Watching Traffic for an Anomaly: Data Visualization using Dimensionality Reduction

Neal Patwari, Alfred O. Hero III

University of Michigan, Ann Arbor MI, USA
Dept. of Electrical Engineering and Computer Science
http://www.engin.umich.edu/~npatwari

Workshop on Internet Signal Processing
November 11, 2004

Problem Formulation

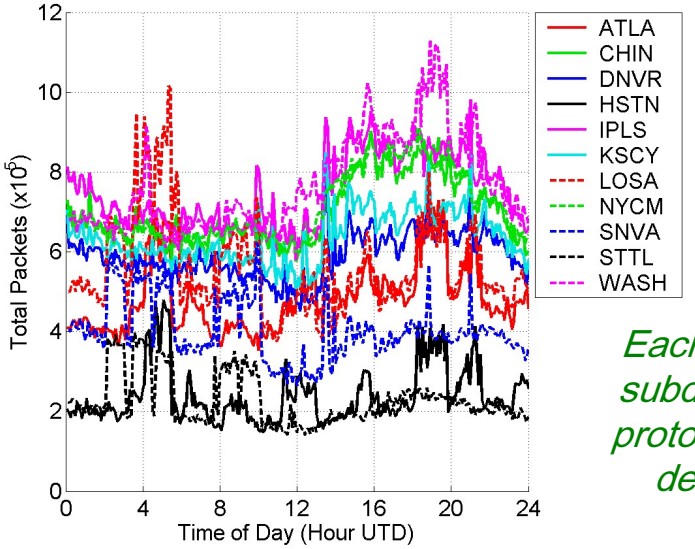
- ‘Bad’ events change traffic over space & time
  - How do you see spatial & temporal characteristics?
- Motivation: Watch changing correlations over space
  - Map the routers based on traffic data ‘closeness’
  - Very close routers = very high correlation
- Goals:
  1. Show dramatic changes in correlation
  2. Show ‘where’ to look in an anomaly

Traffic Measurements

- From NetFlow, aggregate traffic in Δ=5 min intervals
  - Total Packets, Flows or Octets
  - By Port/Protocol (eg. top few appls.)
  - By Source or Destn AS

  → Multidim. vector meas't possible at each router, time

Abilene backbone network: 11 routers

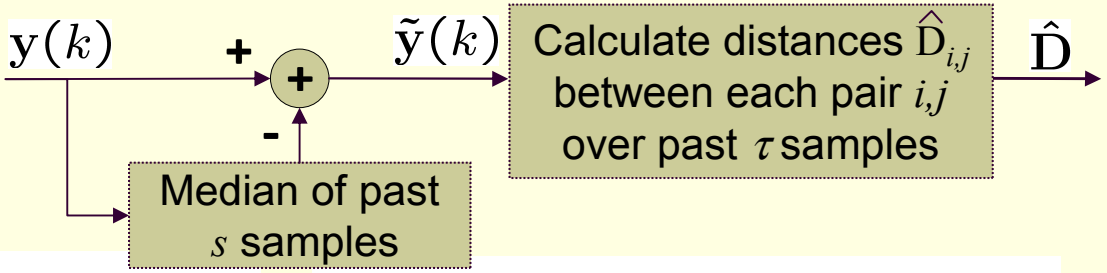


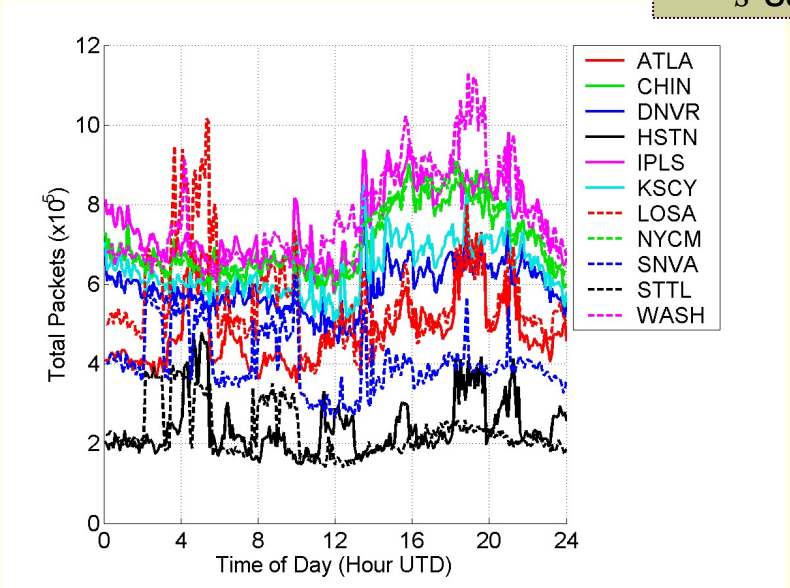
Total packets at each router over time for 11 June '04

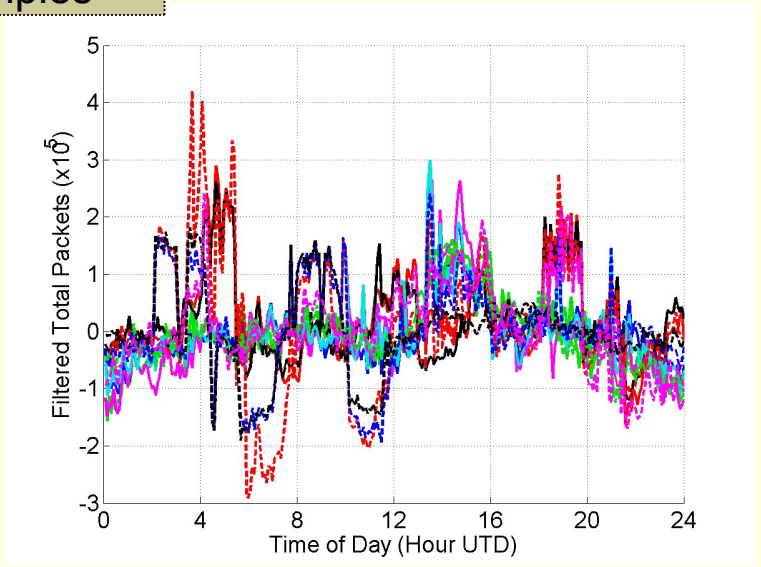
Each plot could be subdivided by port/ protocols, source or destination AS

Approach and Methodology

1. Filter traffic data to remove running mean







2. Estimate distances using L_2 norm (τ past): $\hat{D}_{i,j}^2 = \sum_{\substack{t= \\ k-\tau+1}}^{k} \left\| \tilde{y}_i(t) - \tilde{y}_j(t) \right\|^2$

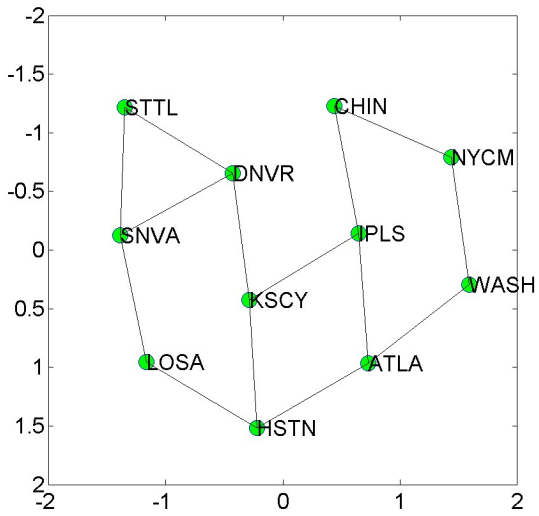
   Or another decreasing fcn of the correlation

Approach and Methodology

3. Pick non-zero weights $w_{i,j}$ for K nearest-neighbors: Eg. $e^{-\gamma \hat{D}_{i,j}}$
4. Find coordinates $\{\mathbf{z}_k\}_k$ which minimize the weighted cost function:

$$\arg\min_{\{\mathbf{z}_i\}} \left\{ \sum_{i,j} \underbrace{w_{i,j}}_{\text{Weight}} \underbrace{\left(\|\mathbf{z}_i - \mathbf{z}_j\| - \hat{D}_{i,j} \right)^2}_{\text{Stress}} + \sum_k \underbrace{r_k}_{\text{Weight}} \underbrace{\|\mathbf{z}_k - \bar{\mathbf{z}}_k\|^2}_{\text{Prior}} \right\}$$

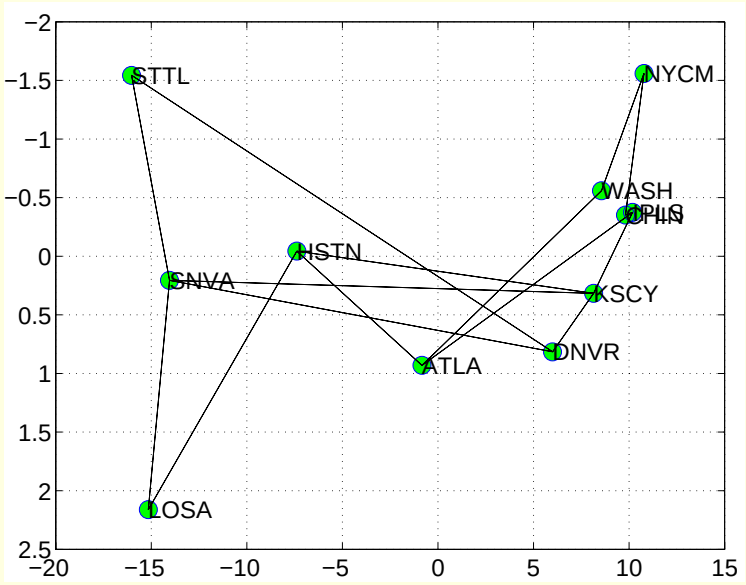
- Distributed, Weighted Multidimensional Scaling (dwMDS)
  - Localized data sharing
  - Weights distances according to expected accuracy
  - Distributed minimization
  - Majorization method guarantees improvement at each round

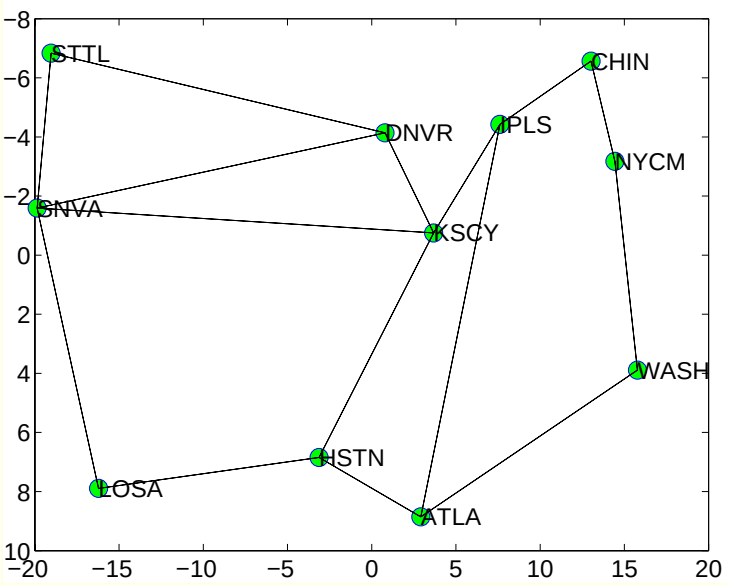


Example Prior on coordinates: equal-distance links

Preliminary Results

- June 11, 2004: For $\tau = 288$ data plotted previously





- MDS-generated map

MDS overly weights long-range distances

- dwMDS-generated map

$(r = 0, K = 5, w_{ij} = 1)$

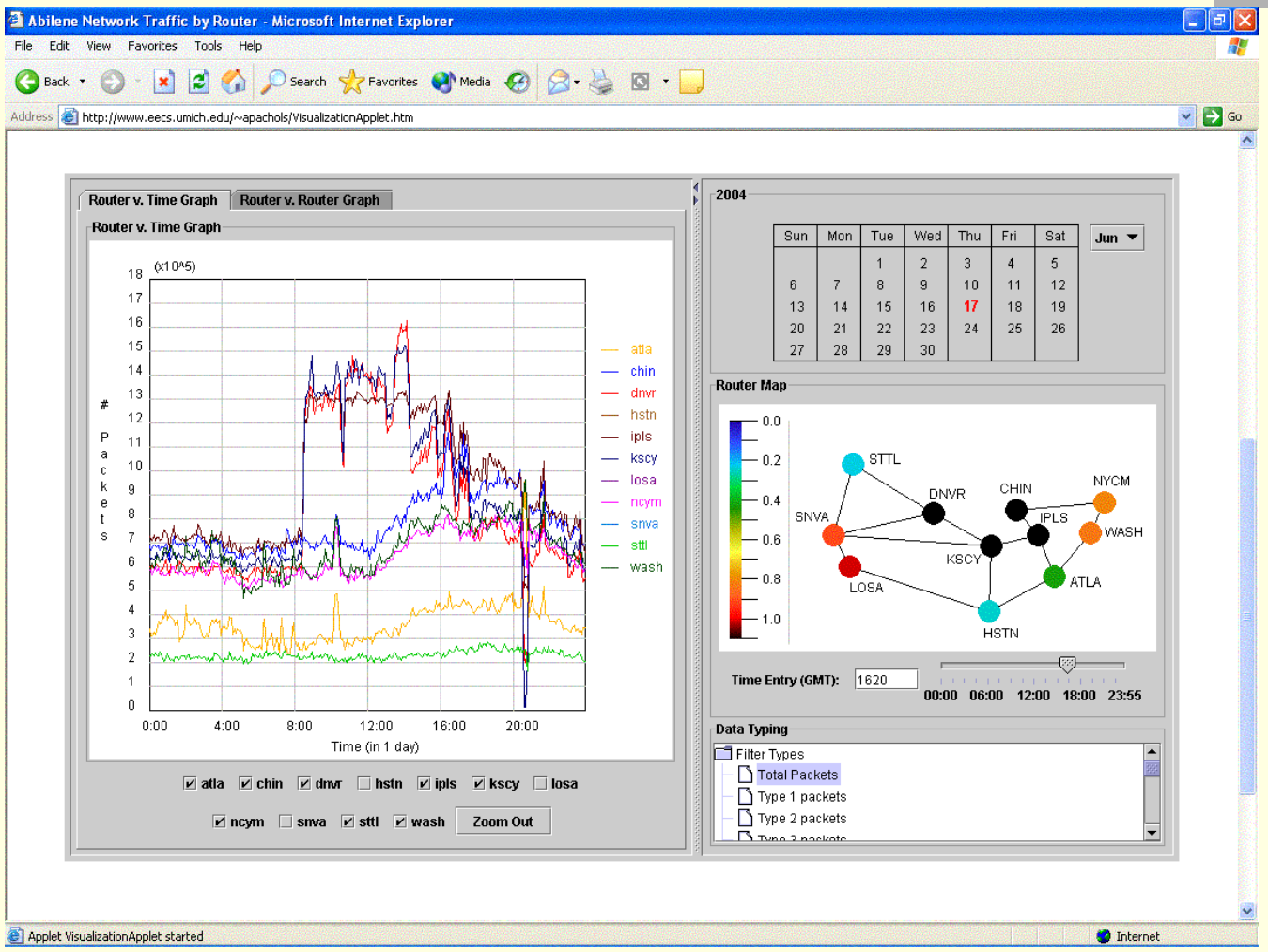
Validation

- Video of 6 – 12 June ’04
  - 16 hour memory (200-dim vectors)
  - New map estimated each 20 minutes

Next Steps

- Apply to larger networks
  - Test K-nearest-neighbors, distributed calculation
- Use higher-dimensional data
  - Visualization becomes more important as dim. Increases
  - Change in distribution of traffic will affect map
  - Eg: Flows, Octets, and Packets
  - Eg: Top n Applications (like FlowScan)
  - Eg: Source/Dest AS
  - Eg: Link data or OD-flow data vs. router data
  - Use Transformed Data (Wavelet, Spectral, …)
- Verify vs. known anomalies
- Implement in real-time web Applet

Space-Time Visualization Applet



Plan: Implement the dynamic correlation-map in an accessible, multifunction visualization tool.

- http://www.eecs.umich.edu/~apachols/VisualizationApplet.htm **Try it!**