Inference and Signal Processing for Networks

ALFRED O. HERO III
Depts. EECS, BME, Statistics
University of Michigan - Ann Arbor
http://www.eecs.umich.edu/~hero

Students: Clyde Shih, Jose Costa Neal Patwari, Derek Justice, David Barsic
Eric Cheung, Adam Pocholski, Panna Felsen

Outline

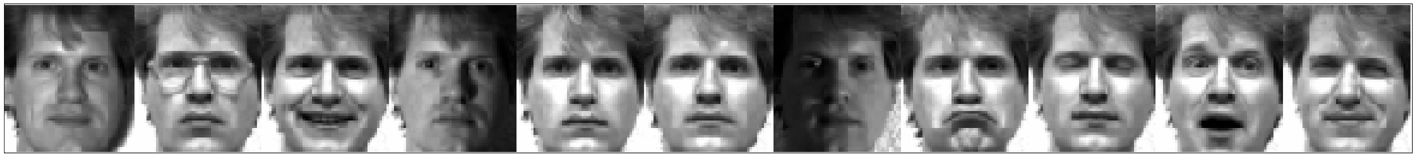
1. ***Dealing with the data cube***
2. ***Challenges in multi-site Internet data analysis***
3. ***Dimension reduction approaches***
4. ***Conclusion***

My Current Research Areas

- Dimension reduction, manifold learning and clustering
  - Information theoretic dimensionality reduction (Costa)
  - Information theoretic graph approaches to clustering and classification (Costa)
- Ad hoc networks
  - Distributed detection and node-localization in wireless sensor nets (Costa, Patwari)
  - Distributed optimization and distributed detection (Blatt, Patwari)
- Administered networks
  - Spatio-temporal Internet traffic analysis (Patwari)
  - Tomography (Shih)
  - Topology discovery (Shih, Justice)
- Adaptive resource allocation and scheduling in networks
  - Sensor management for tracking multiple targets (Kreucher)
  - Sensor management for acquiring smart targets (Blatt)
- Inference on gene regulation networks
  - Gene and gene pair filtering and ranking (Jing, Fleury)
  - Confident discovery of dependency networks (Zhu)
- Imaging
  - Image and volume registration (Neemuchwala)
  - Tomographic reconstruction from projections in medical imaging (Fessler)
  - Quantum imaging, computational microscopy and MRFM (Ting)
  - Multi-static radar imaging with adaptive waveform diversity (Raich, Rangajaran)

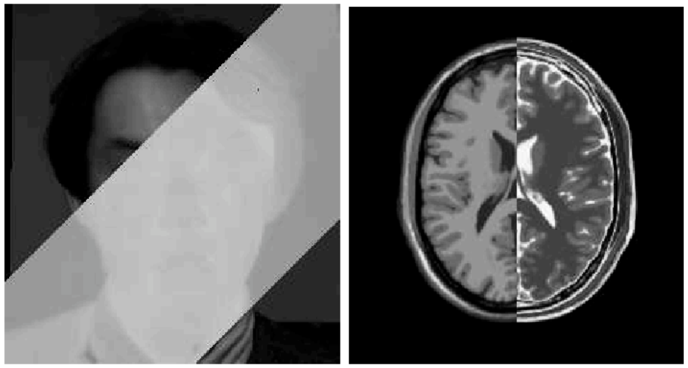
Applications

- Characterization of face manifolds (Costa)

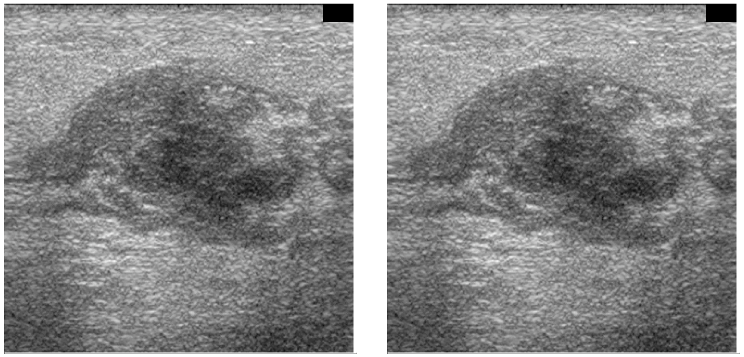


  - The set of face images evolve on a lower dimensional imbedded manifold in 128x128 =16384 dimensions

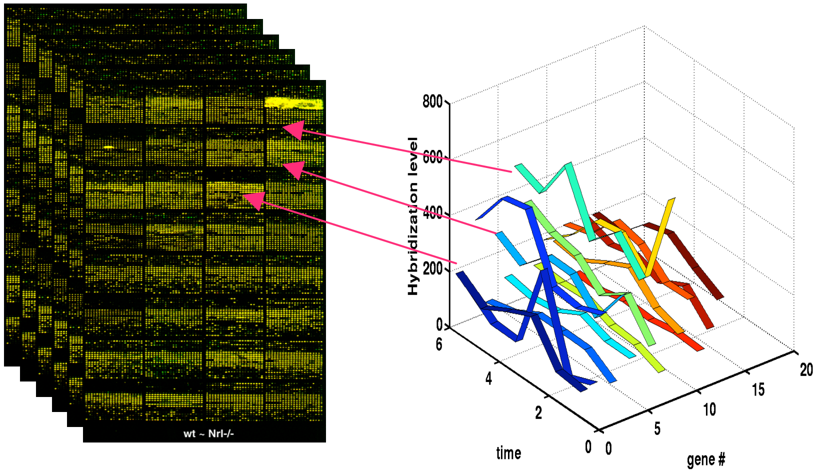
- Handwriting (Costa)   - Pattern Matching(Neemuchwala)



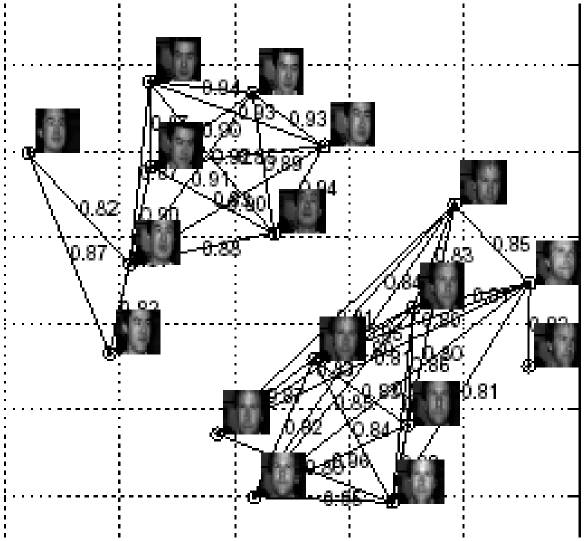
Applications



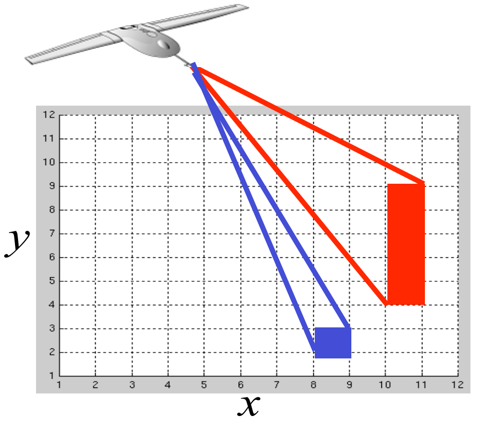
Ultrasound Breast Registration (Neemuchwala)



Gene microarray analysis (Zhu)



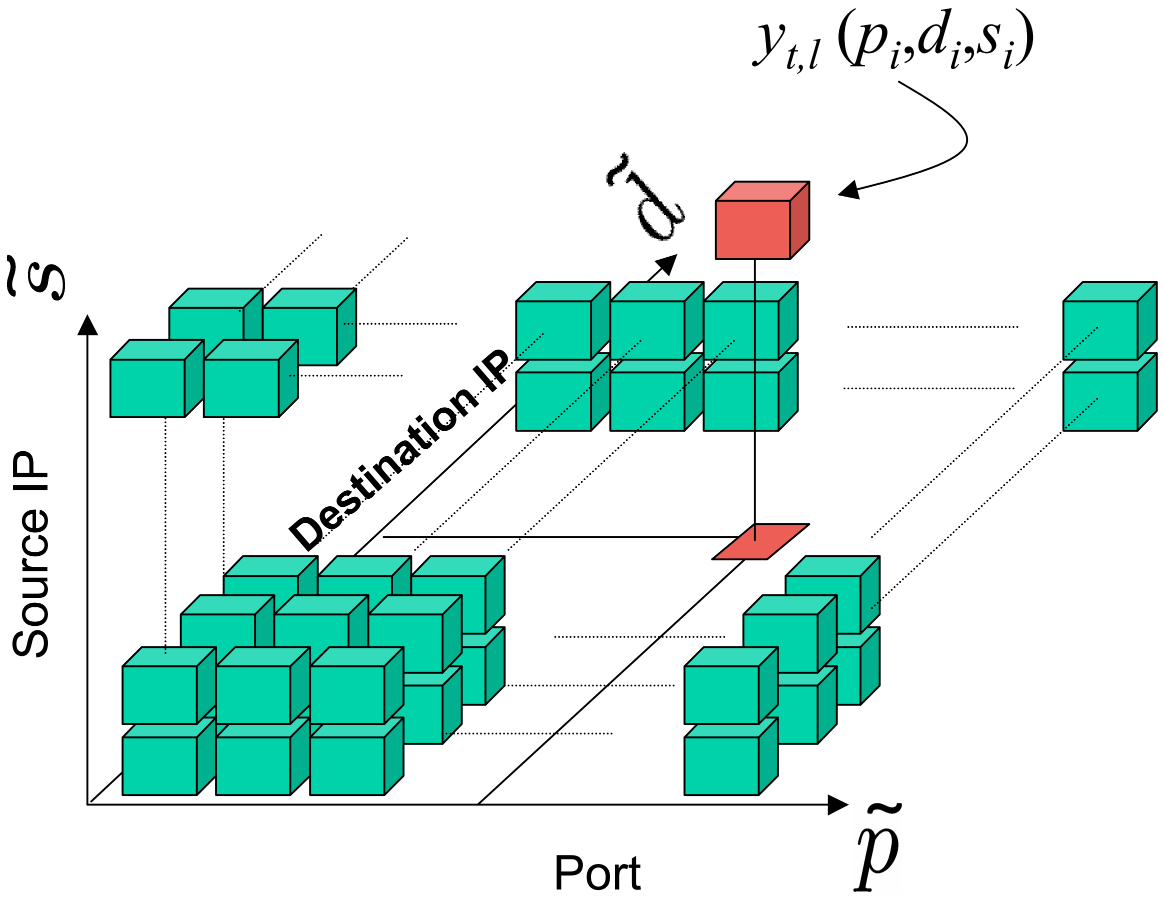
Clustering and classification (Costa)



$$\mathbf{x} = [x, y, \dot{x}, \dot{y}]^T$$

Adaptive scheduling of measurements (Kreucher)

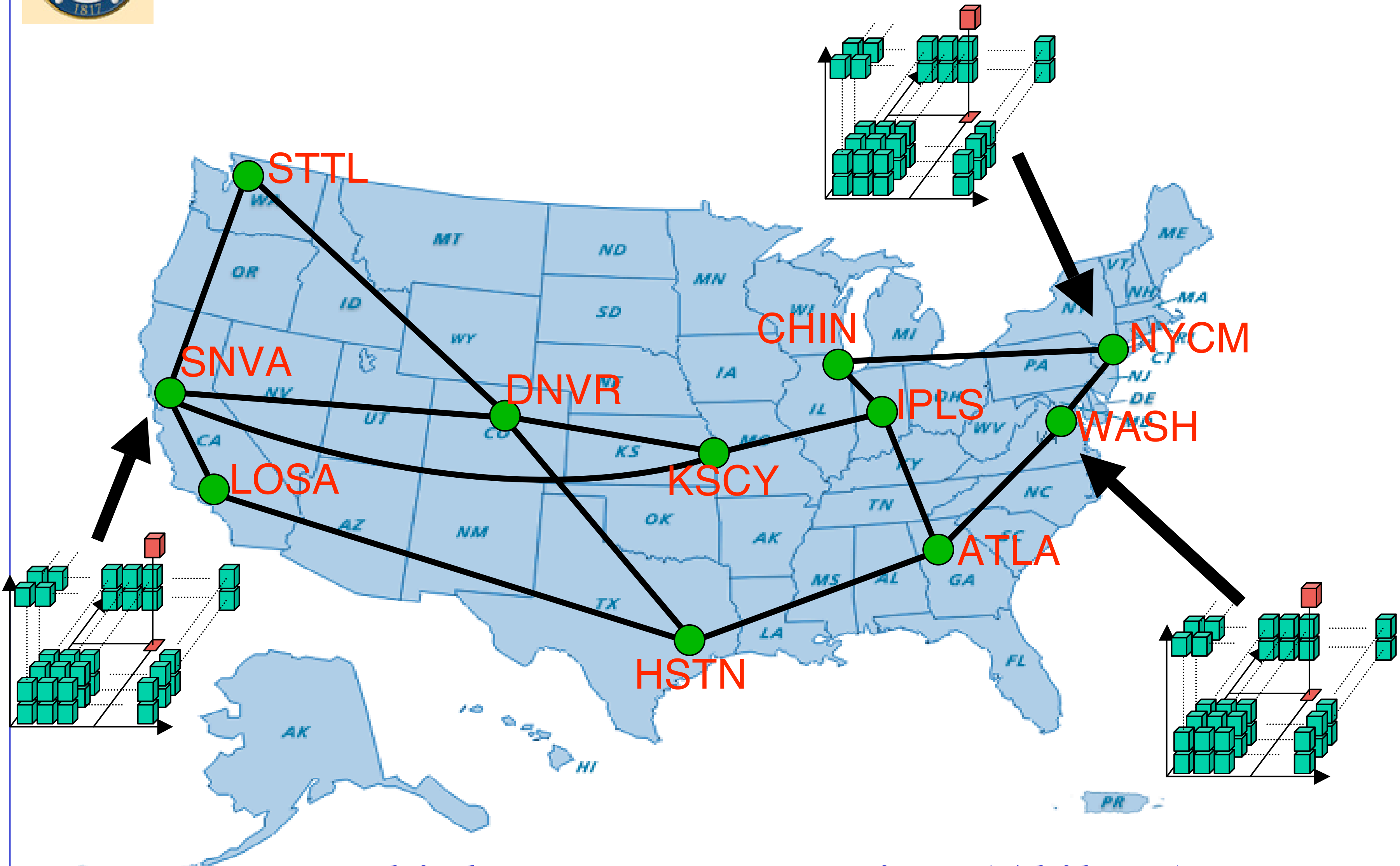
1. Dealing with the data cube



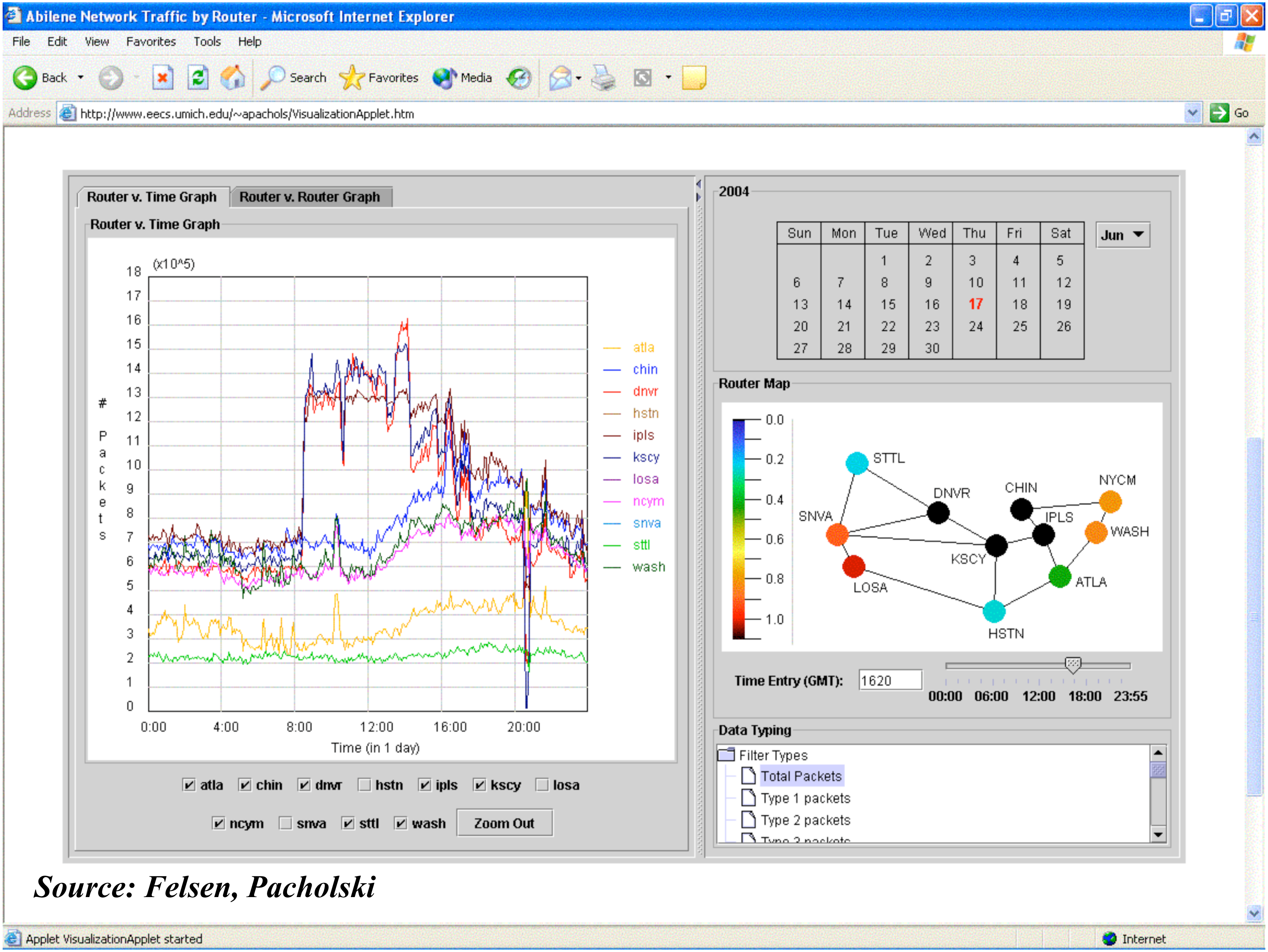
Single measurement site (router)

Ports, applications, protocols > dozens of dimensions

Dealing with the data cube



Multiple measurement sites (Abilene)



Abilene Network Traffic by Router - Microsoft Internet Explorer

Address: http://www.eecs.umich.edu/~apachols/VisualizationApplet.htm

Router v. Time Graph | Router v. Router Graph

Router v. Time Graph

☑ atla ☑ chin ☑ dnvr □ hstn ☑ ipls ☑ kscy □ losa

☑ ncym □ snva ☑ sttl ☑ wash Zoom Out

2004

Sun	Mon	Tue	Wed	Thu	Fri	Sat
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	**17**	18	19
20	21	22	23	24	25	26
27	28	29	30			

Jun

Router Map

Time Entry (GMT): 1620

Data Typing

- Filter Types
  - Total Packets
  - Type 1 packets
  - Type 2 packets
  - Type 3 packets

Applet VisualizationApplet started

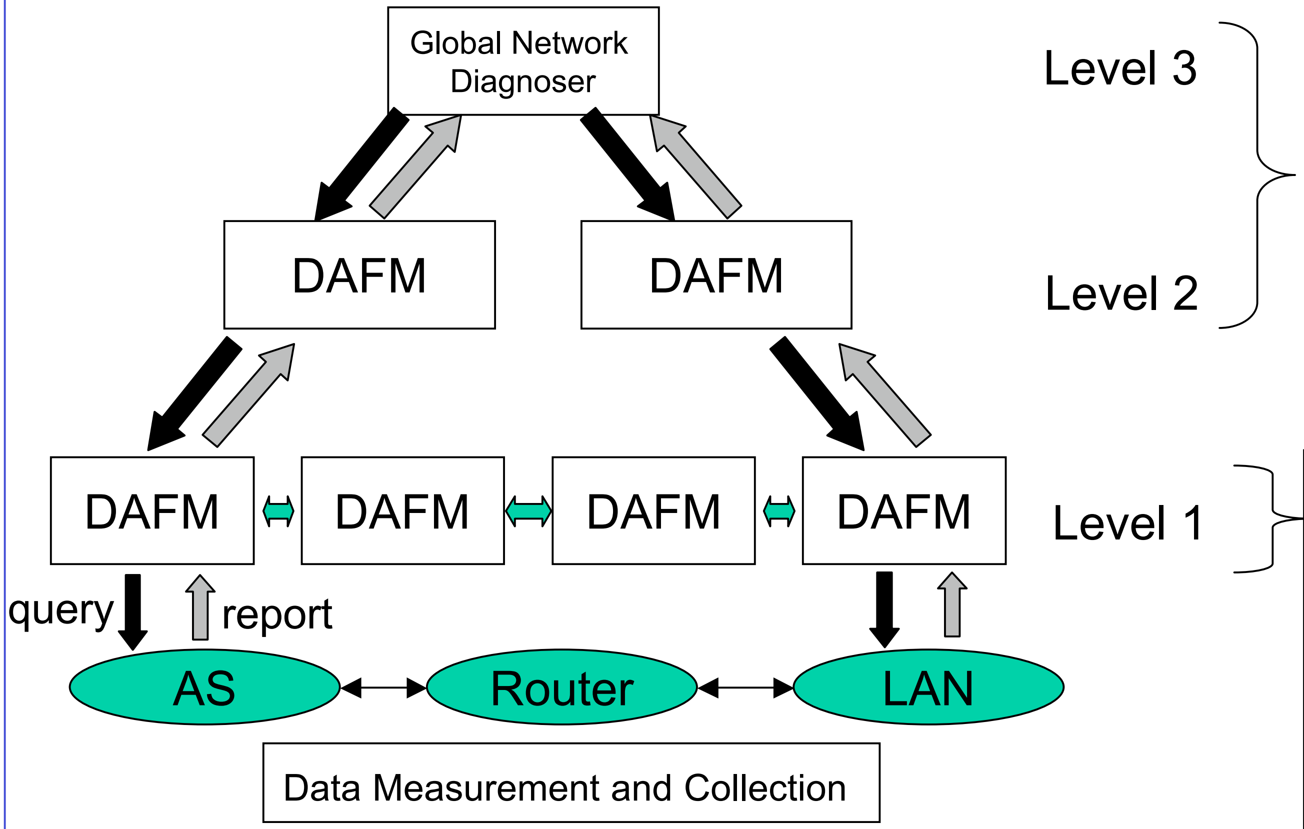
Internet

Source: Felsen, Pacholski

2. Internet SP Challenges

- What makes multisite Internet data analysis hard from a SP point of view?
  - Bandwidth is always limited
  - Sampling will never be adequate
    - Spatial sampling: cannot measure all link/node correlations from passive measurements at only a few sites
    - Temporal sampling: full bit stream cannot be captured
    - Category sampling: only a subset of all field variables can be monitored at a time
  - Measurement data is inherently non-stationary
  - Standard modeling approaches are difficult or inapplicable for such massive data sets
  - Little ground truth data is available to validate models
- General robust and principled approach is needed:
  - Adopt hierarchical multiresolution modeling and analysis framework
  - Task-driven dimension reduction

Hierarchical Network Measurement Framework



Event-driven models

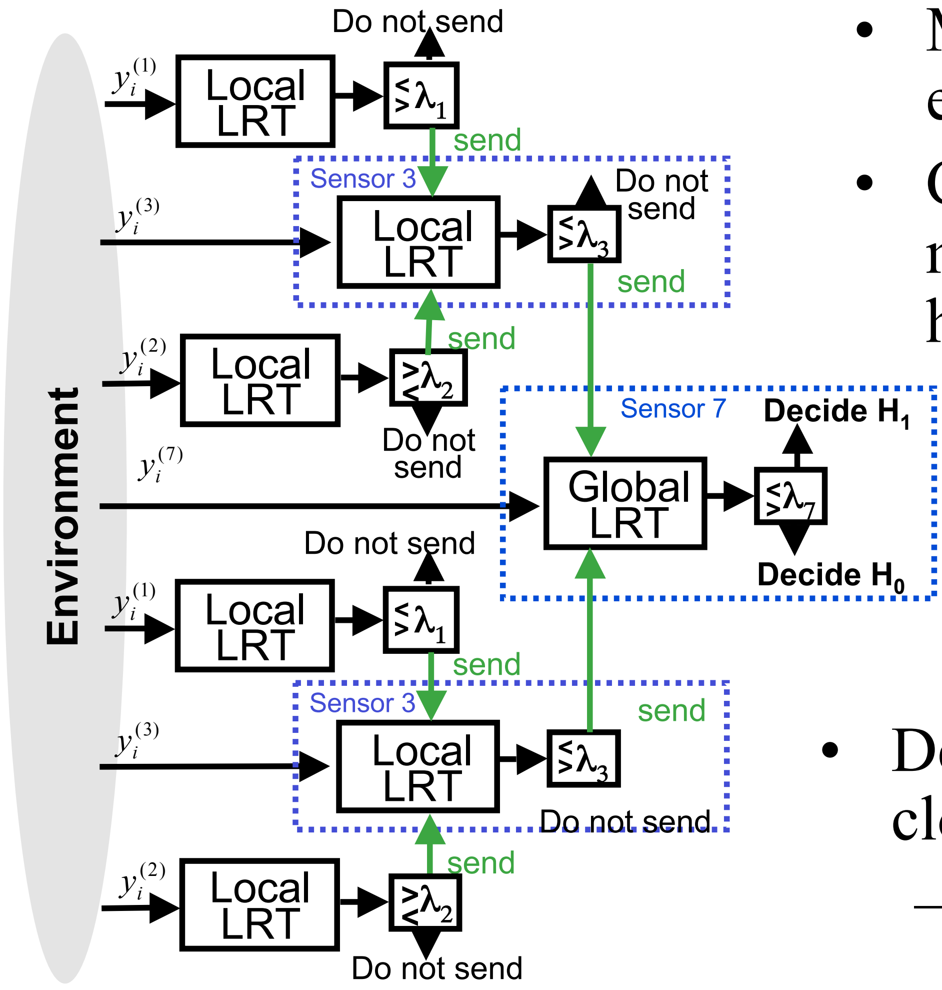
- Modular diagnosis
- Active querying
- Distributed detection

Spatio-temporal models and systems

- Feature extraction
- Dimension reduction
- Tomography
- On-line traffic analysis

Legend: DAFM - Data aggregation and filtering module
AS – Autonomous System
LAN – Local Area Network

Example: distributed anomaly detection

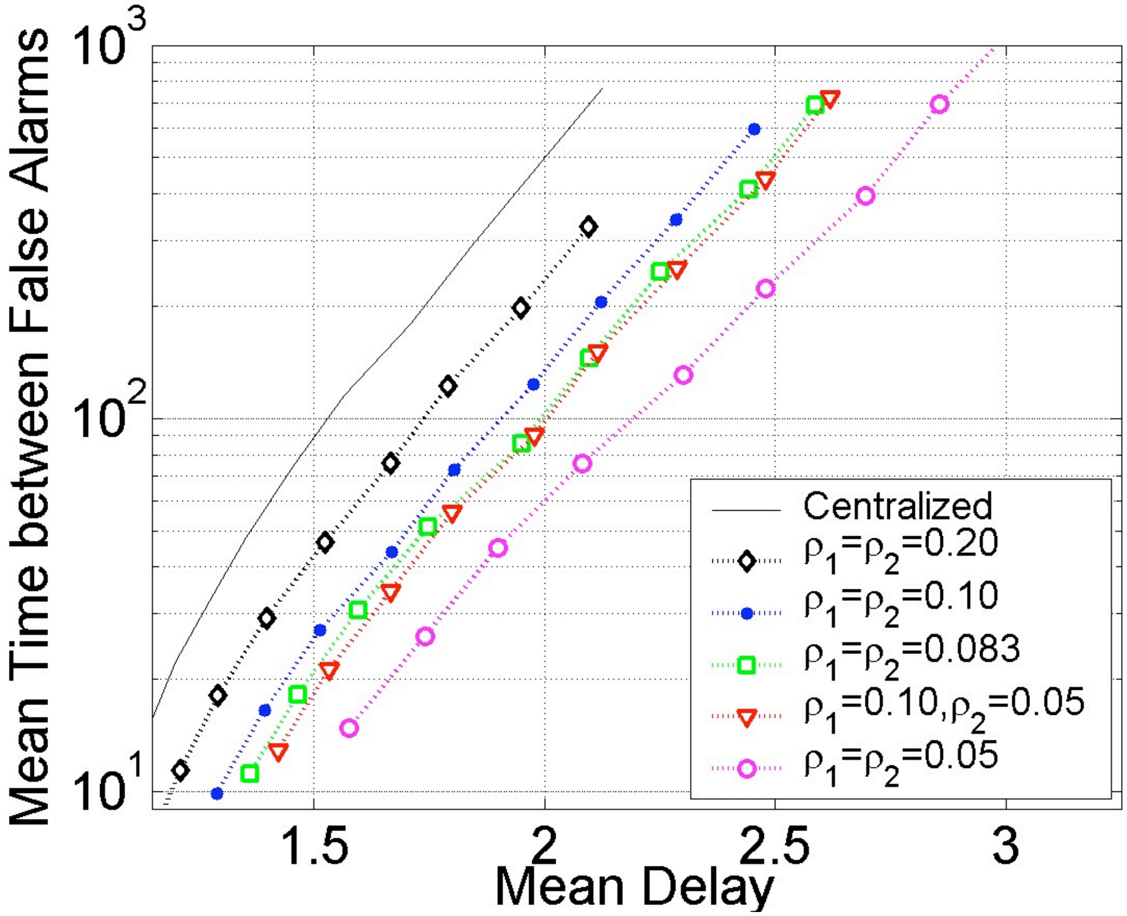


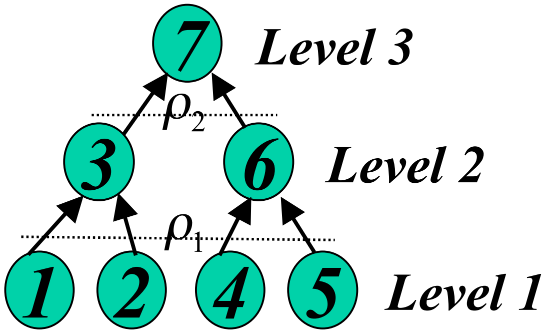
- Multi-hop is desirable for energy efficiency, cost
- Censored test can be iterated to match arbitrary multi-hop 'tree' hierarchy
  - ∀ $\rho = 1$ ↔ centralized
  - $0 < \rho < 1$ ↔ data fusion, reduce data bottleneck at the root
- Detection performance can be close to optimal [1]
  - Even $\rho = 0.01$ sensors greatly improve performance

[1] N. Patwari, A.O. Hero III, "Hierarchical Censoring for Distributed Detection in Wireless Sensor Networks", IEEE ICASSP '03, April 2003.

Example: distributed anomaly detection

- Parameter ρ selected to constrain mean time btwn false alarms



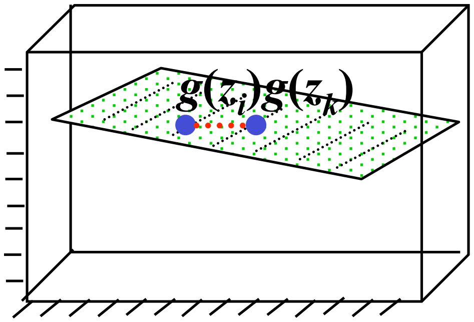


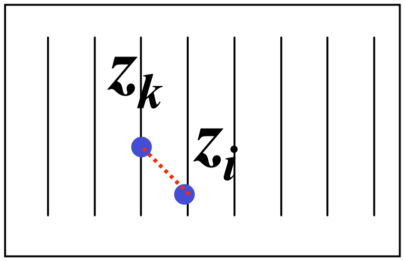
Research Issues

- Broad questions
  - Anomaly detection, classification, and localization
    - Model-driven vs data-driven approaches
    - Partitioning of information and decisionmaking (Multiscale-multiresolution decision trees)
    - Learning the “Baseline” and detecting deviations
    - Feature selection, updating, and validation
  - Multi-site measurement and aggregation
    - Remote monitoring: tomography and topology discovery
    - Multi-site spatio-temporal correlation
    - Distributed optimization/computation
  - Dynamic spatio-temporal measurement
    - Sensor management: scheduling measurements and communication
    - Passive sensing vs. active probing
    - Adaptive spatio-temporal resolution control
  - Dimension reduction methods
    - Beyond linear PCA/ICA/MDS…

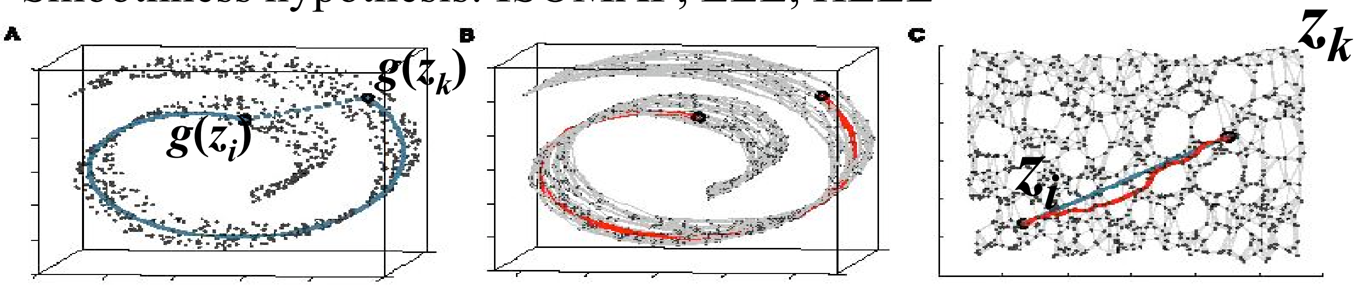
3. Dimension Reduction

- Manifold domain reconstruction from samples: “the data manifold”
  - Linearity hypothesis: PCA, ICA, multidimensional scaling (MDS)





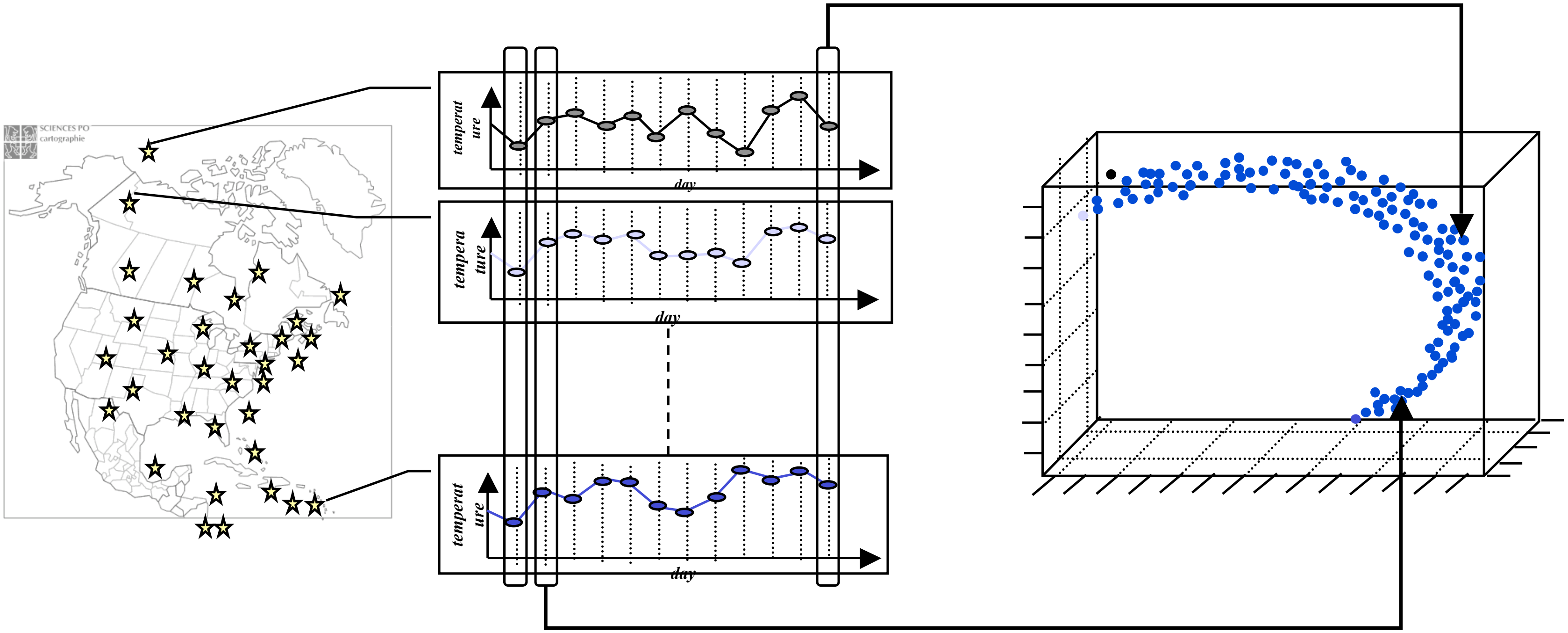
  - Smoothness hypothesis: ISOMAP, LLE, HLLE



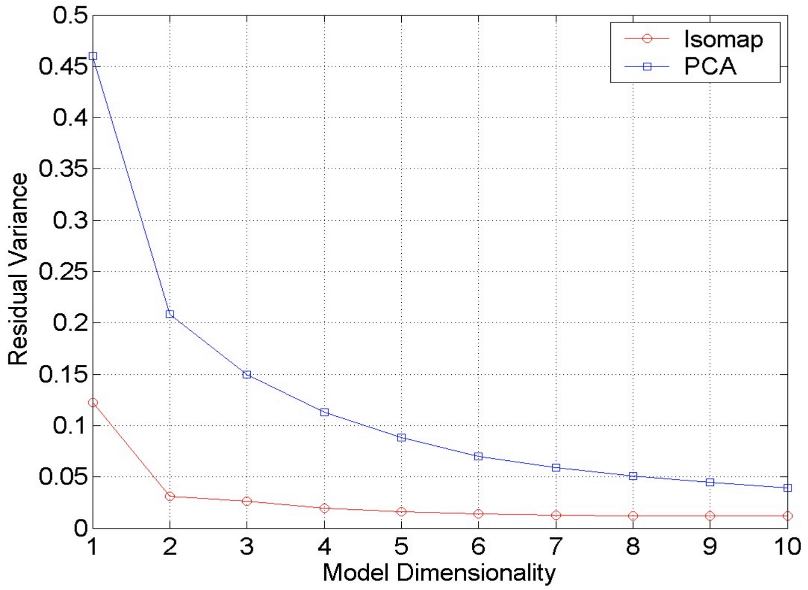
- Dimension estimation: infer degrees of freedom of data manifold
- Infer entropy, relative entropy of sampling distribution on manifold

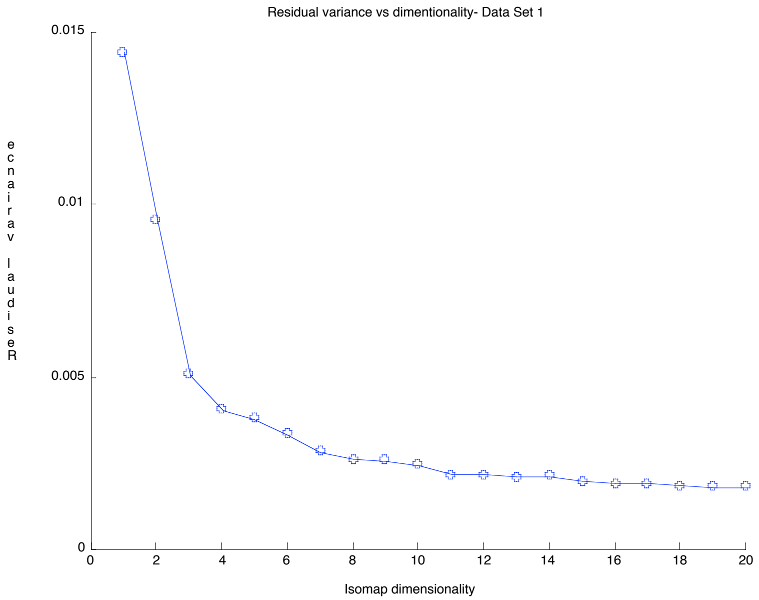
Application: Internet Traffic Visualization

- Spatio-temporal measurement vector:



Key problem: dimension estimation

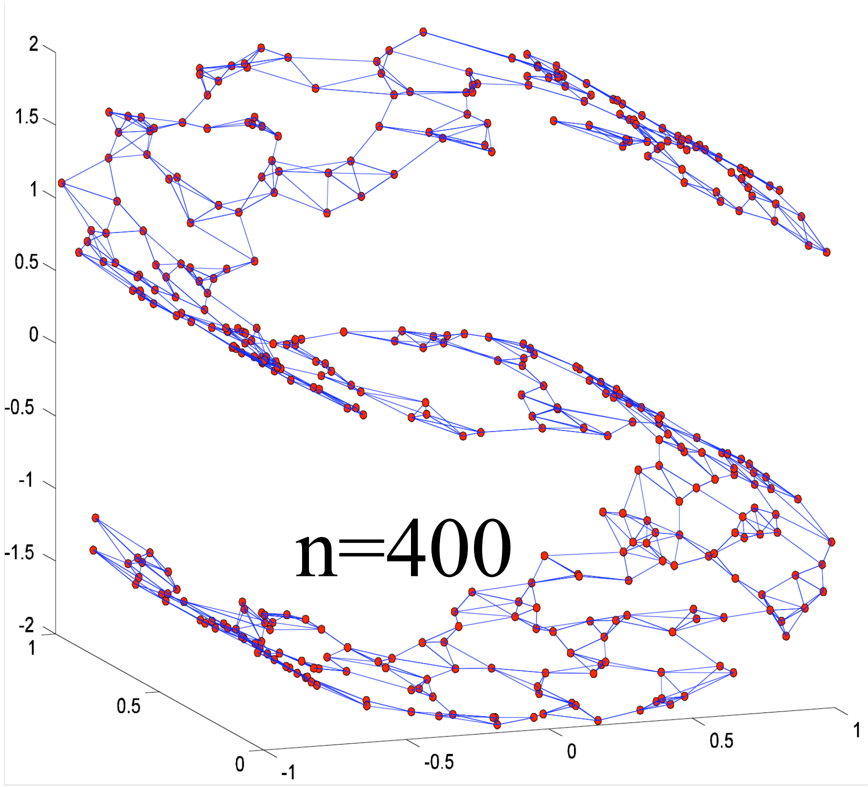


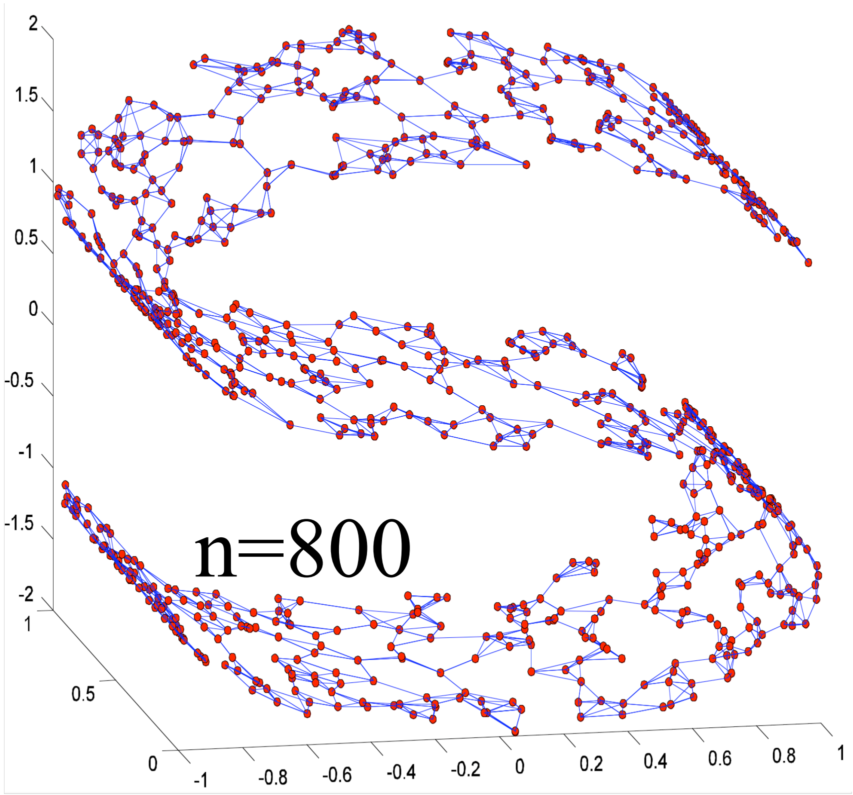


Residual fitting curves
for 11x21 = 231 dimensional
Abilene Netflow data set

ISOMAP residual curve
for 41+11=51 dimensional
Abilene OD link data
(Lakhina,Crovella, Diot)

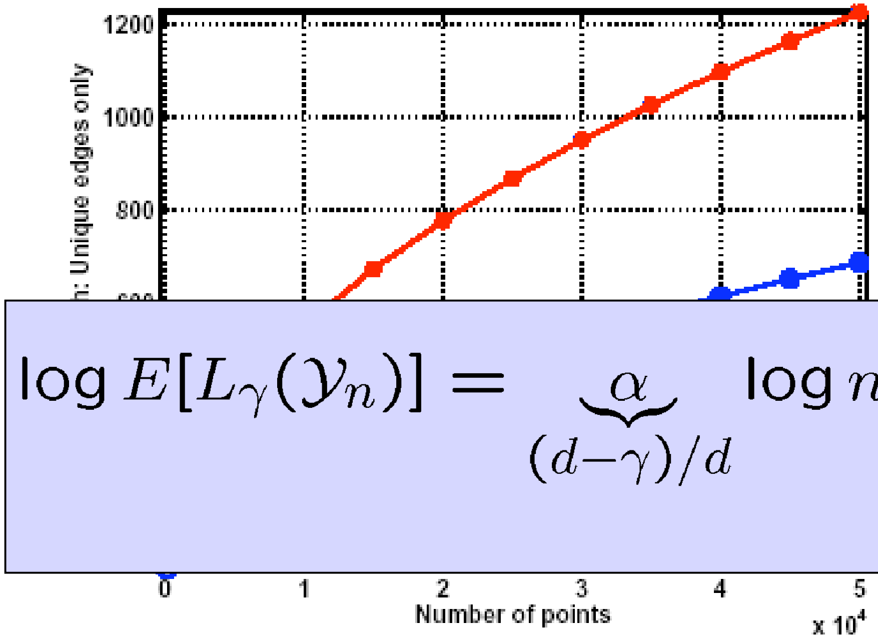
GMST Rate of convergence=dimension, entropy

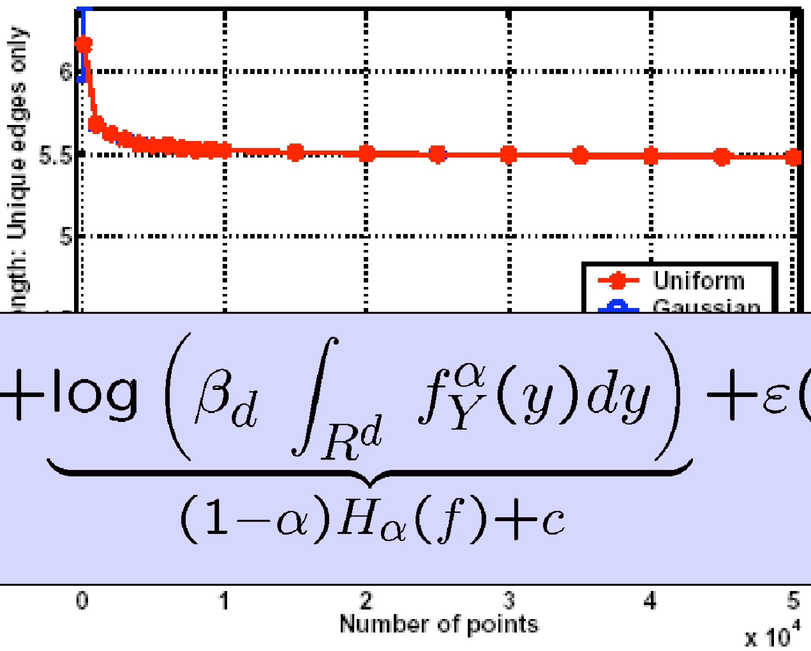




Rate of increase in length functional of MST should be related to the intrinsic dimension of data manifold

BHH Theorem



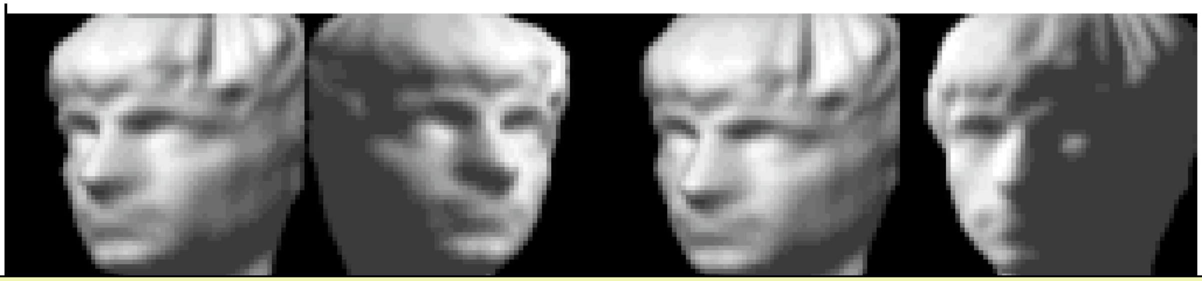


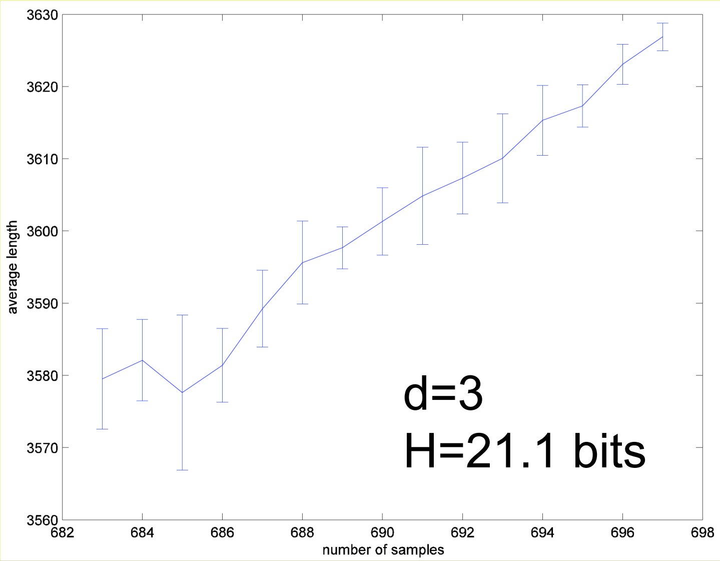
$$\log E[L_\gamma(\mathcal{Y}_n)] = \underbrace{\alpha}_{(d-\gamma)/d} \log n + \underbrace{\log\left(\beta_d \int_{R^d} f_Y^\alpha(y)dy\right)}_{(1-\alpha)H_\alpha(f)+c} + \varepsilon(n)$$

Extended BHH Theorem (Costa&Hero):

$$L_\gamma(\mathcal{Y}_n)/n^\alpha \to \beta_d \underbrace{\int_S f_Y^\alpha(y)dy}_{H_\alpha(f_Y)} \qquad \alpha = (d-\gamma)/d$$

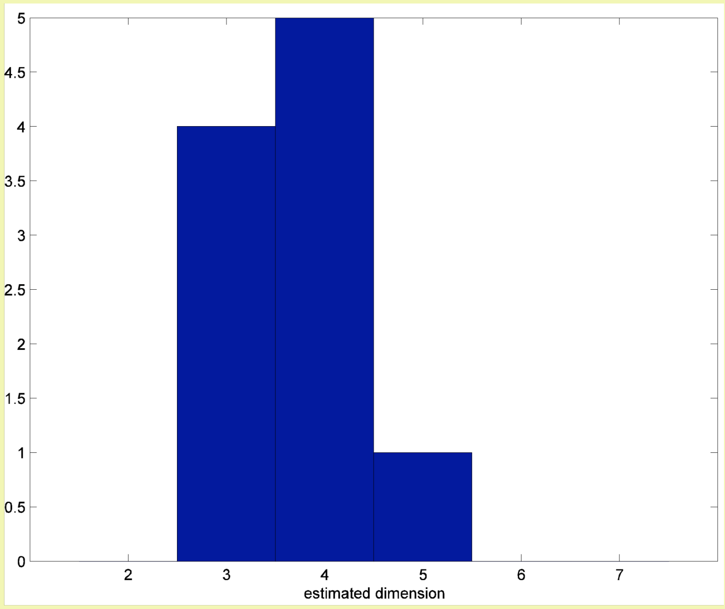
Application: ISOMAP Database





d=3
H=21.1 bits

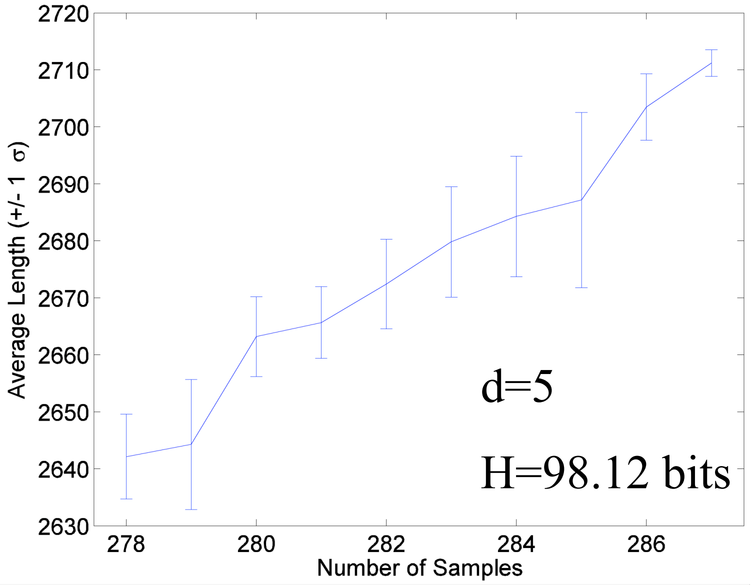
Mean GMST Length Function



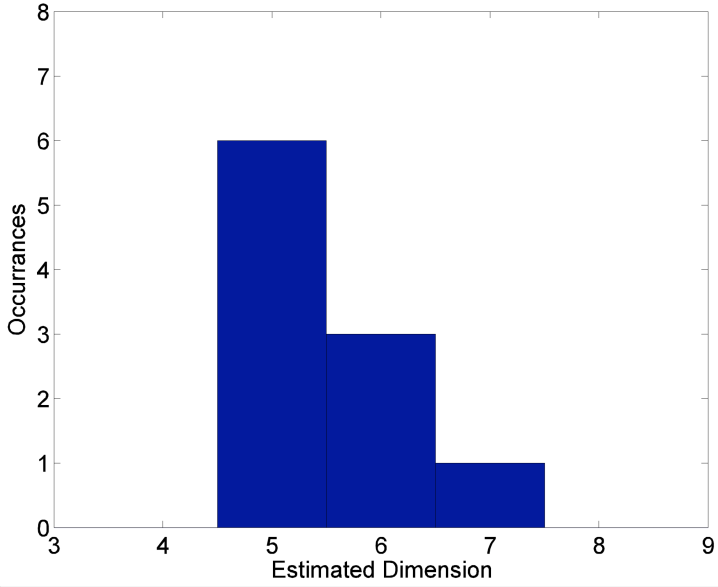
Resampling Histogram of d hat

Illustration: Abilene Netflow

- 11 routers and 21 applications = each sample lives in 231 dimensions
- 24 hour data block divided into 5 min intervals = 288 samples

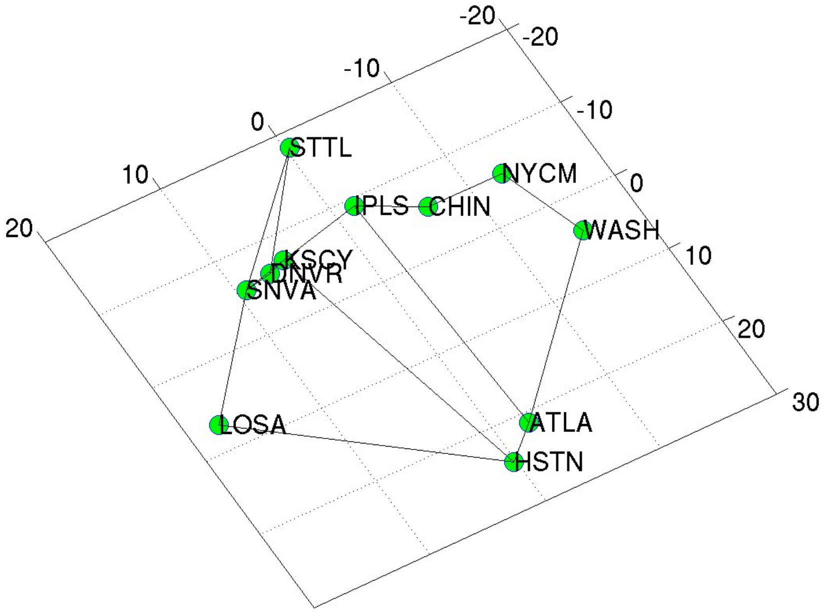


Mean GMST Length Function

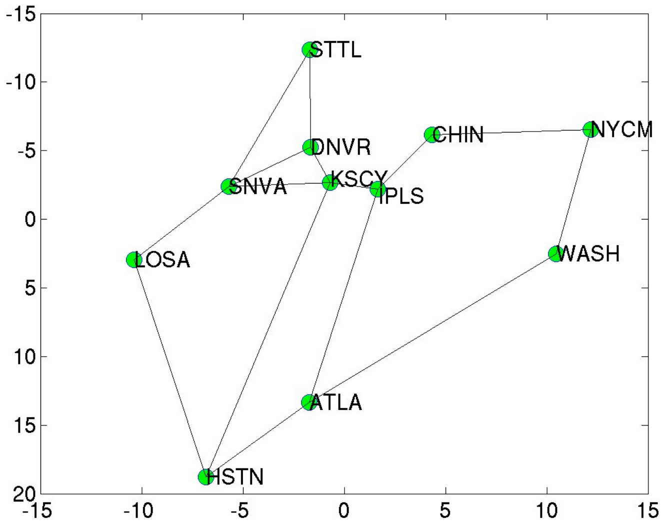


Resampling histogram of d hat

dwMDS embedding/visualization



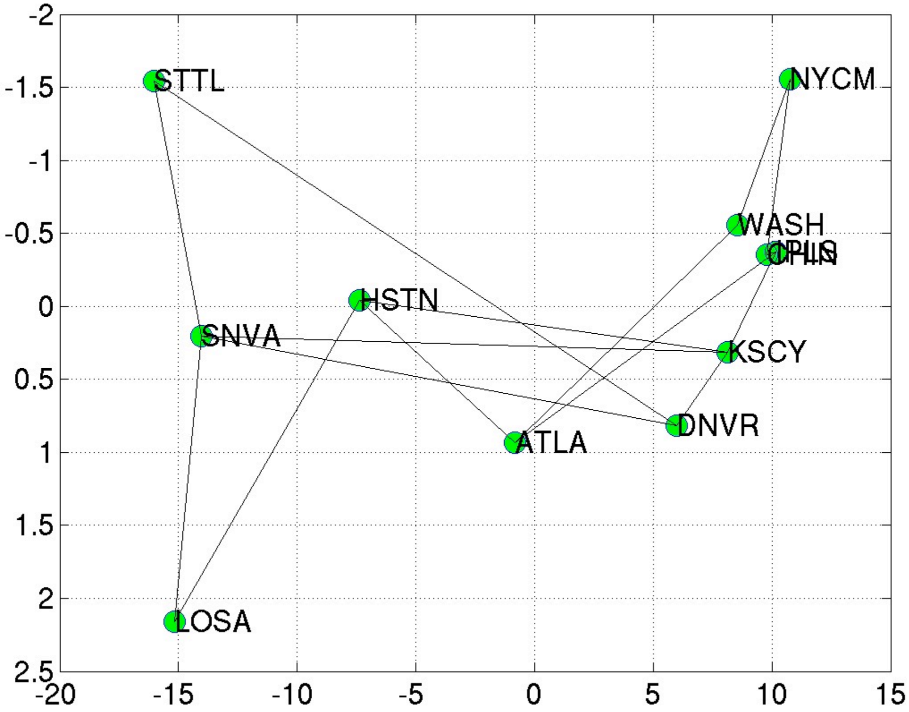
Abilene Network Isomap
(Centralized computation)



Abilene Network DW MDS
(Distributed computation)

Data: total packet flow over 5 minute intervals 10 june ’04
Isomap(Tennbaum): k=3, 2D projection, L2 distances
DW MDS(Costa&Patwari&Hero): k=5, 2D projection, L2 distances

dwMDS embedding/visualization



Abilene Network MDS (linear)
(Centralized computation)

Data: total packet flow over 5 minute intervals 10 june ’04
MDS: 2D projection, L2 distances

4. Conclusions

- Interface of SP, control, info theory, statistics and applied math is fertile ground for network measurement/data analysis
- SP will benefit from scalable hierarchical multiresolution modeling and analysis framework
  - Multiresolution modeling, communication, decisionmaking
- Task-driven dimension reduction is necessary
  - Go beyond linear methods (PCA/ICA)
    - What is goal? Estimation/Detection/Classification?
    - Subspace constraints (smoothness, anchors)?
    - Out-of-sample updates?
    - Mixed dimensions?
- Validation is a critical problem: annotated classified data or ground truth data is lacking.