UPPSALA UNIVERSITET

Data Injection Attacks against Feedforward Controllers

André Teixeira

Assistant Professor

Uppsala University, Sweden

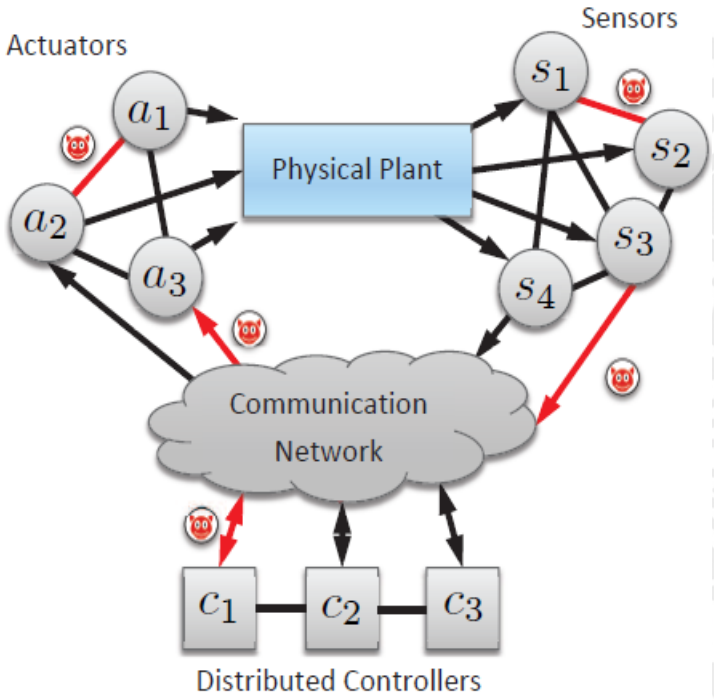
andre.teixeira@angstrom.uu.se

www.andre-teixeira.eu

Cyber-Secure and Resilient Networked Control Systems

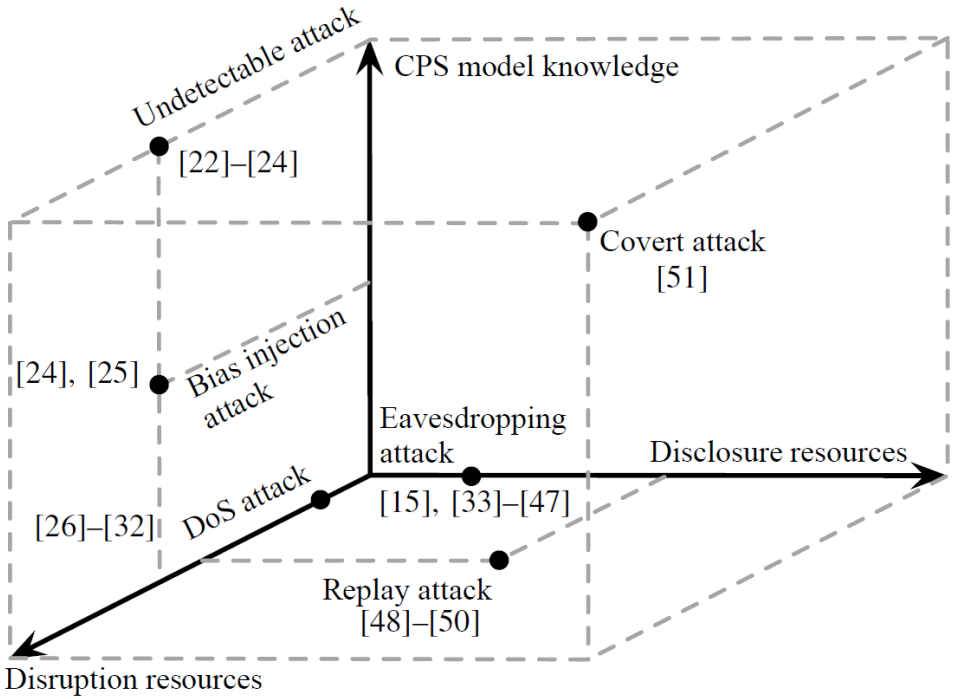
- Networked control systems are to a growing extent based on **open communication and software technology**

- Leads to **increased vulnerability** to cyber-threats with many potential points of attacks
- Cyber-attacks can have dramatic physical impact



- **How to model adversaries and attacks?**
- **How to compute impact of attacks?**
- How to measure vulnerability?
- How to design protection and detection mechanisms?

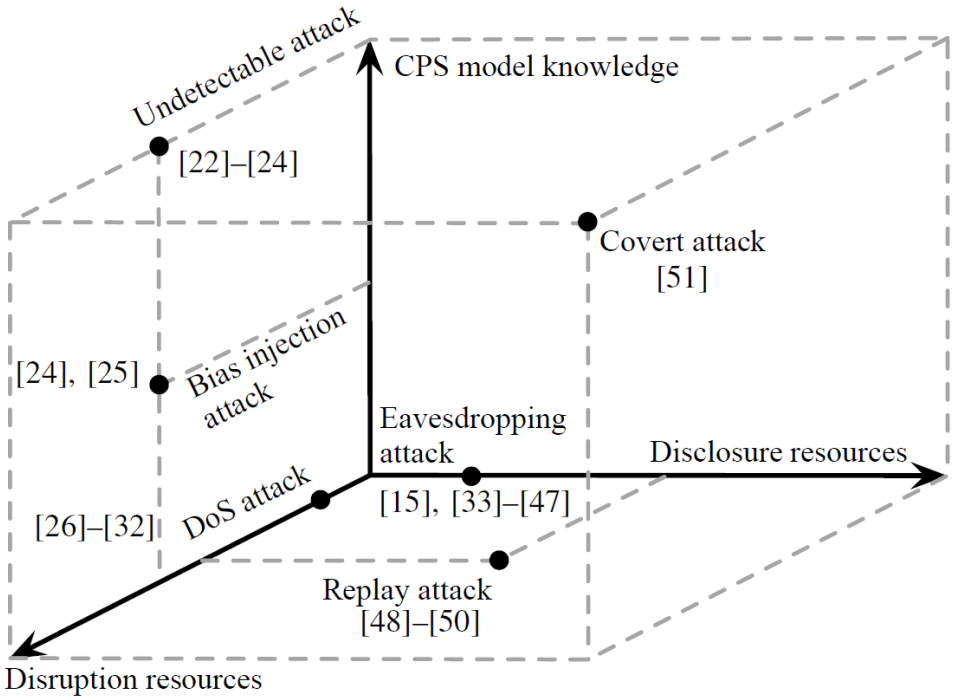
Secure Control Systems



[Chong *et al.*, ECC, 2019]

Secure Control Systems

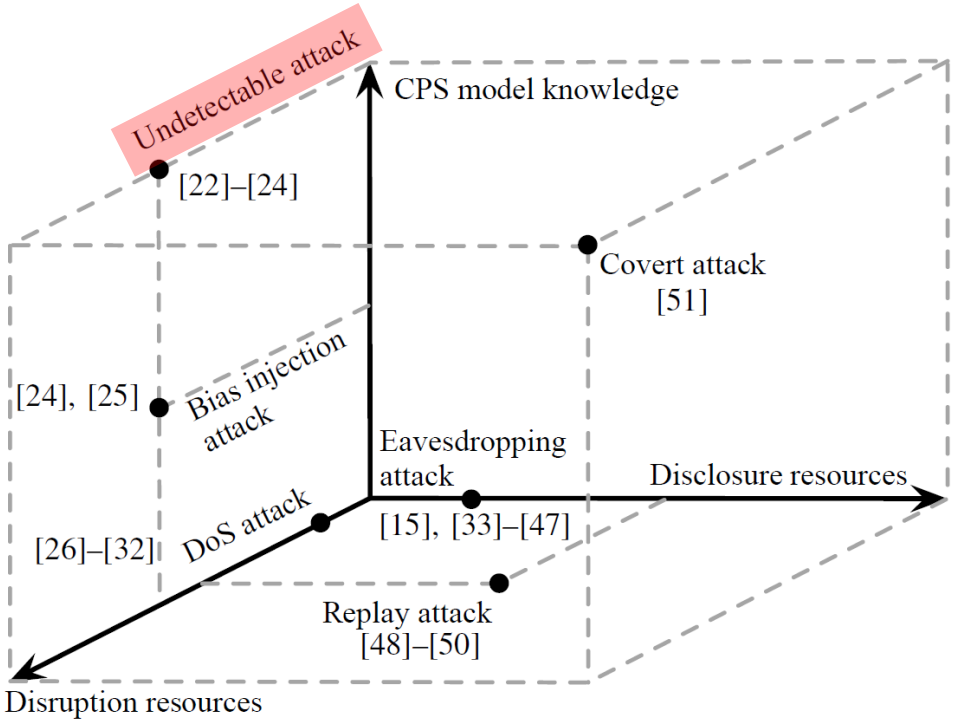
- Exciting field, plenty of open questions
  - Tutorial session at ECC



[Chong *et al.*, ECC, 2019]

Secure Control Systems

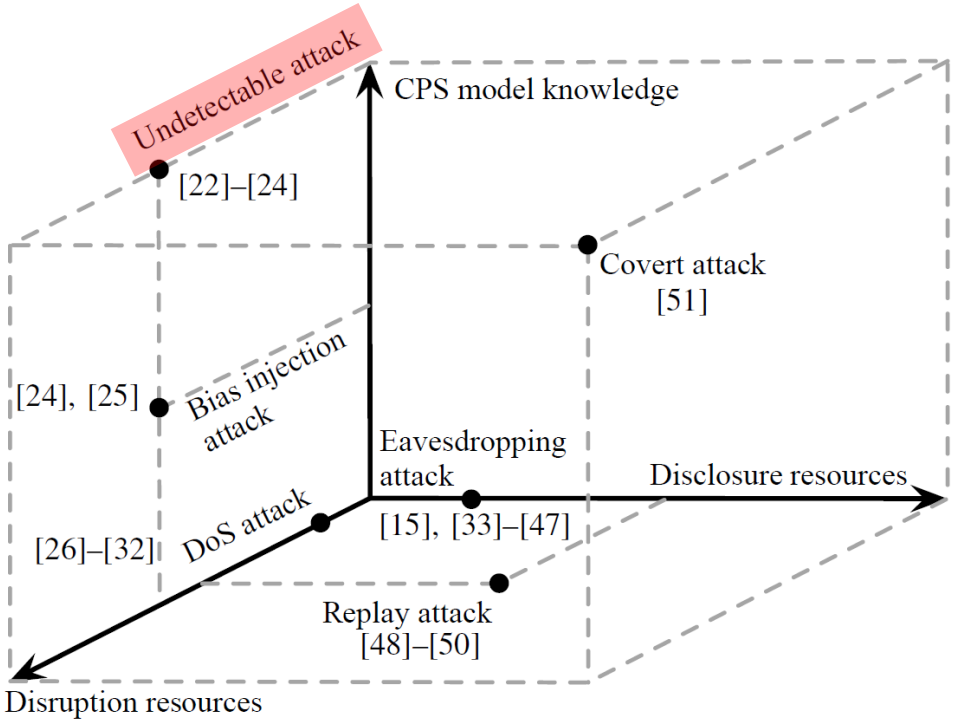
- Exciting field, plenty of open questions
  - Tutorial session at ECC
- Undetectable attacks have been investigated
- Focus on attacks on sensors and/or actuators



[Chong *et al.*, ECC, 2019]

Secure Control Systems

- Exciting field, plenty of open questions
  - Tutorial session at ECC
- Undetectable attacks have been investigated
- Focus on attacks on sensors and/or actuators
- No results w.r.t. attacks on disturbance measurements



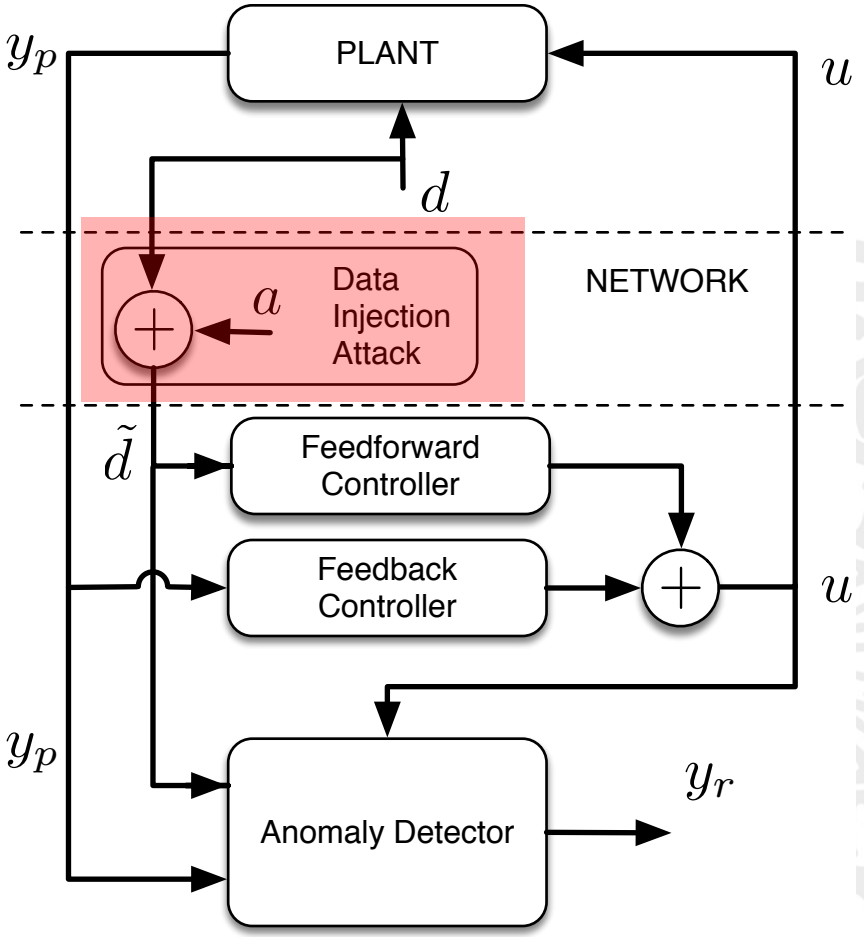
[Chong *et al.*, ECC, 2019]

Data Injection Attacks against Feedforward Controllers

- Disturbance measurement is corrupted
- A physical disturbance *d* may be present, but is unknown

- Main questions:
  - What attacks are (**un**)**detectable**?

  - What is the **impact** of attacks on the plant measurements and internal states?



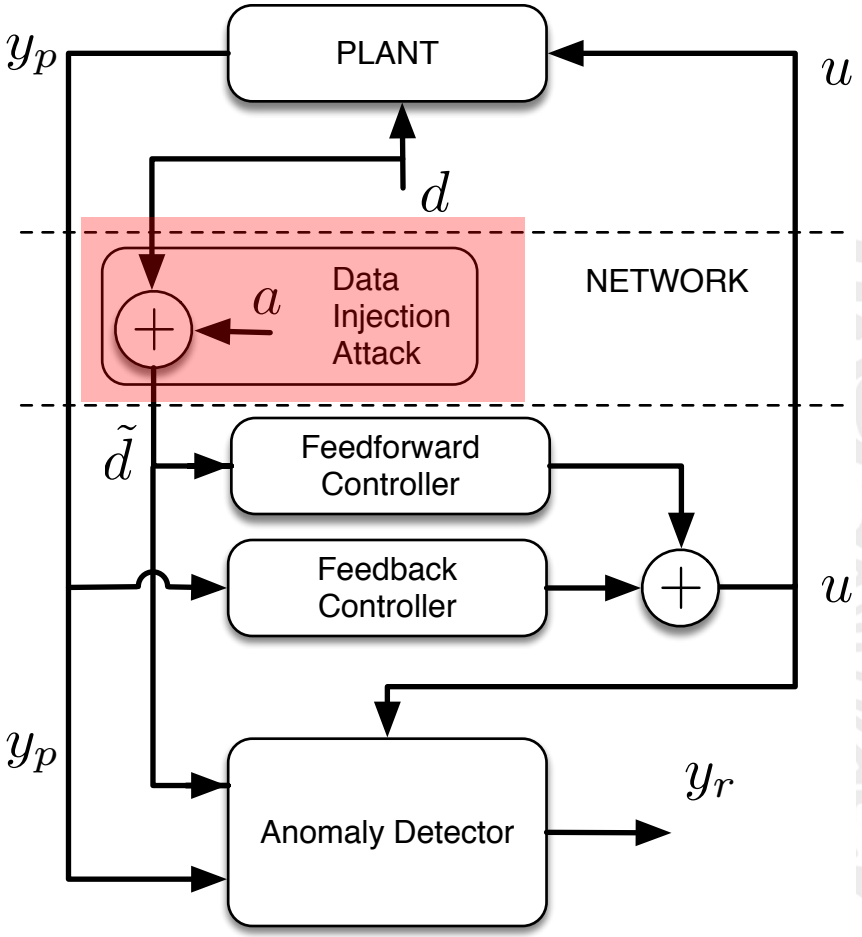
Full closed-loop model

$$\mathcal{P}: \begin{cases} x_p[k+1] = A_p x_p[k] + B_p u[k] + F_p d[k] + \eta[k] \\ y_p[k] = C_p x_p[k] + G_p d[k] + \xi[k] \\ u[k] = u_y[k] + u_d[k] \end{cases}$$

$$\mathcal{F}_d: \begin{cases} x_d[k+1] = A_d x_d[k] + B_d \tilde{d}[k] \\ u_d[k] = C_d x_d[k] + D_d \tilde{d}[k] \end{cases}$$

$$\mathcal{C}: \begin{cases} x_c[k+1] = A_c x_c[k] + B_c y_p[k] \\ u_y[k] = C_c x_c[k] + D_c y_p[k] \end{cases}$$

$$\mathcal{R}: \begin{cases} x_r[k+1] = A_r x_r[k] + B_r u[k] + K_r y_p[k] + F_r \tilde{d}[k] \\ y_r[k] = C_r x_r[k] + D_r u[k] + E_r y_p[k] + G_r \tilde{d}[k] \end{cases}$$



Open-loop model

- Analysis based on the open-loop model (plant + FF controller)

$$\mathcal{P}: \begin{cases} x_p[k+1] = A_p x_p[k] + B_p u[k] + F_p d[k] \\ y_p[k] = C_p x_p[k] + G_p d[k] \\ u[k] = u_y[k] + u_d[k] \end{cases}$$

$$\mathcal{F}_d: \begin{cases} x_d[k+1] = A_d x_d[k] + B_d \tilde{d}[k] \\ u_d[k] = C_d x_d[k] + D_d \tilde{d}[k] \end{cases}$$

$$\begin{aligned} x_{pd}[k+1] &= A_{pd} x_{pd}[k] + B_{pd} \tilde{d}[k] + F_{pd} d[k] \\ y_p[k] &= C_{pd} x_{pd}[k] + G_{pd} d[k], \end{aligned}$$

$$x_{pd}[k] \triangleq \left[x_p^\top[k] \; x_d^\top[k] \right]^\top$$

$$\tilde{d}[k] = d[k] + a[k]$$

- Gives results that hold for any LTI controller & anomaly detector
- It is straightforward to include the controller (by augmenting the plant)

Undetectable attacks

- Definition of undetectable attacks:

A data injection attack on the disturbance measurement, $a[k]$, occurring at $k = 0$ is said to be 0-stealthy with respect to an *arbitrary anomaly detector*, with inputs $u[k]$, $y_p[k]$, and $\tilde{d}[k]$, if there exists a disturbance $d[k]$ and initial conditions $x_p[0]$ and $x_d[0]$ such that $y_p[k] = 0$ and $\tilde{d}[k] = 0$ for all k.

[Pasqualetti et al, TAC, 2013], [Sandberg and Teixeira, SoSCYPS, 2016]

- Intuitively: the attack *mimics* an "*virtual*" disturbance + a transient
- Can be posed as a *output-zeroing problem / zero-dynamics*:

$$x_{pd}[k+1] = A_{pd}x_{pd}[k] + \begin{bmatrix} B_{pd} + F_{pd} & B_{pd} \end{bmatrix} \begin{bmatrix} d^a[k] \\ a[k] \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} y_p[k] \\ \tilde{d}[k] \end{bmatrix} = \begin{bmatrix} C_{pd} \\ 0 \end{bmatrix} x_{pd}[k] + \begin{bmatrix} G_p & 0 \\ I & I \end{bmatrix} \begin{bmatrix} d^a[k] \\ a[k] \end{bmatrix}$$

Feedforward controller - disturbance rejection

- Performance output: $z[k] \triangleq C_z x_p[k] + G_z d[k]$
- Definition of perfect disturbance rejection:

The controller achieves perfect (asymptotic) disturbance rejection with respect to the performance output $z[k]$ if $z[k] = 0 \; \forall k$ ($\lim_{k\to\infty} z[k] = 0$).

- Naturally leads to a characterization based on output-zeroing / zero-dynamics

$$x_{pd}[k+1] = A_{pd} x_{pd}[k] + (B_{pd} + F_{pd}) d[k]$$
$$0 = z[k] = C_{zd} x_{pd}[k] + G_{zd} d[k],$$

UPPSALA UNIVERSITET

Conclusions so far…

Conclusions so far…

- Detectability relates to zero-dynamics from **attack & disturbance** to plant's measurement output

- Disturbance rejection relates to zero-dynamics from **disturbance** to performance output

Conclusions so far…

- Detectability relates to zero-dynamics from **attack & disturbance** to plant's measurement output
- Disturbance rejection relates to zero-dynamics from **disturbance** to performance output
- Zero-dynamics connects detectability with disturbance rejection
- Zero-dynamics can be used to analyze attacks in terms of
  - detectability
  - impact on performance output / measurements (and states)

(Un)Detectability results

- Thm. 1: (known model of plant and feedforward controller)
  - a 0-stealthy attack is an invariant zero of $(A_{pd}, F_{pd}, C_{pd}, G_{pd})$

$$x_{pd}[k+1] = A_{pd}x_{pd}[k] + \begin{bmatrix} B_{pd} + F_{pd} & B_{pd} \end{bmatrix} \begin{bmatrix} d^a[k] \\ a[k] \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} y_p[k] \\ \tilde{d}[k] \end{bmatrix} = \begin{bmatrix} C_{pd} \\ 0 \end{bmatrix} x_{pd}[k] + \begin{bmatrix} G_p & 0 \\ I & I \end{bmatrix} \begin{bmatrix} d^a[k] \\ a[k] \end{bmatrix}$$

- Thm. 2: (known model of plant only)
  - A 0-stealthy attack is an invariant zero of (A_p, F_p, C_p, G_p)
  - a[k] mimics a *virtual* disturbance that results in a zero output signal (i.e., naturally rejected by the open-loop system)

- Results hold for arbitrary LTI controllers & anomaly detector

Impact analysis - the role of the feedforward controller

Impact analysis - the role of the feedforward controller

- Suppose that no physical disturbance is present $(d[k] = 0)$

Impact analysis - the role of the feedforward controller

- Suppose that no physical disturbance is present $(d[k] = 0)$
- FF controller has perfect disturbance rejection w.r.t. $z = y_p$
  - Cor. 2: suppose that a[k] is a **non-vanishing 0-stealthy** attack. Then a[k] mimics a *virtual* disturbance that is perfectly rejected w.r.t. $z = y_p$, and the attack results in a **vanishing** measurement signal.

Impact analysis - the role of the feedforward controller

- Suppose that no physical disturbance is present $(d[k] = 0)$
- FF controller has perfect disturbance rejection w.r.t. $z = y_p$
  - Cor. 2: suppose that a[k] is a **non-vanishing 0-stealthy** attack. Then a[k] mimics a *virtual* disturbance that is perfectly rejected w.r.t. $z = y_p$, and the attack results in a **vanishing** measurement signal.

- FF controller has perfect disturbance rejection w.r.t. '$z \neq y_p$'
  - Cor. 1: suppose that a[k] is a **non-vanishing 0-stealthy** attack. Then a[k] mimics a *virtual* disturbance that results in a **non-vanishing** measurement signal.

Impact analysis - the role of the feedforward controller

- Suppose that no physical disturbance is present $(d[k] = 0)$
- FF controller has perfect disturbance rejection w.r.t. $z = y_p$
  - Cor. 2: suppose that a[k] is a **non-vanishing 0-stealthy** attack. Then a[k] mimics a *virtual* disturbance that is perfectly rejected w.r.t. $z = y_p$, and the attack results in a **vanishing** measurement signal.

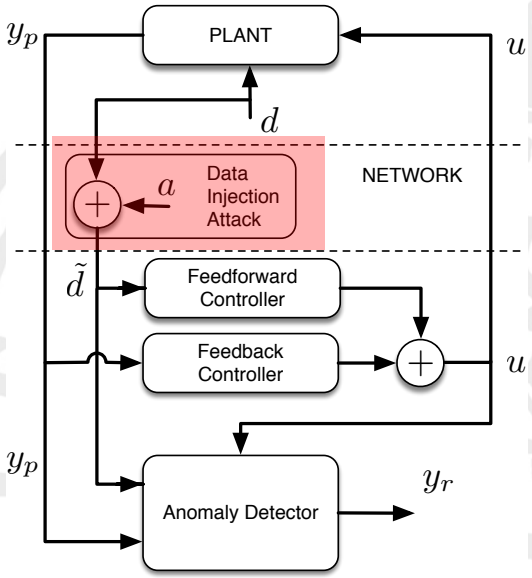
- FF controller has perfect disturbance rejection w.r.t. '$z \neq y_p$'
  - Cor. 1: suppose that a[k] is a **non-vanishing 0-stealthy** attack. Then a[k] mimics a *virtual* disturbance that results in a **non-vanishing** measurement signal.

- In both cases, state estimates will be non-vanishing.
- These results capture the impact on the measurement and state estimates

Numerical Examples

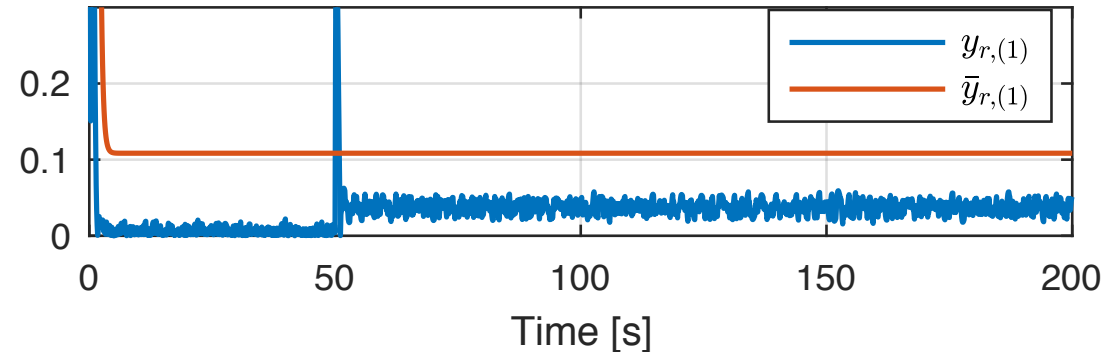
- Unstable plant (2 states, one sensor/actuator) + square-wave disturbance
  - Measurement: $Y_p(z) = G_{yu}(z)U(z) + G_{yd}(z)D(z)$
  - Performance output: $Z(z) = G_{zu}(z)U(z) + G_{zd}(z)D(z)$
- Feedback controller + constant reference
- Anomaly detector with robust threshold

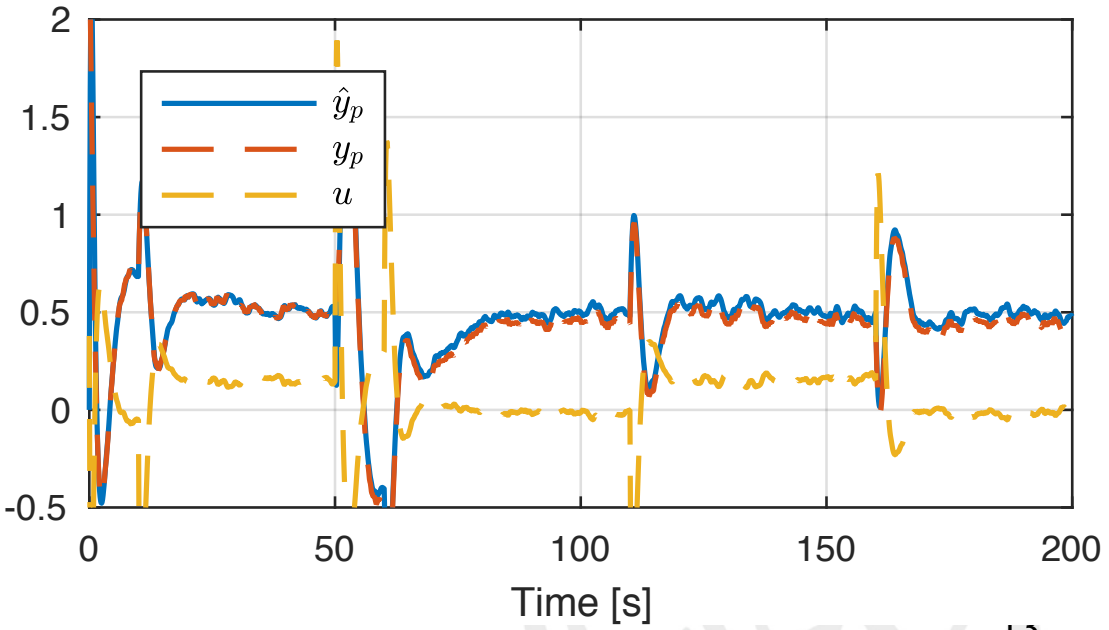
- FF controller: $F_d(z) = -\text{“}G_{zu}^{-1}(z)G_{zd}(z)\text{”}$
  - Case 1: $z = y_p$, detectable attack
  - Case 2: $z = y_p$, undetectable attack
  - Case 3: ‘$z \neq y_p$’, undetectable attack



Detectable attack

- Attack begins at 50s
- Constant attack
$$a[k] = 0.5$$
- Attack detected due to sharp spike





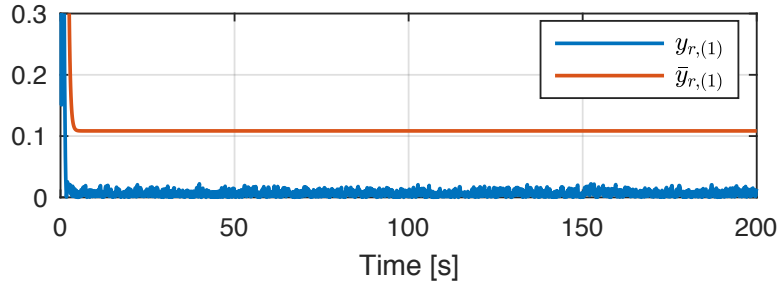
Undetectable attack with $z = y_p$.

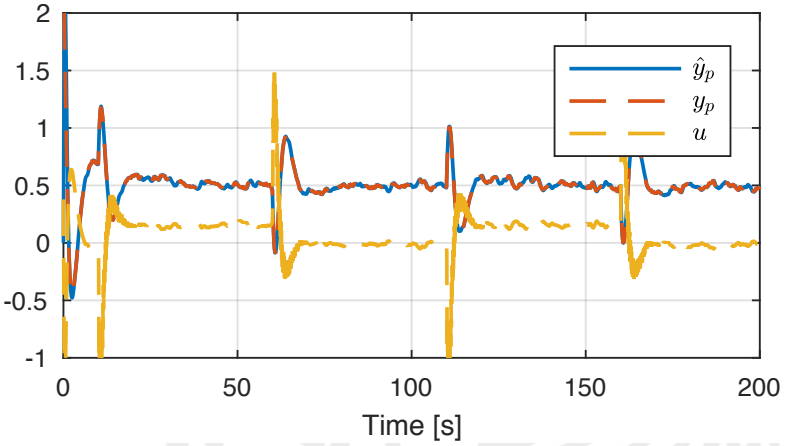
- Attack exploiting an unstable zero of (A_p, F_p, C_p, G_p)

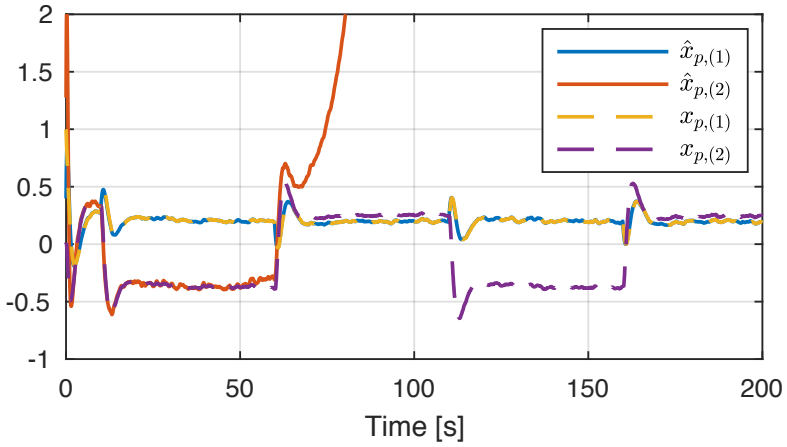
$$a[k] = \lambda^{k-k_a} x_a$$

$$x_a = -0.01,\ \lambda = 1.0141$$

- Attack is not detected
- Vanishing effect on measurement
- Non-vanishing state estimation error







- FF controller is not reacting to the "*virtual*" disturbance on z
  - it is naturally rejected

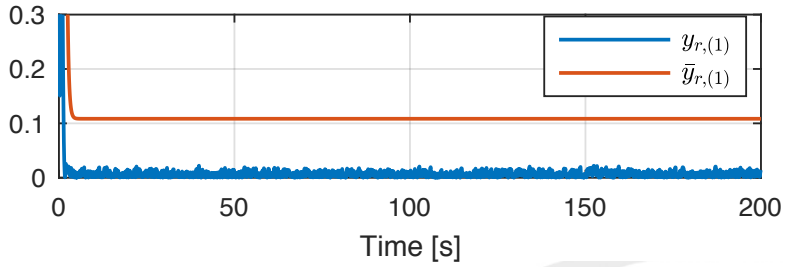
Undetectable attack with $z \neq y_p$

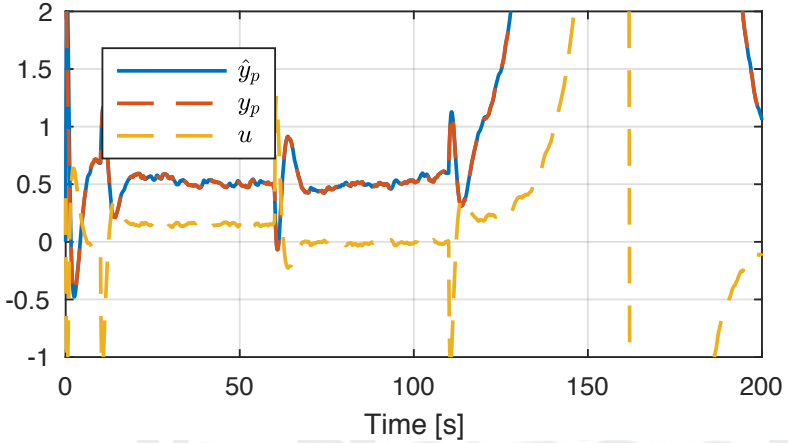
- Attack exploiting an unstable zero of (A_p, F_p, C_p, G_p)

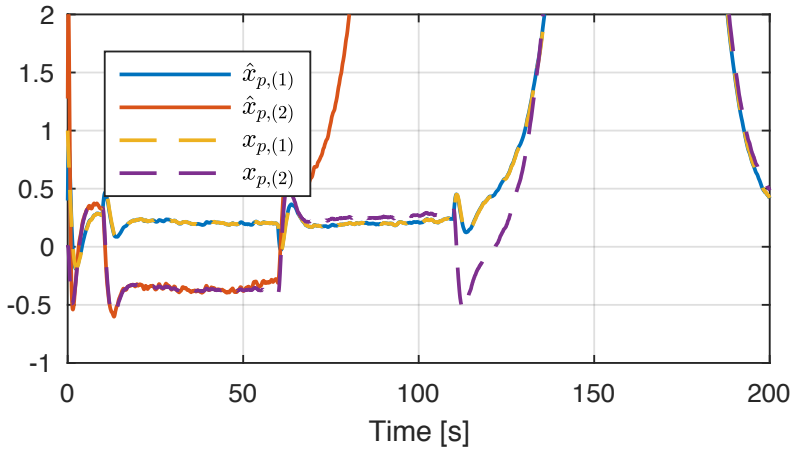
$$a[k] = \lambda^{k-k_a} x_a$$

$$x_a = -0.01,\ \lambda = 1.0141$$

- Attack is not detected
- Non-vanishing effect on measurement and states







- FF controller is compensating the "*virtual*" disturbance on z

Summary and Future Work

- Summary
  - Data injection attacks on disturbance measurements are investigated through analysis of zero-dynamics
  - Undetectable attacks must mimic a virtual disturbance that follows the zero dynamics
  - Impact on plant measurement depends on feedforward controller
  - Estimates of internal states are significantly affected

- Future work
  - Incorporate known disturbance models
  - Investigate the behaviour under specific disturbance rejection strategies
  - Use watermarking strategies to detect attacks

Thank you!

www.andre-teixeira.eu