Can Enterprise Operations Software Really Work in Real Time?

Minos Garofalakis

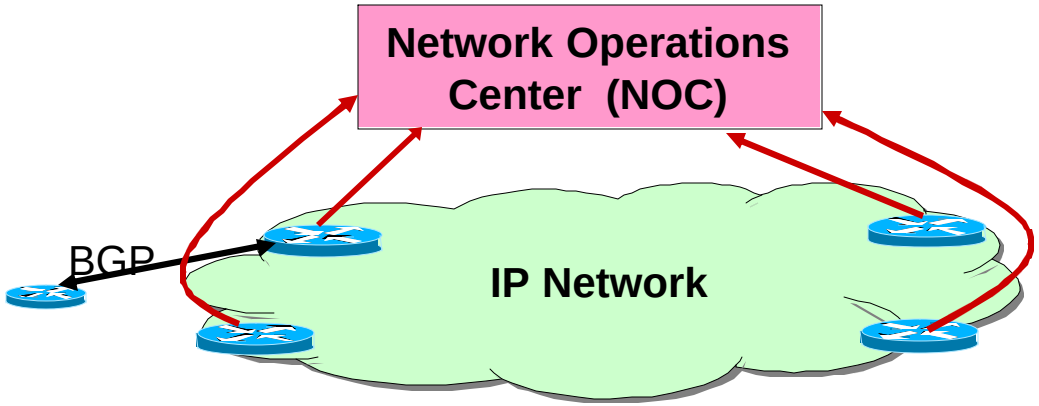
Intel Research Berkeley

minos.garofalakis@intel.com

It Depends ☺

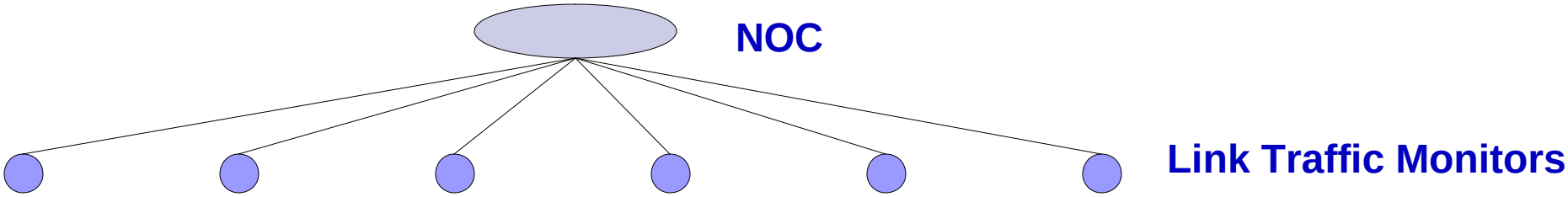
- Seems like an impossible task!
  - Large amounts of data
  - Complex data processing/analysis requirements
  - Physically-distributed operations
  - …
- Approximation (ε,δ, …) to the rescue!
  - For several tasks (pattern mining, anomaly detection, etc.) precise answers are not always required
  - Can trade-off *space, processing, and network* costs with answer quality
    - (Near) real-time operation

Enterprise Network Monitoring



- Massive, continuous measurement streams from hundreds of locations
  - Large enterprises (e.g., Intel) collect 100s of GBs each day!
- Need real-time monitoring for exceptional ***network-wide*** conditions
  - Anomalies, DDoS attacks, …
- Cannot afford to continuously centralize the streams

Distributed PCA Tracking



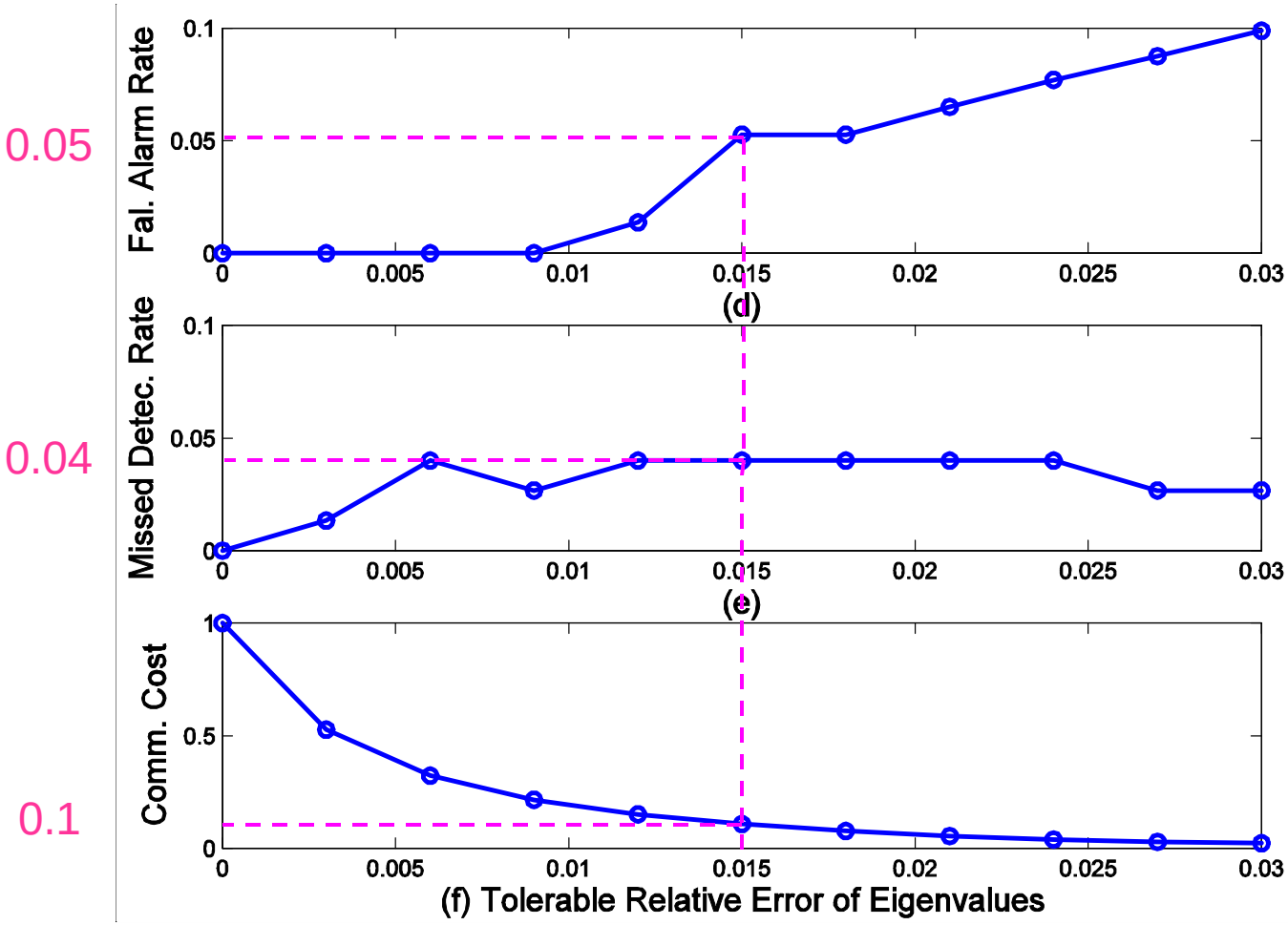
$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2n} \\ \cdots & \cdots & & \cdots & \cdots \\ x_{m1} & x_{m2} & x_{m3} & \cdots & x_{mn} \end{bmatrix} \downarrow \text{time window} = Y$$

- Threshold the total energy of the low PCA coefficients of Y = Robust indicator of network-wide anomalies [Lakhina et al.'04]
- Optimize for *common case*: Filter most traffic locally
  - NOC only sees approximate "filtered" version of data
  - Effect on PCA quantified through *stochastic matrix perturbation theory* [Huang et al.'06]

Preliminary Results

- Implemented and tested approximate PCA anomaly monitoring over real-life network data
  - Can reduce communication overhead by over an order of magnitude
    - Increase false positives/negatives by less than 3%
- Approximation can help with real-time distributed data-stream analysis
  - well, in some cases… ☺

Preliminary Results



- *Approximation can help with real-time distributed streaming analysis*