Detection of Cyber-Physical Faults and Intrusions from Physical Correlations

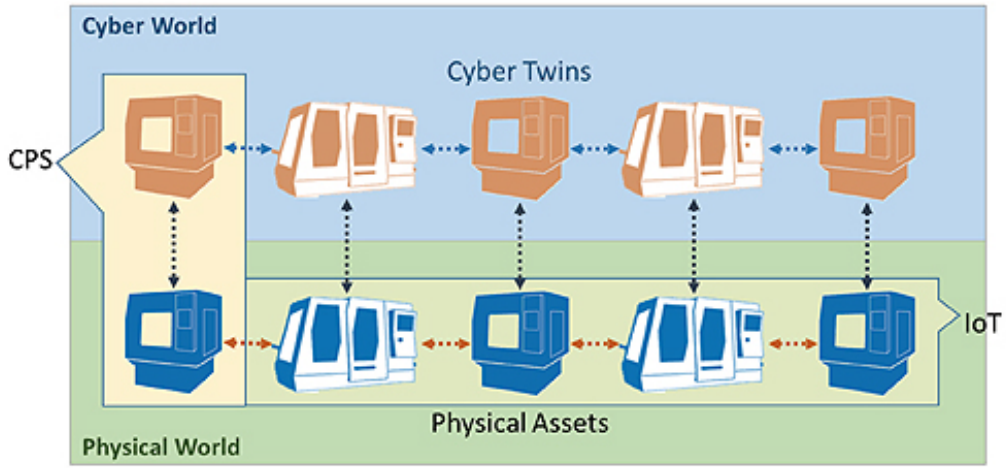
Andrey Lokhov, Nathan Lemons, Thomas McAndrew, Aric Hagberg, Scott Backhaus

Theoretical Division/Center for Nonlinear Studies
Los Alamos National Laboratory

ODD 4.0, August 14, 2016

Outlier detection: protection of cyber-physical systems

Examples of CPS: smart grids, avionics, autonomous control systems, ...



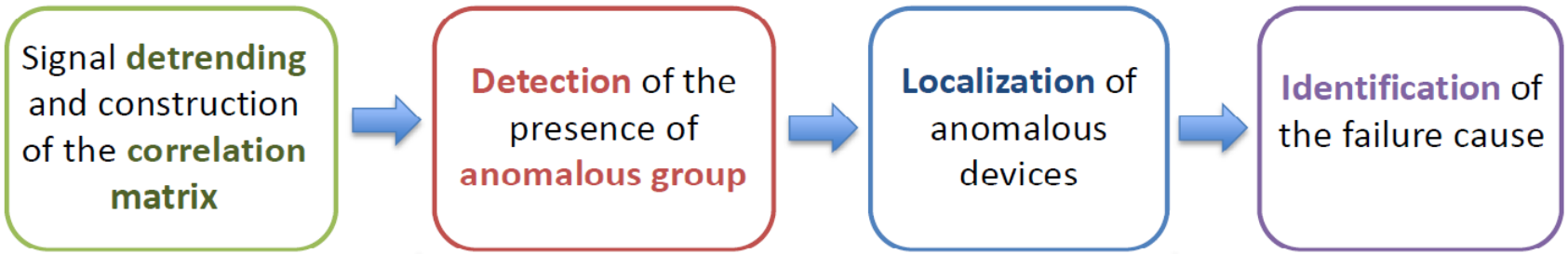
Control from the cyber part $\leftrightarrow$ feedback from sensors in the physical part

- ✓ Intrusions in cyber system $\rightarrow$ signatures in physical network
- ✓ **Task:** detection of these signatures with minimal assumptions

Setting and steps of solution

General setting: using sensor measurements of physical signals, detect and localize failures and intrusions in cyber-physical system

- ✓ From analysis of individual signals to exploration of correlations
- ✓ **Require:** scalability, generality, robustness, low complexity

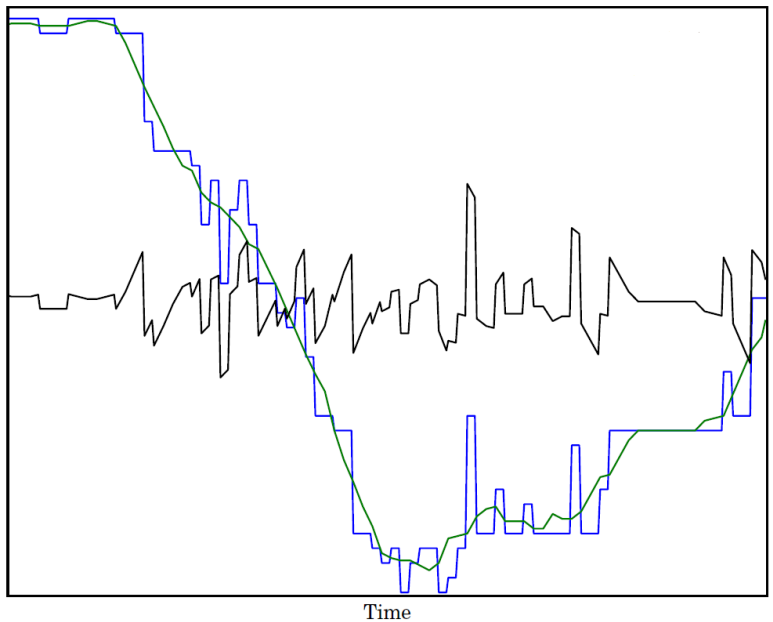


(generality) (scalability) (identification limits)

Signal detrending

Given: N heterogeneous data streams $X_i(t)$.

$$X_i(t) = \underbrace{Y_i(t)}_{trace} + \underbrace{N_i(t)}_{noise} + \underbrace{S_i(t)}_{signal}$$

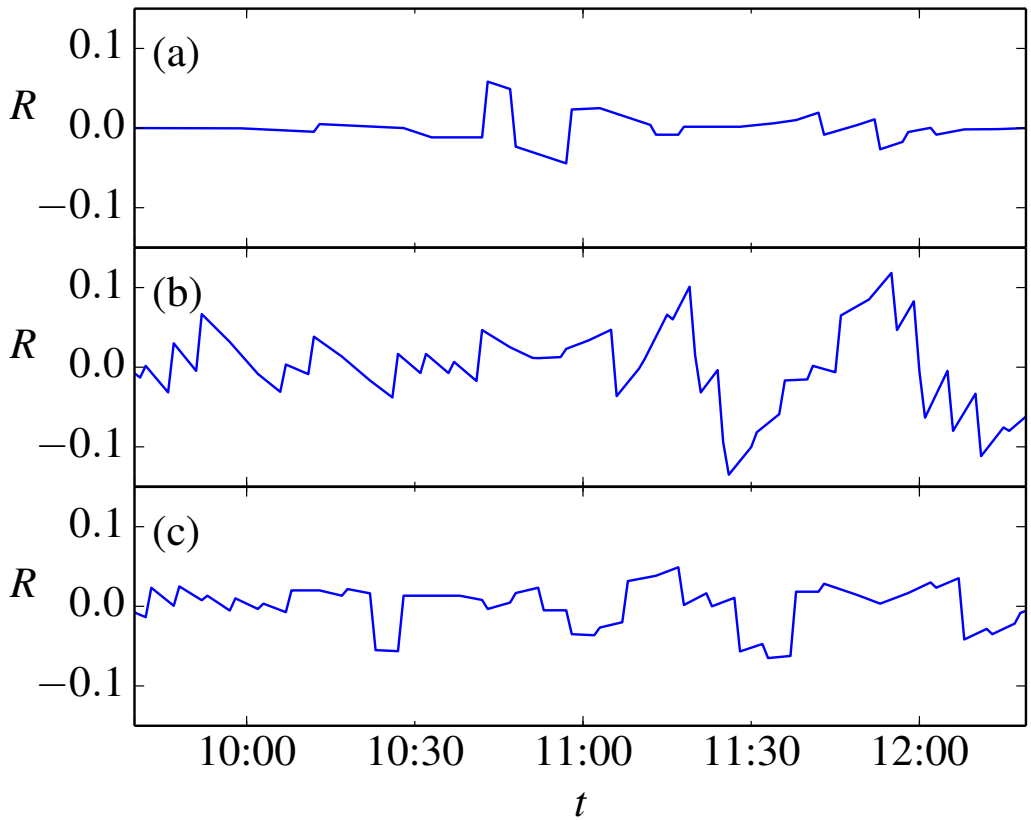


Observation: $R_i(t) = X_i(t) - Y_i(t)$ correlated for $i \in U$ s.t. $S_i(t) \neq 0$

Signal detrending

$Y_i(t)$ unknown $\rightarrow$ rolling mean $\bar{X}_i(t) = \frac{1}{\tau_{\mathrm{av}}} \sum_{t'=t-\tau_{\mathrm{av}}/2}^{t+\tau_{\mathrm{av}}/2} X_i(t')$

Testing correlations between $R_i(t) = X_i(t) - \bar{X}_i(t)$.



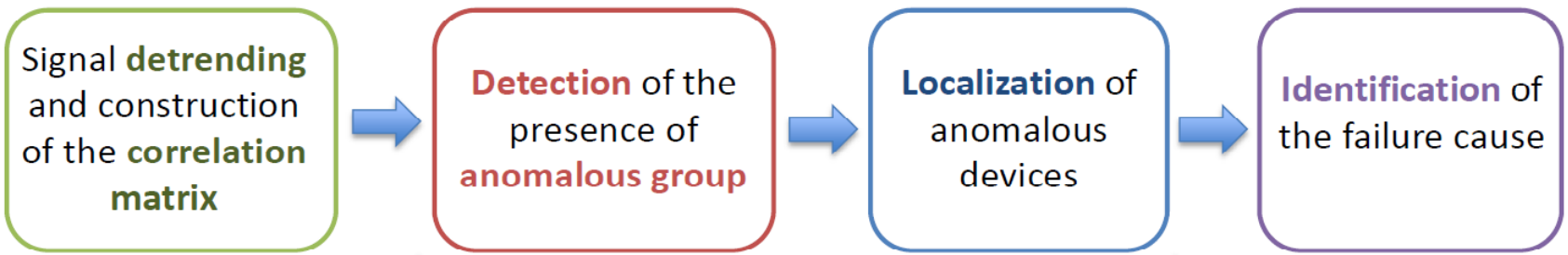
Construction of the correlation matrix

$M_{ij}(t) = \text{corr}(R_i(t), R_j(t))$ over window $[t - \tau_{\text{corr}}, t]$, $M_{ii}(t) = 0$

Normal operations: $\mathbb{E}[M_{ij}(t)] = 0$

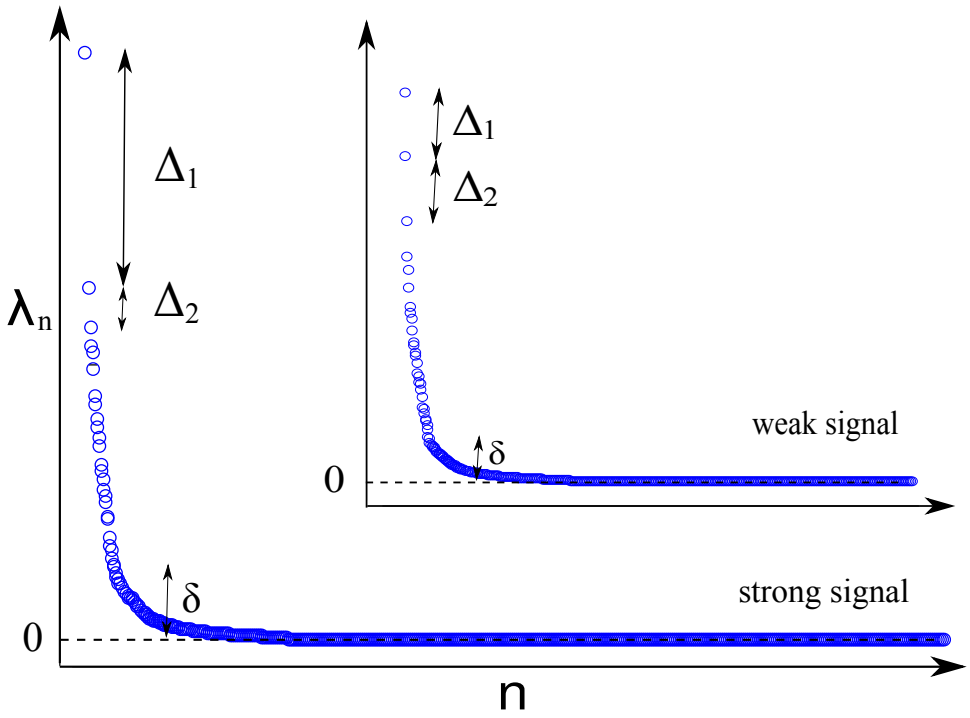
Attack or failure: $\exists U \subset V$ such that $S_i(t) \neq 0$ for $i \in U$, and

$$\mathbb{E}[M_{rs}(t)] = m_{rs} > 0$$

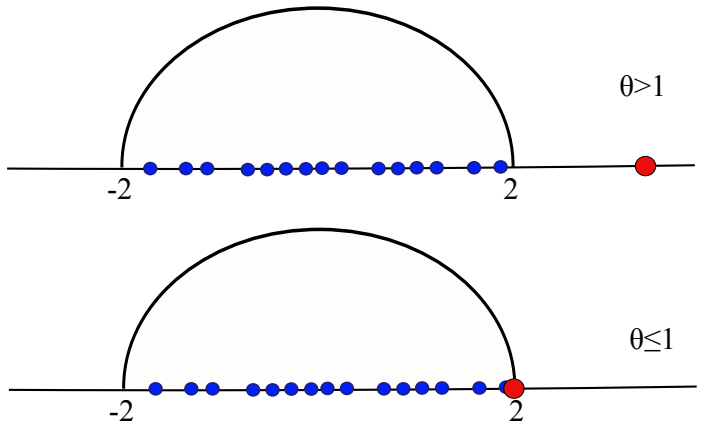


Detection: is there any significant fault?

Spectral gap indicates the presence of an anomalous subgroup



Ideal case: $M = \theta uu^T + W$

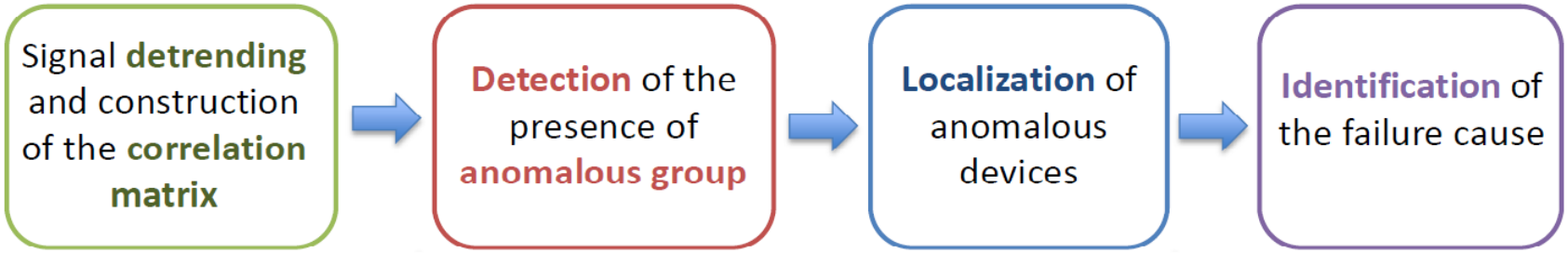


[Benaych-Georges & Rao, 2011]

Estimation of the noise characteristic scale $\delta = \sqrt{\frac{1}{N-2}\sum_{i\geq 2}\Delta_i^2}$

Detection criterion: $\boxed{\Delta_1 > \Delta_2 + \delta}$

Localization: low-rank approximation



Approximation: rank-1 matrix perturbed by noise

$$\begin{bmatrix} m & 0 & 0 & m & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ m & 0 & 0 & m & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} + \widehat{W}$$

$$\widehat{M} = \arg\min_{\widehat{M}} \|M - \widehat{M}\|_F \text{ s.t. } \mathrm{r}(\widehat{M}) = 1$$

SVD: $\widehat{M} = \sigma q q^T$.

Sparse PCA: k largest elements in q

[Zhang *et al.*, 2002], [Papailiopoulos *et al.*, 2013]

Size of the anomalous submatrix?

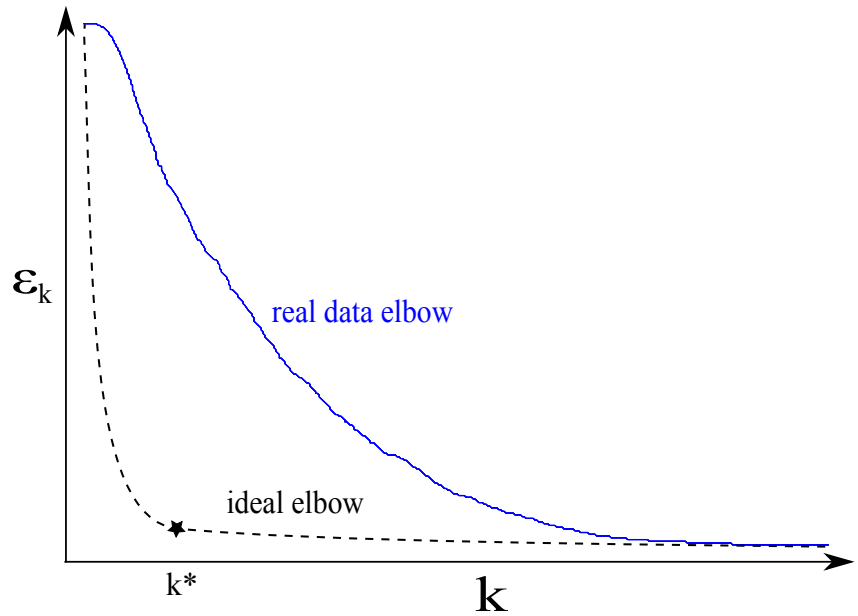
$$\varepsilon_k = \| M - \sigma_k q_k q_k^T \|_F$$

k unknown, so we test $k^* = \sqrt{N}$

Practically achievable bound

[Hajek *et al.*, 2015]

[Deshpande & Montanari, 2015]



Localization: biclustering methods

Finding large average submatrix without low rank assumption

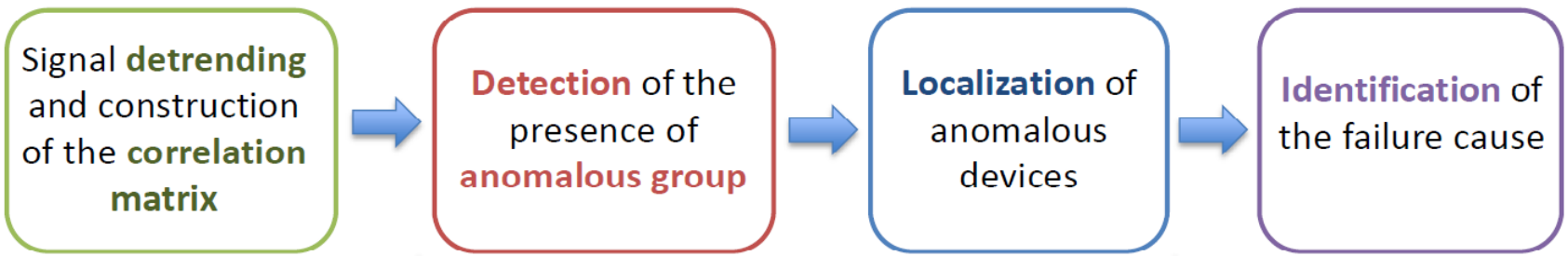
Algorithm 1: iterate until convergence best rows and columns

[Shabalin *et al.*, 2009]

Algorithm 2: greedily add best rows and columns one-by-one

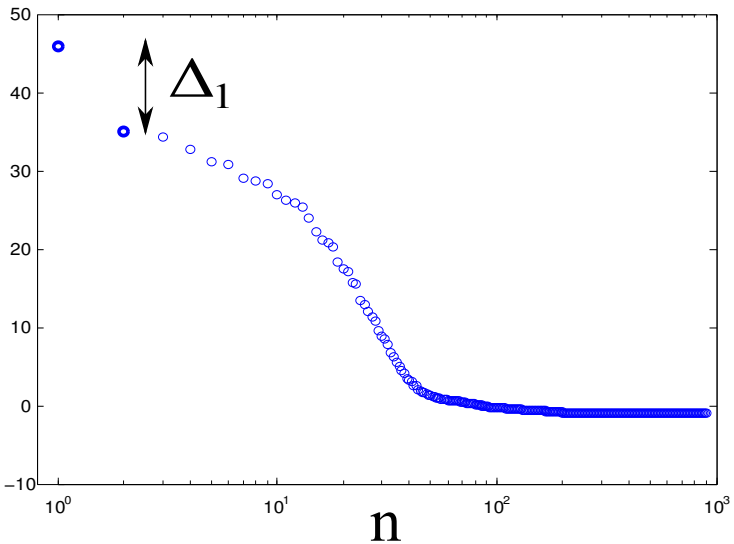
[Gamarnik & Li, 2016]

- ✓ Converge to local optimum
- ✓ L warm starts
- ✓ Again, the strategy is to test $k^* = \sqrt{N}$



First test: synthetic data

$N = 900$ random walks, with $k = 50$ among them correlated

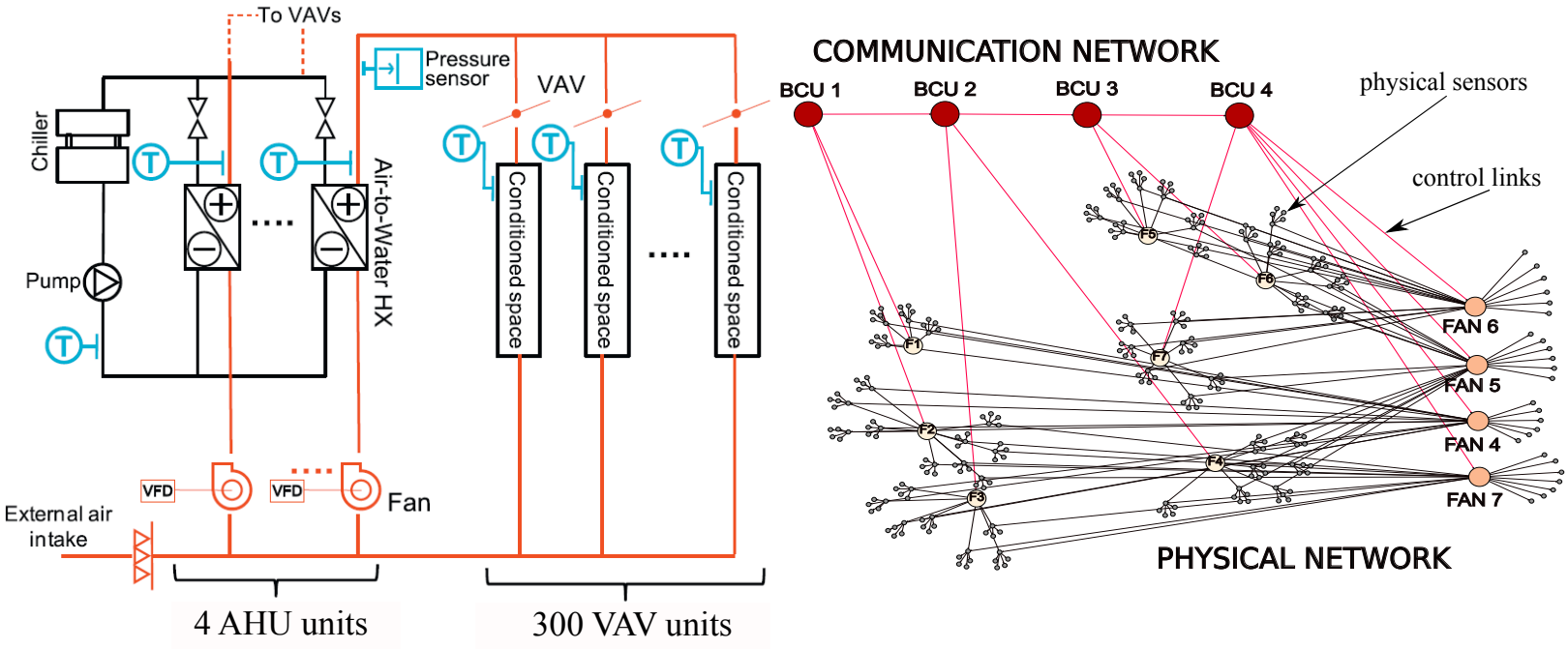


$k^* = 30$: all correlated signals correctly identified

$k = 50$: 5 mistakes by PCA, 1 by biclustering (but requires $L = 3 \cdot 10^4$)

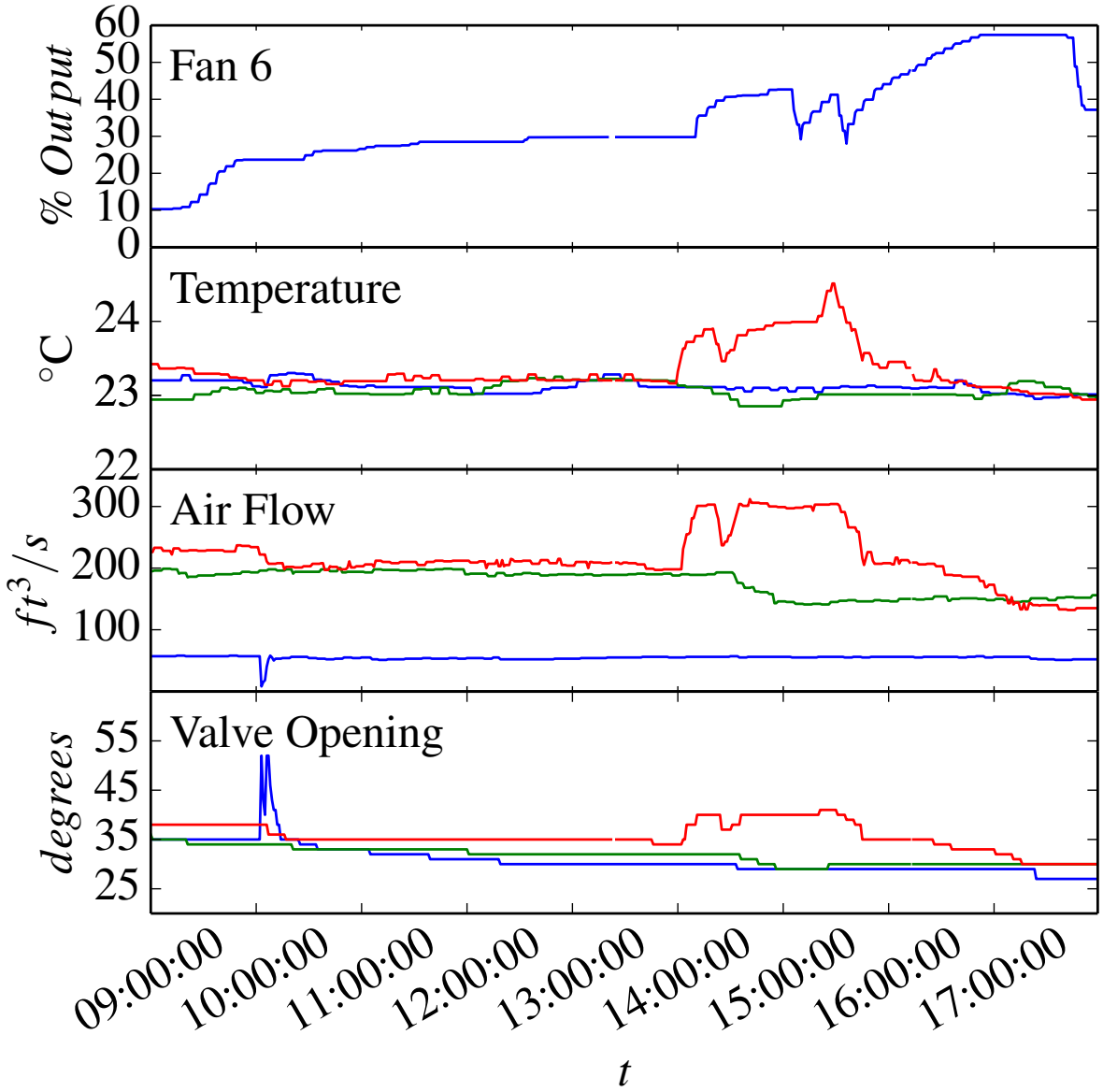
Experiments with real data: AC system in a large building

30,000 m^2 office building, about 900 sensors in conditioned spaces

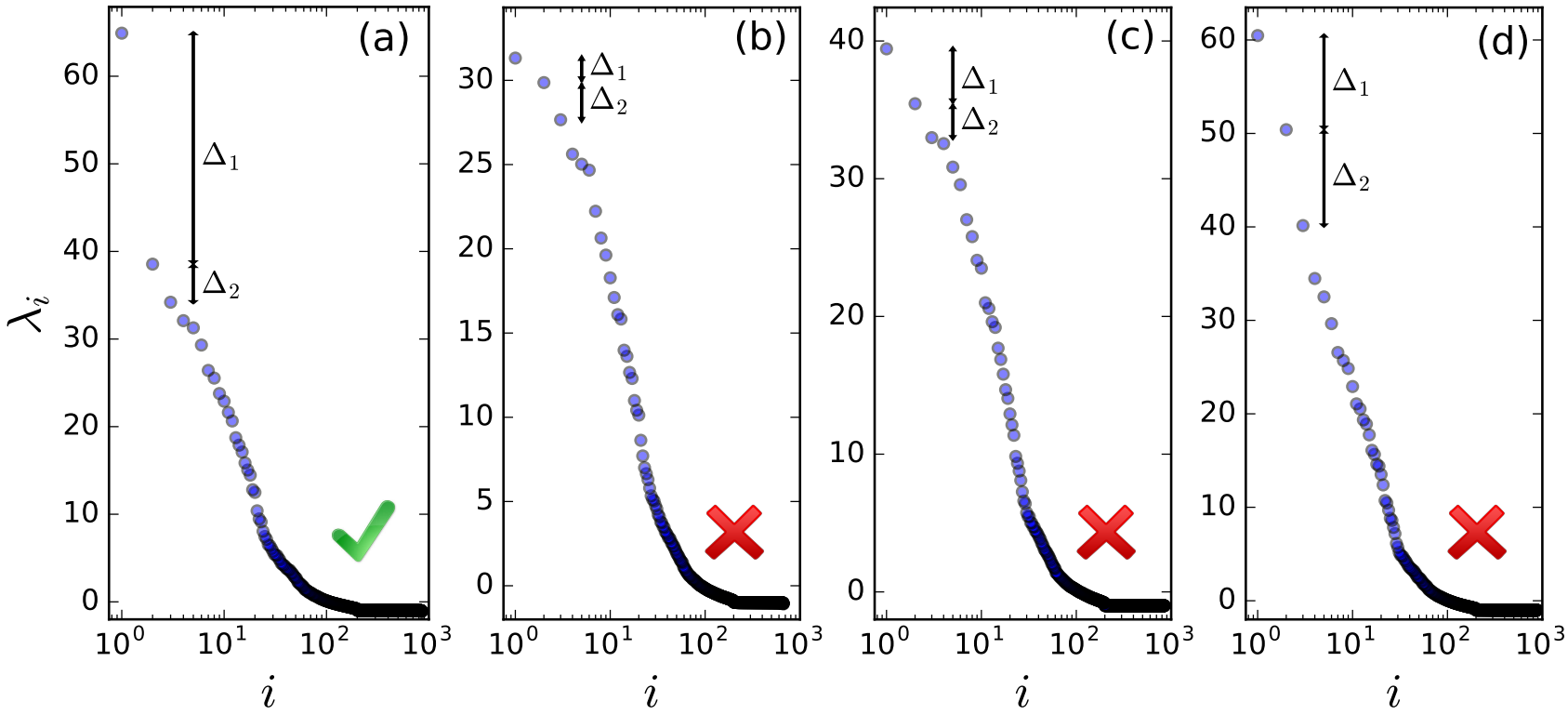


1000 heterogeneous data streams sampled once per minute

Experiments with real data: faulty fan behavior



Experiments with real data: algorithm performance

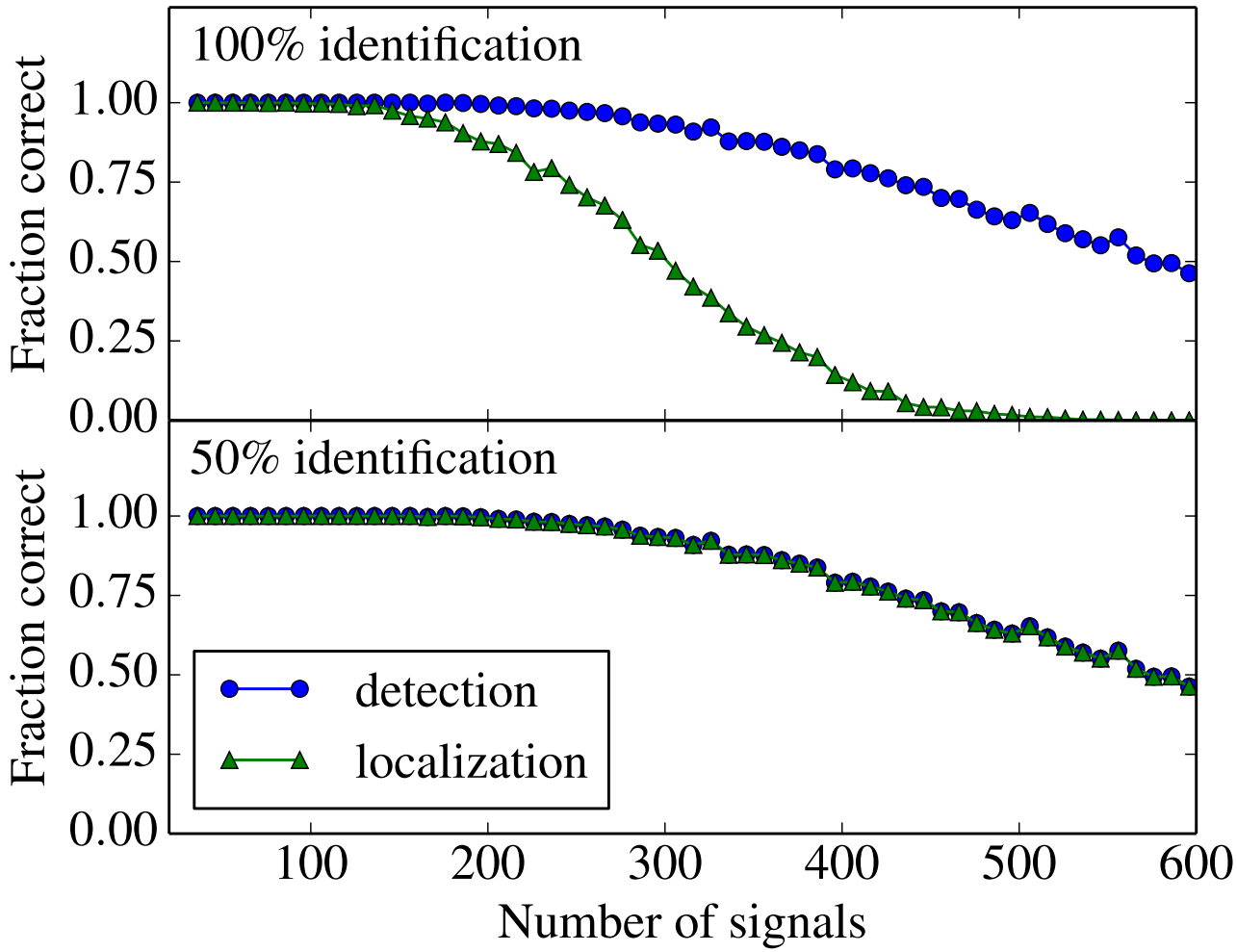


(a) Anomalous signals included (b) Anomalous signals excluded
(c) No oscillations (d) Oscillations over a longer time period

✓ All $k^* = 30$ localized sensors are related to the anomalous fan

Identification limits from controlled experiments

Playing with set points related to $k = 16$ sensors; detection and localization performance as a function of N total streams considered



Conclusion and path forward

- ✓ Protocol for identification of anomalous sensors
- ✓ Insight for localization of failure source
- ✓ Continuation of controlled experiments with other attacks
- ✓ Inclusion of (some) communication data from the cyber part