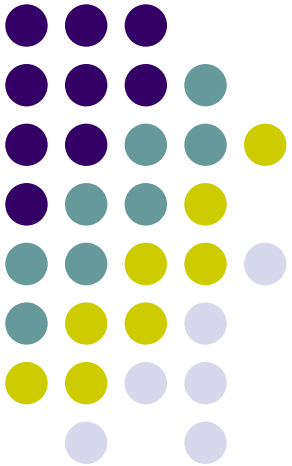
Sensitivity of PCA for Traffic Anomaly Detection

Evaluating the robustness of current best practices

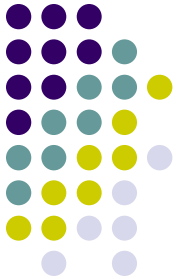


Haakon Ringberg[1], Augustin Soule[2],
Jennifer Rexford[1], Christophe Diot[2]
[1]Princeton University, [2]Thomson Research

Outline

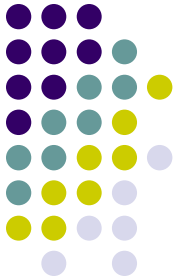
- Context
  - Background and motivation
  - Bigger picture
  - PCA (subspace method) in one slide
- Challenges with current PCA methodology
- Conclusion & future directions

Background



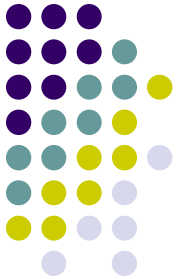
- Promising applications of PCA to AD
  - [Lakhina et al, SIGCOMM 04 & 05]
- But we weren't nearly as successful applying technique to a new data set
  - Same source code
- What were we doing wrong?
  - Unable to tune the technique

Bigger Picture



- Many statistical techniques evaluated for AD
  - *e.g.*, Wavelets, PCA, Kalman filters
  - Promising early results
- But questions about performance remain
  - What did the researchers have to do in order to achieve presented results?

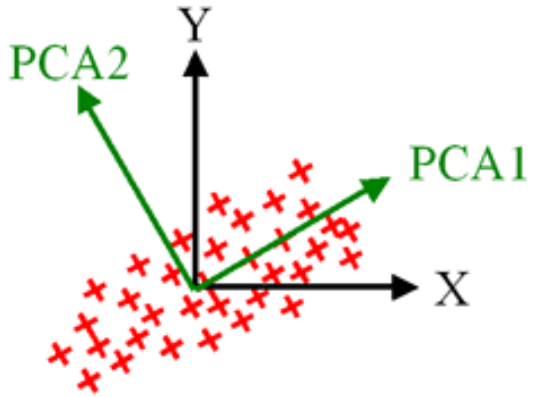
Questions about techniques

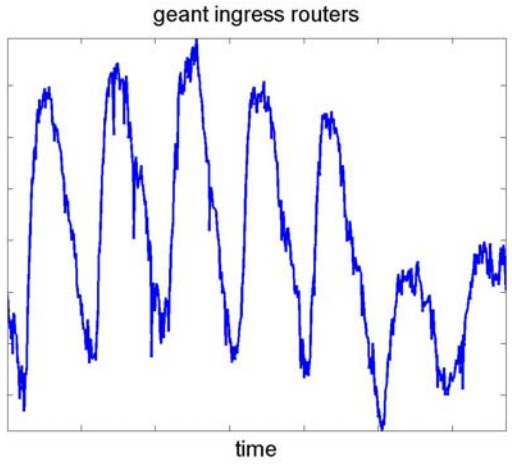


- “Tunability” of technique
  - Number of parameters
  - Sensitivity to parameters
  - Interpretability of parameters
- Other aspects of robustness
  - Sensitivity to drift in underlying data
  - Sensitivity to sampling
- Assumptions about the underlying data

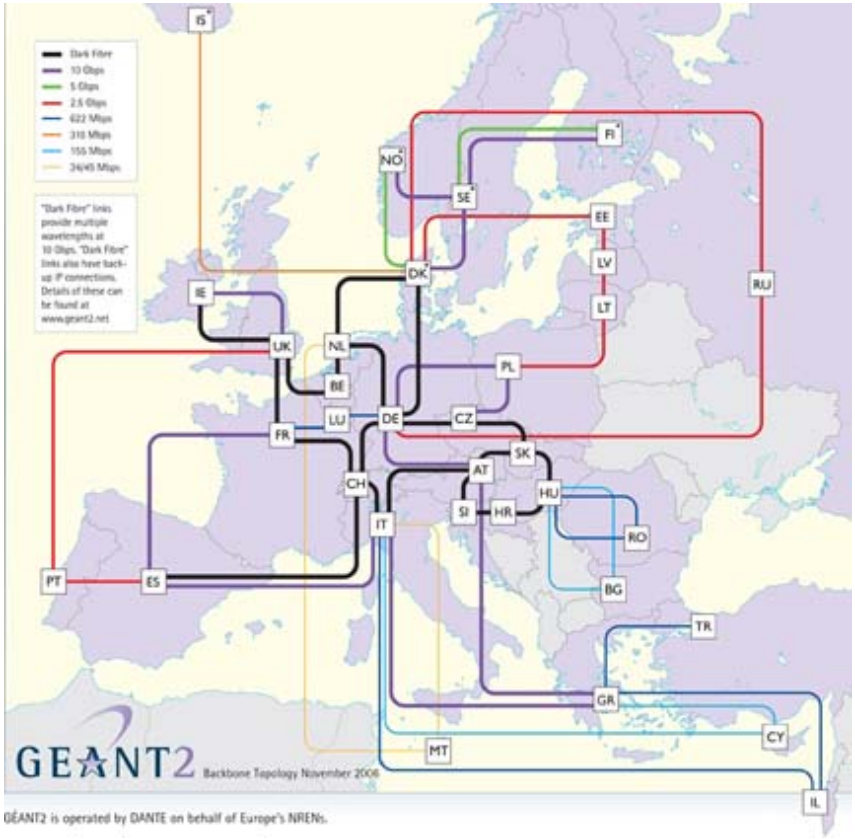
Principal Components Analysis (PCA)

- PCA transforms data into new coordinate system
- Principal components (new bases) ordered by captured variance
- The first *k* (top_k) tend to capture periodic trends
  - *normal subspace*
  - vs. *anomalous subspace*





Data used



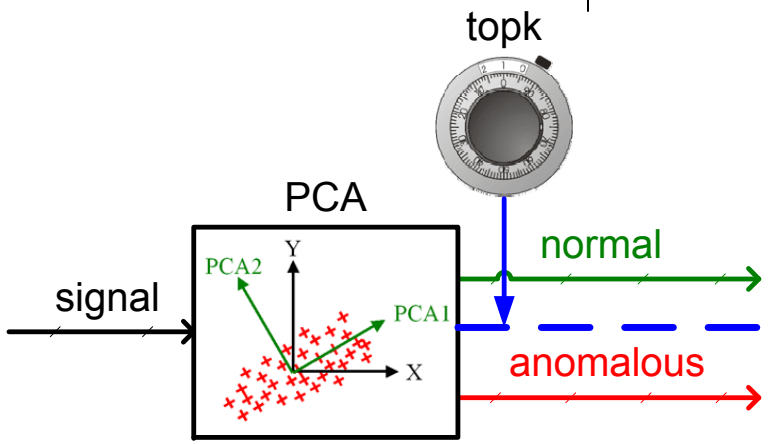
- Géant and Abilene networks
- IP flow traces
- 21/11 through 28/11 2005
- Detected anomalies were manually inspected

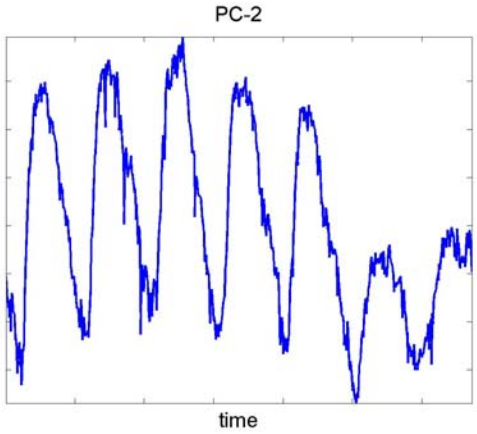
Outline

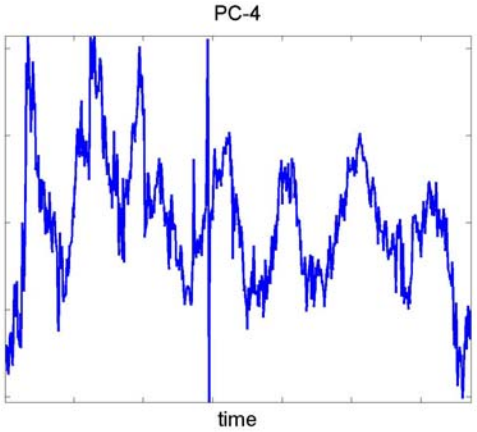
- Context
- Challenges with current PCA methodology
  - Sensitivity to its parameters
  - Contamination of normalcy
  - Identifying the location of detected anomalies
- Conclusion & future directions

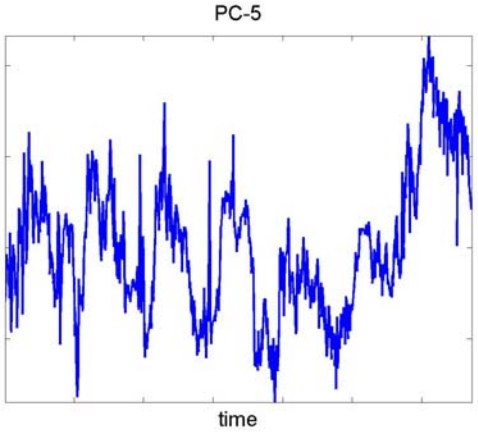
Sensitivity to top_k

- Where is the line drawn between normal and anomalous?
- What is *too* anomalous?

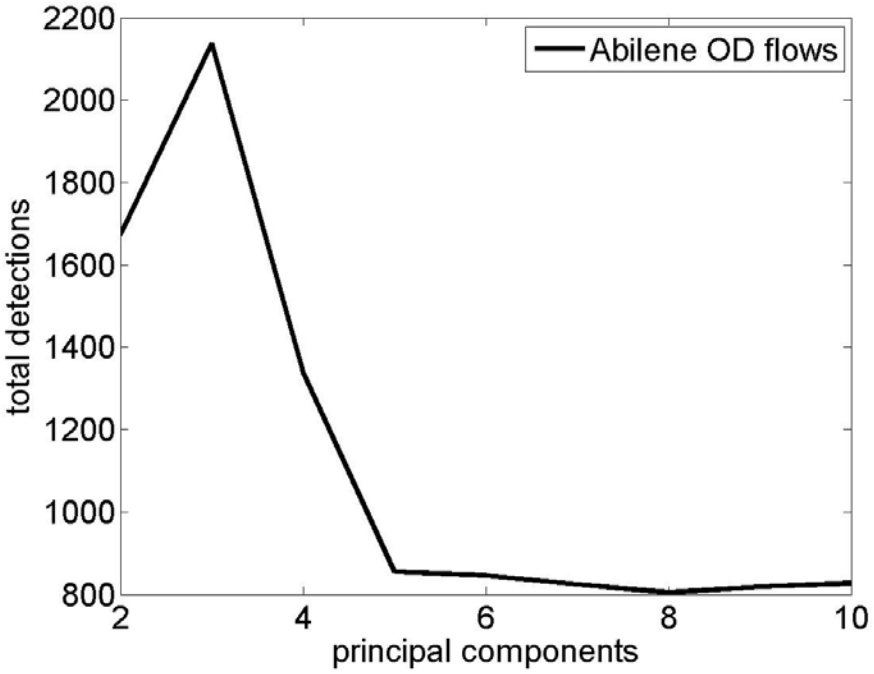






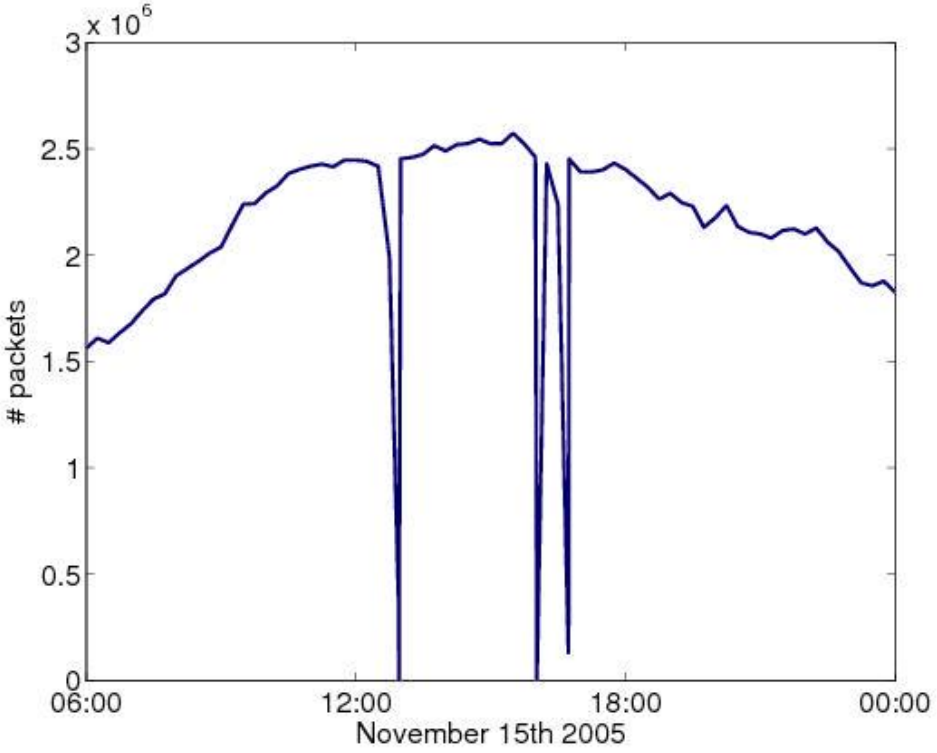


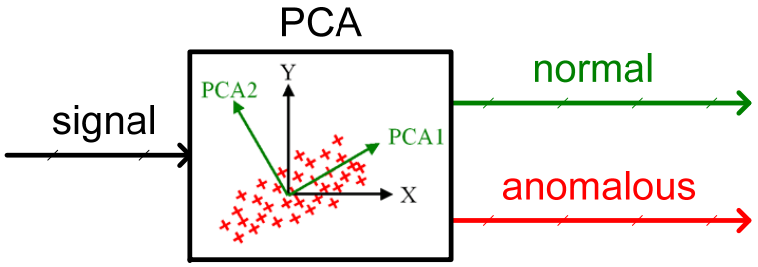
Sensitivity to top_k



- Very sensitive to top_k
  - Total detections and FP
- Not an issue if top_k were tunable
- Tried many methods
  - 3σ deviation heuristic
  - Cattell's Scree Test
  - Humphrey-Ilgen
  - Kaiser's Criterion
- None are reliable

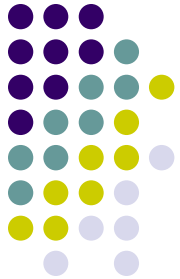
Contamination of normalcy





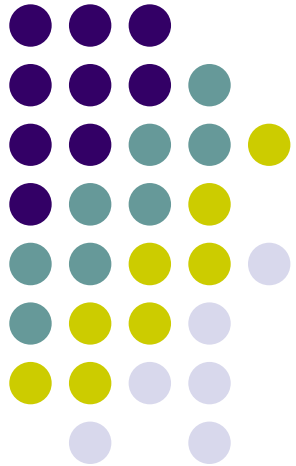
- Large anomalies may be included among top_k
- Invalidates assumption that top PCs are periodic
- Pollutes definition of normal
- In our study, the outage to the left affected 75/77 links
  - Only detected on a handful!

Conclusion & future directions



- PCA (subspace method) methodology issues
  - Sensitivity to top_k parameter
  - Contamination of normal subspace
  - Identifying the location of detected anomalies
- Generally: room for rigorous evaluation of statistical techniques applied to AD
  - Tunability, robustness
- Assumptions about underlying data
  - Under what conditions does method excel?

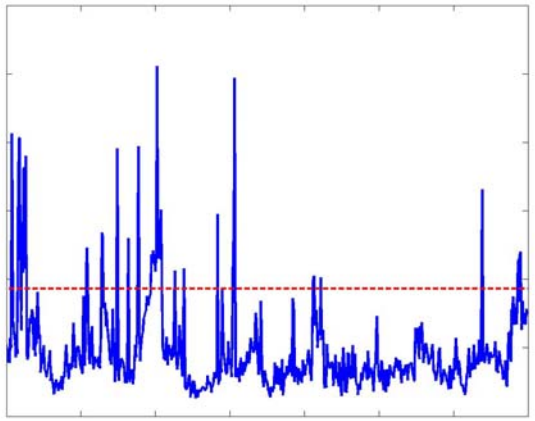
Thanks!
Questions?

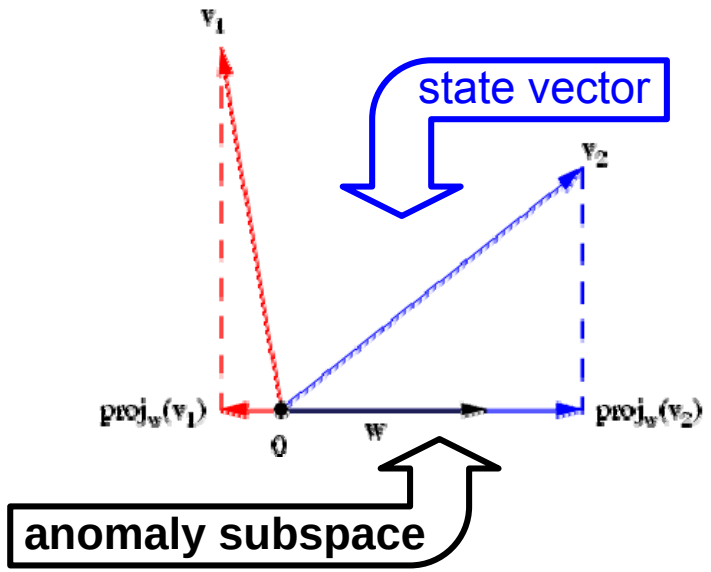


Haakon Ringberg
Princeton University Computer Science
http://www.cs.princeton.edu/~hlarsen/

Identifying anomaly locations

- Spikes when state vector projected on anomaly subspace
  - But network operators don't care about this
  - They want to know where it happened!
- How do we find the original location of the anomaly?





Identifying anomaly locations

- Previous work used a simple heuristic
  - Associate detected spike with *k* flows with the largest contribution to the state vector *v*
- No clear *a priori* reason for this association