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 $128 \times 128 = 16384$ dimensions

- Handwriting (Costa)
- Pattern Matching (Neemuchwala)
Applications

Ultrasound Breast Registration (Neemuchwala)

Gene microarray analysis (Zhu)

Clustering and classification (Costa)

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)
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

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

Event-driven models
- Modular diagnosis
- Active querying
- Distributed detection

Spatio-temporal models and systems
- Feature extraction
- Dimension reduction
- Tomography
- On-line traffic analysis
Example: distributed anomaly detection

- Multi-hop is desirable for energy efficiency, cost
- Censored test can be iterated to match arbitrary multi-hop ‘tree’ hierarchy
  \[ \forall \rho = 1 \leftrightarrow \text{centralized} \]
  \[ 0 < \rho < 1 \leftrightarrow \text{data fusion, reduce data bottleneck at the root} \]
- Detection performance can be close to optimal [1]
  - Even \( \rho = 0.01 \) sensors greatly improve performance

Example: distributed anomaly detection

- Parameter $\rho$ selected to constrain mean time between false alarms

![Graph showing mean time between false alarms vs. mean delay, with different levels and parameters $\rho_1$ and $\rho_2$](image)
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 $11 \times 21 = 231$ 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
Theorem

Extended BHH Theorem (Costa&Hero):

\[
L_\gamma(Y_n)/n^\alpha \to \beta_d \int_S f_Y^\alpha(y) dy \quad \alpha = (d - \gamma)/d
\]
Application: ISOMAP Database

• http://isomap.stanford.edu/datasets.html
• Synthesized 3D face surface
• Computer generated images representing 700 different angles and illuminations
• Subsampled to 64 x 64 resolution (D=4096)
• Disagreement over intrinsic dimensionality – d=3 (Tenenbaum) vs d=4 (Kegl)

Resampling Histogram of $d$ hat

Mean GMST Length Function

d=3
H=21.1 bits
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
wMDS 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

WISP: Nov. 04
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.