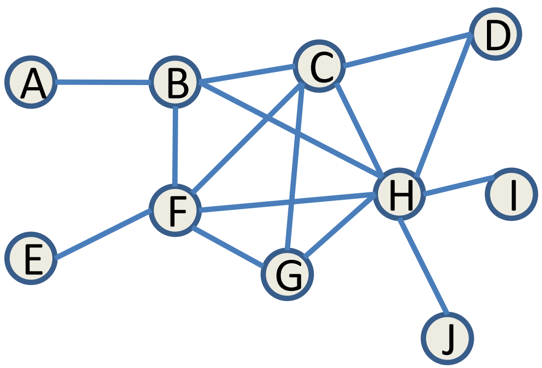


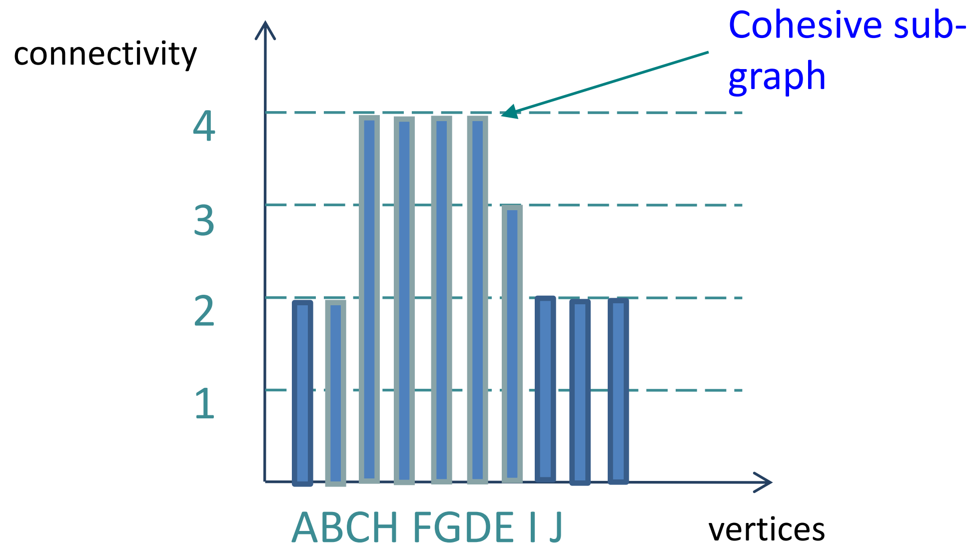
ζ_{max}(a) = 5

The algorithmic challenge is to approximate these efficiently [SIGMOD 2008]

CSV algorithm on a synthetic graph

From graph to plot

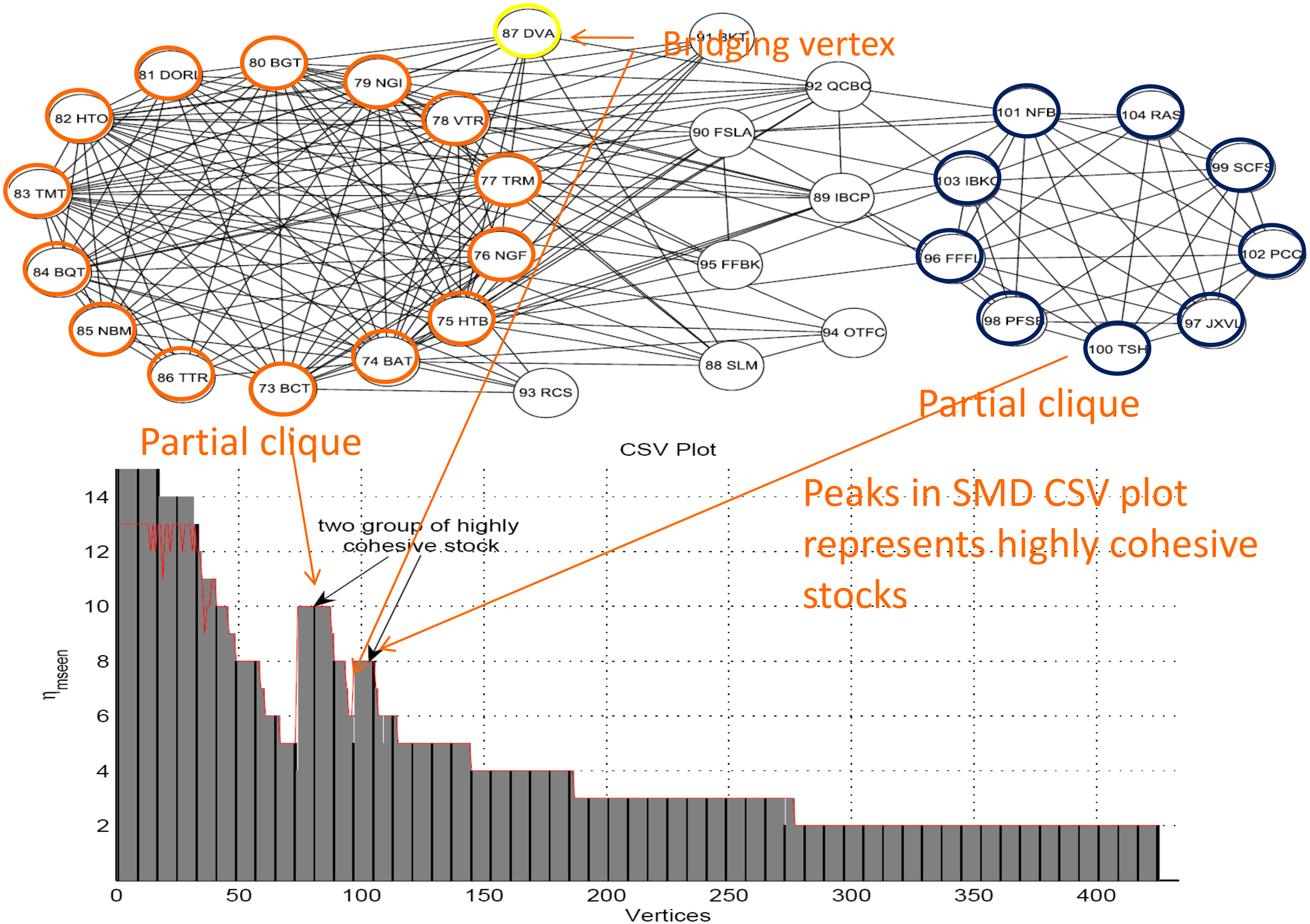




Visit every vertex accordingly to produce a plot.

Peaks represent cohesive sub-graphs.

SMD: Stock Market Data



DIP: Database of interacting proteins

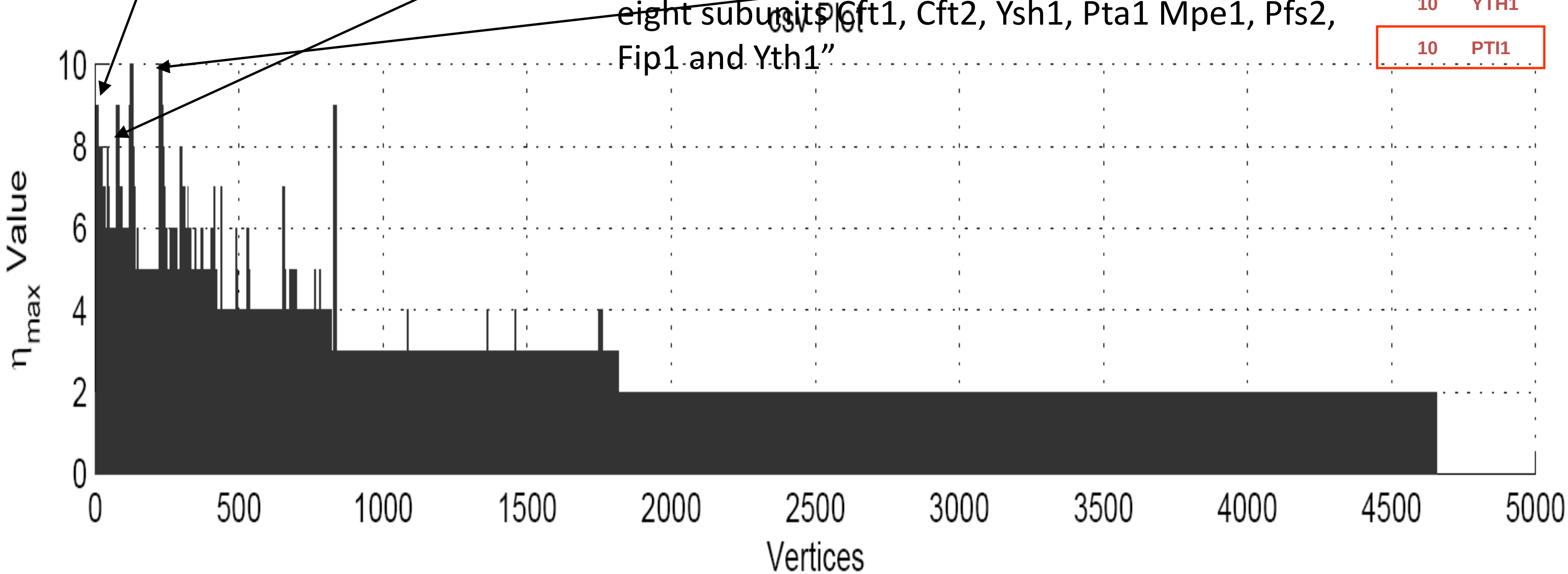
9 LSM8
9 LSM2
9 DCP1
9 LSM6
9 LSM3
9 LSM4
9 PAT1
9 LSM7
9 LSM5

8 SMD3
8 PRP4
8 PRP8
8 PRP6
8 LUC7
8 SMX2
8 SNP1
8 STO1
8 NAM8
8 SNU71
8 PRP31
8 YHC1

9 PFS2
10 RNA14
10 FIP1
10 REF2
10 CFT1
10 CFT2
10 MPE1
10 GLC7
10 PAP1
10 PTA1
10 YSH1
10 YTH1
10 PTI1

Structure of a nucleotide-bound Clp1-Pcf11 polyadenylation factor
Christian G. Noble, Barbara Beuth, and Ian A. Taylor*. Nucleic Acids Res. 2007 January; 35(1): 87–99.

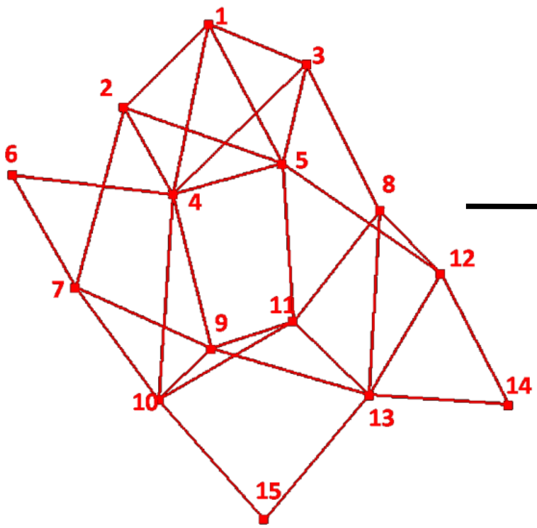
“CPF is also required in both the cleavage and polyadenylation reactions. It contains a core of eight subunits Cft1, Cft2, Ysh1, Pta1 Mpe1, Pfs2, Fip1 and Yth1”



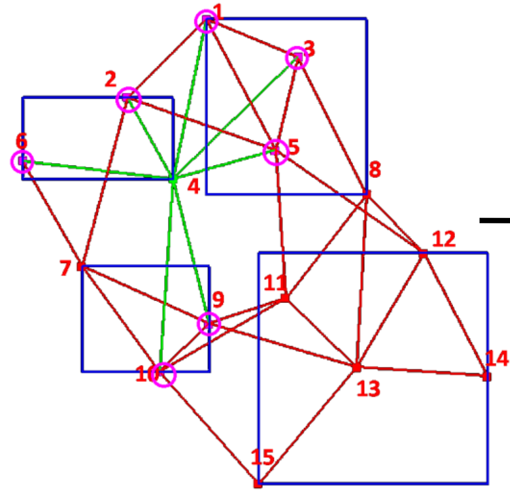
Handling Dynamism: Layout

- Surprisingly there are no good strategies here.
- Design tenets
  - Must maintain cognitive correspondence (mental map)
  - Must have similar “energy profile” to a stand-alone static approach
- Basic Dynamic Layout Strategy
  - Identify and localize changes to graph (e.g quad-tree/R-tree)
  - Compute dirty nodes/regions/bounding boxes
  - Ideally limit re-computation of layout to within bounding boxes that are dirty (guarantees mental map)
  - Produce final output

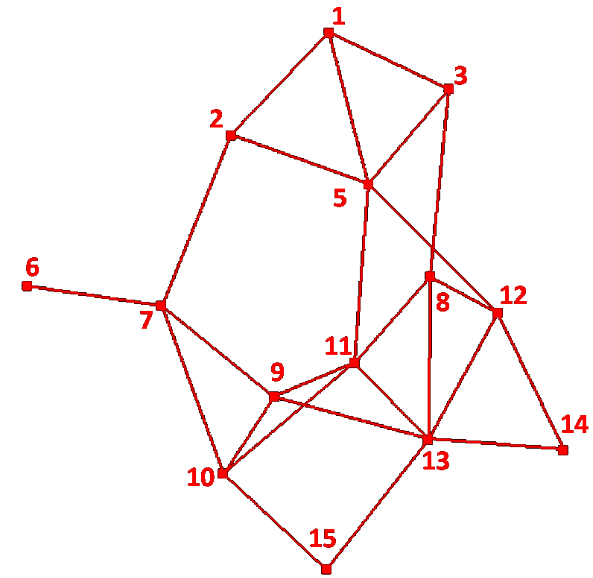
Dynamic Layout Strategy



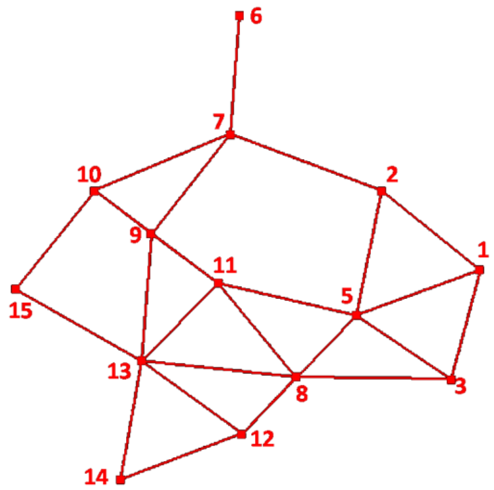
Original Graph G



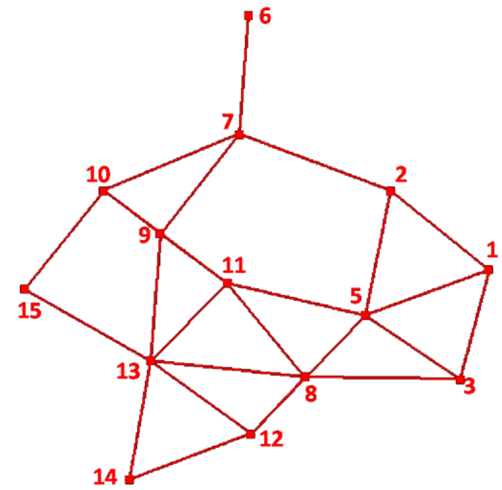
Delete Nd 4, Propogating Updates
Housing within an R-tree



Refined Graph G'

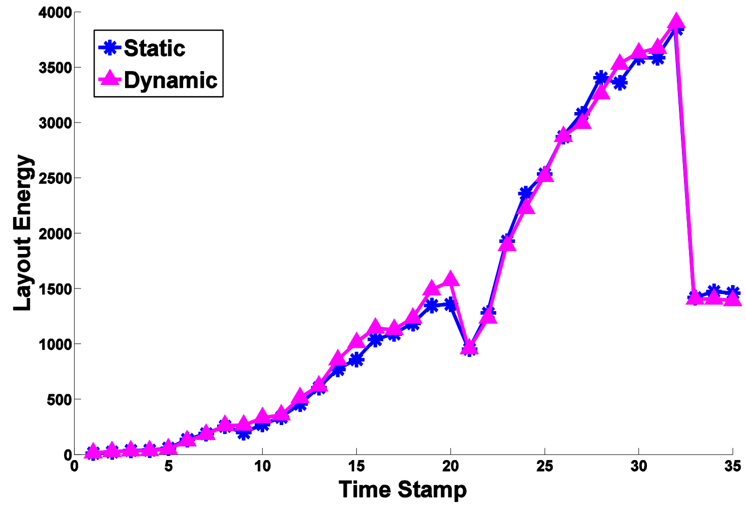


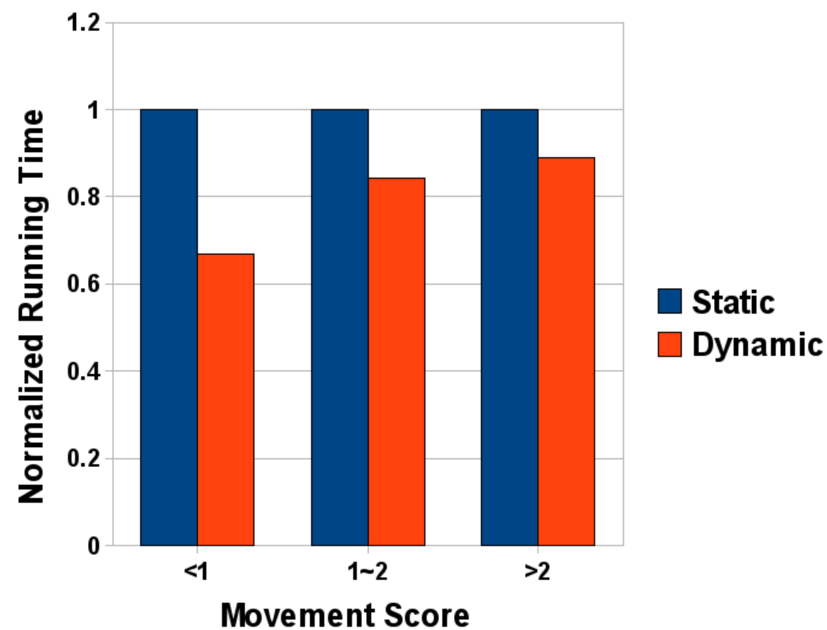
Static Layout of G'
for comparitive purposes



Dynamic Graph Layout: Early Results

- Enron Dataset
- Energy profile of Static (from scratch layout) very similar to our dynamic variant
- Dynamic variant maintains better mental map (not shown)
- Dynamic variant is also more efficient (up to 40% more efficient)





Concluding Remarks

- Visualization is an important facet of the knowledge discovery process
  - Transparency, validation, exploratory, data analysis are some of the roles
  - Central to discovery of actionable and interpretable patterns
- Potential for significant impact
  - Science, Engineering and Medicine
- Under represented in the field inspite of unquestioned utility
- Key challenges: pixel wall, scalability & integration

Exciting area to work in!

General thoughts on Interdisciplinary Collaboration

- Steep learning curve
  - Need to learn domain language
  - Express results in domain language
- Patience, patience, patience
  - Communities are inertia bound
  - Often difficult to make headway
- Potential for incredible rewards
  - Scientific/medical implications
- Good working relationship
  - Among collaborators is an absolute must – equal partners

Thanks for your attention
Questions?

- More details from:
  - srini@cse.ohio-state.edu
  - http://www.cse.ohio-state.edu/~srini
  - http://dmrl.cse.ohio-state.edu
- Most of these results can be found from the above sites
- Acknowledgements:
  - A number of NSF and DOE grants
  - A fantastic bunch of students and collaborators