ROYAL INSTITUTE OF TECHNOLOGY

Quantifying Cyber-Security for Networked Control Systems

André Teixeira, Kin Cheong Sou, **Henrik Sandberg**, Karl H. Johansson

ACCESS Linnaeus Centre,

KTH Royal Institute of Technology

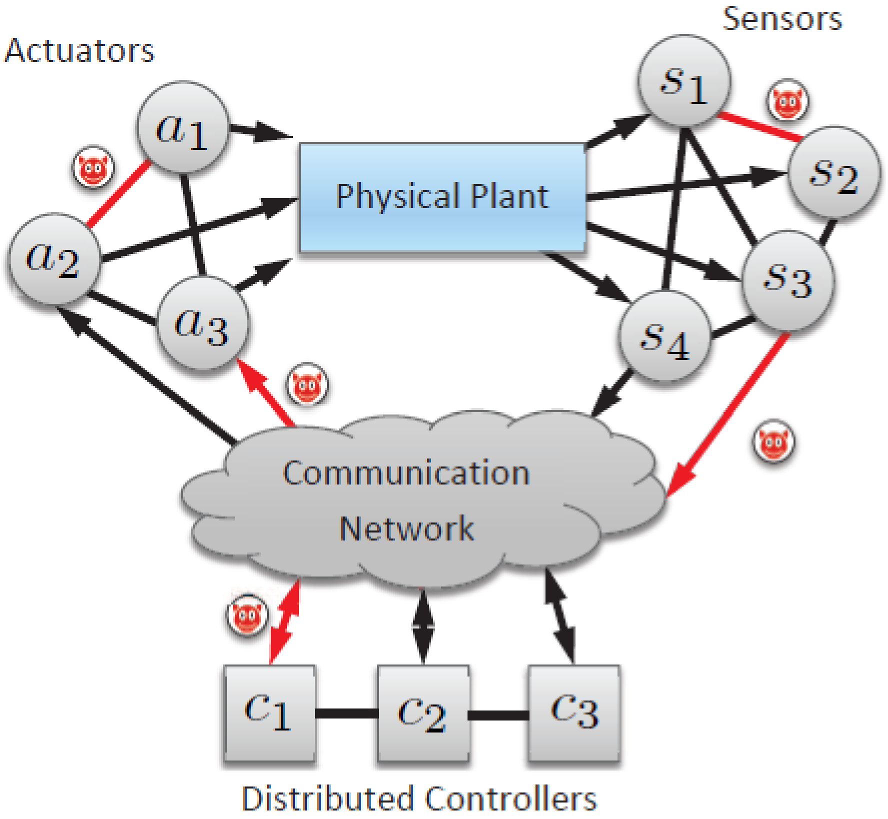
Workshop on Control of Cyber-Physical Systems

Johns Hopkins University

March 20, 2013

Motivation

- Networked control systems are to a growing extent based on commercial off-the-shelf components
- Leads to **increasing vulnerability** to **cyber-physical threats** with many potential points of attacks
- Need for tools and strategies to understand and mitigate attacks in networked control systems
  - Which threats should we care about?
  - What impact can we expect from attacks?
  - Which resources should we protect (more)?



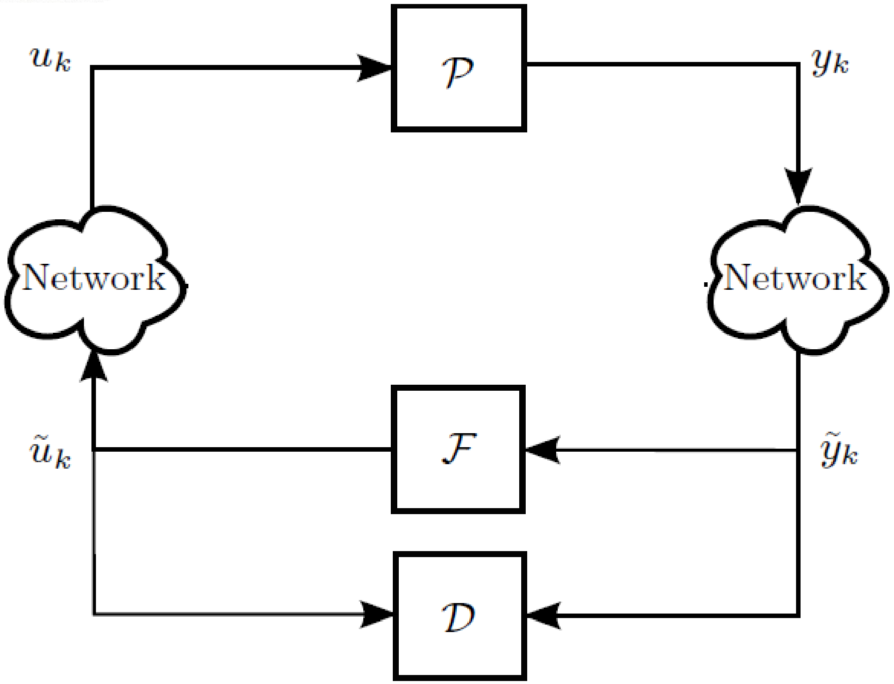
Contributions

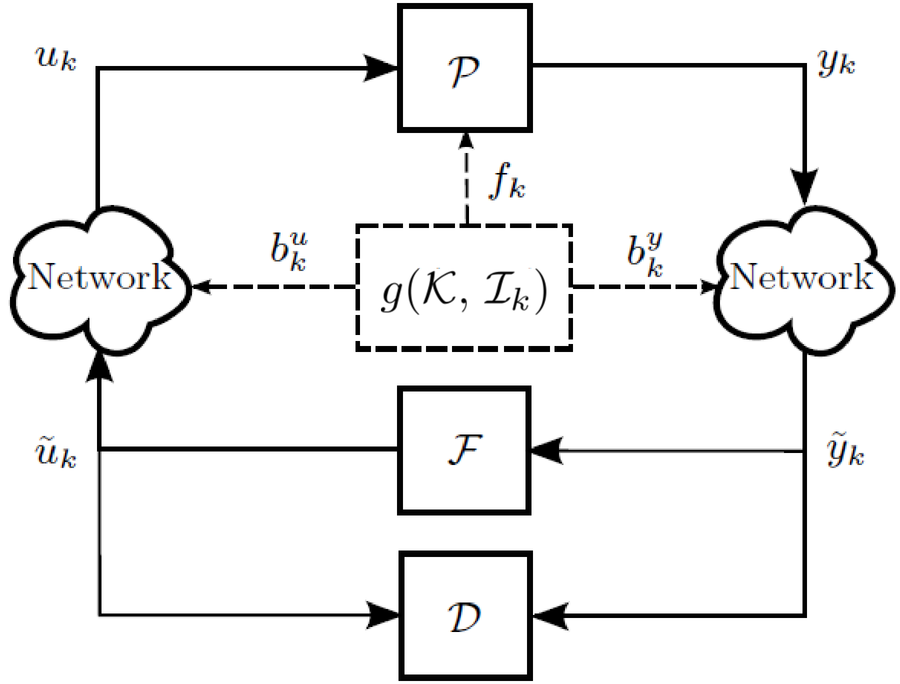
- Tools for quantitative trade-off analysis between attacker's impact and resources
- Extending existing notions for static systems to dynamical systems
- Closed-form solutions and mixed integer linear programming formulations

Outline

- Networked control system and adversary model
- Static and dynamical systems
- Application to the quadruple tank process

Networked Control System under Attack





- Physical plant ($\mathcal{P}$)
- Feedback controller ($\mathcal{F}$)
- Anomaly detector ($\mathcal{D}$)

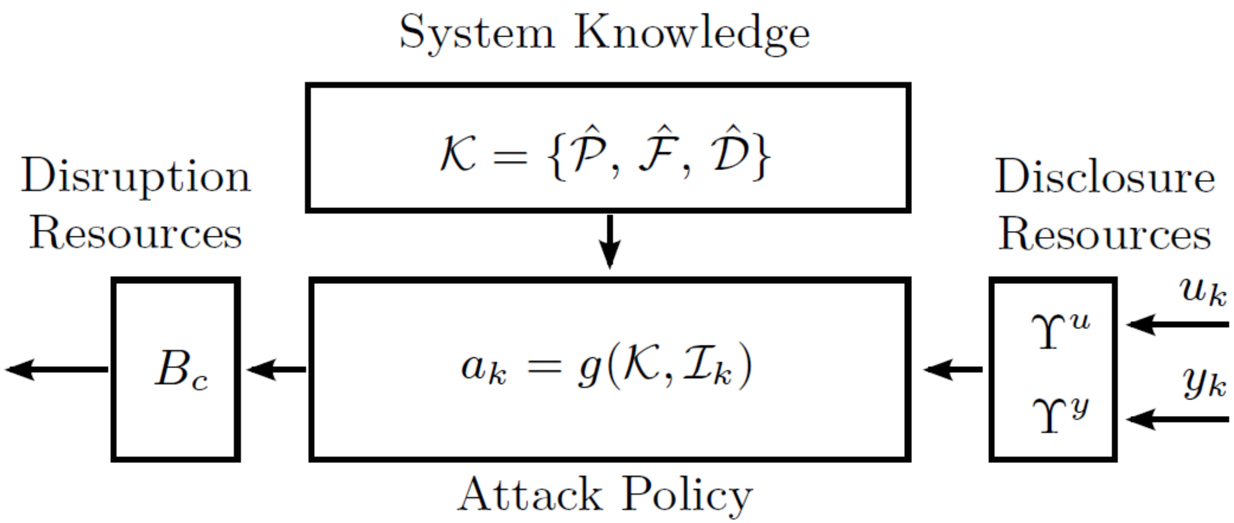
- Disclosure Attacks

- Physical Attacks f_k
- Deception Attacks

$$\tilde{u}_k = u_k + \Gamma^u b_k^u$$

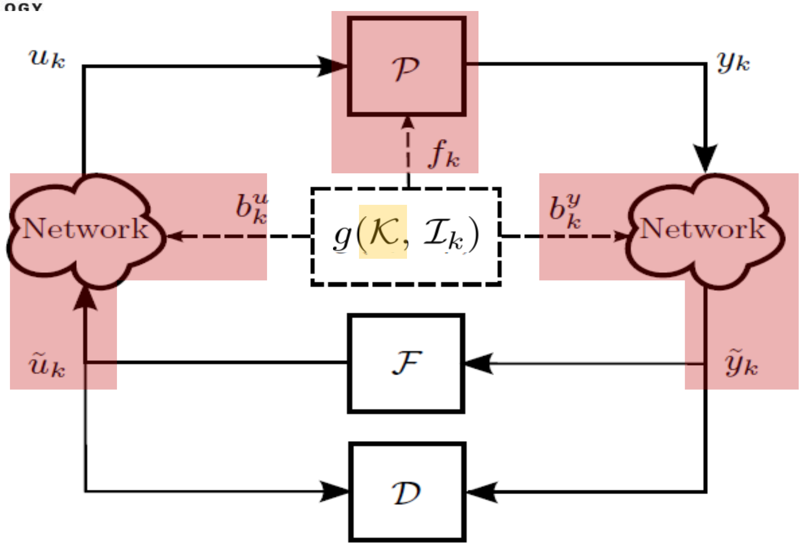
$$\tilde{y}_k = y_k + \Gamma^y b_k^y$$

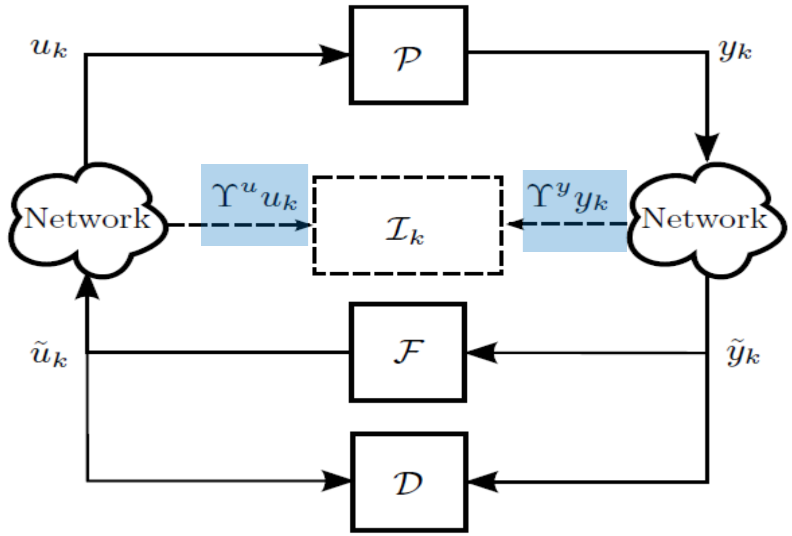
Adversary Model

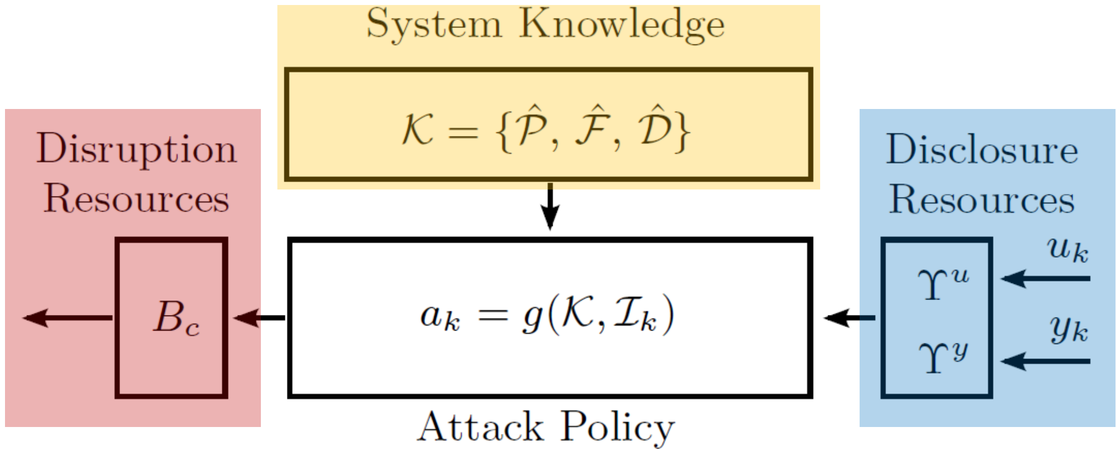


- Adversary's goal is to force the process state into an unsafe region
- Attack should be stealthy, i.e., no alarms
- Adversary constrained by limited resources

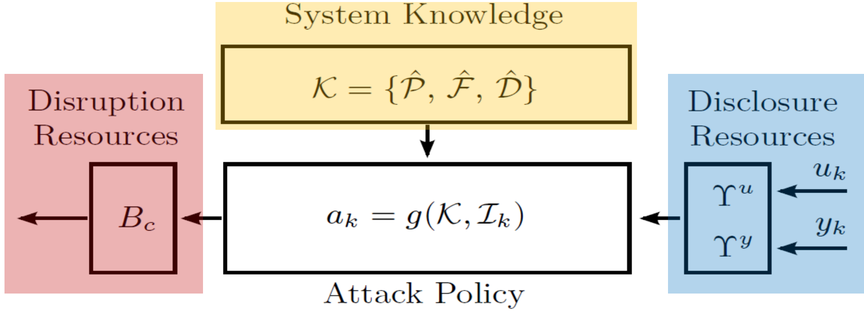
Networked Control System with Adversary Model

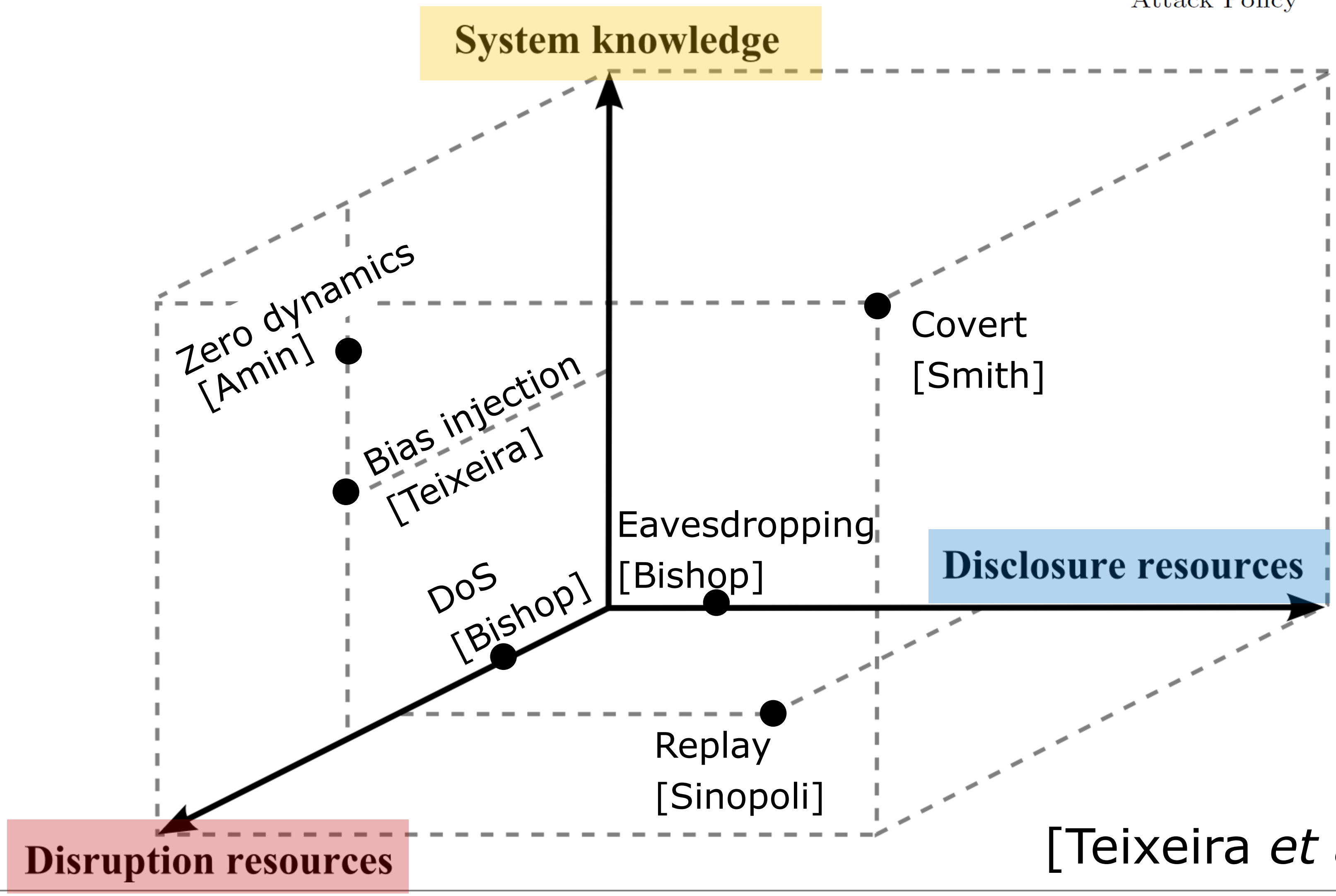






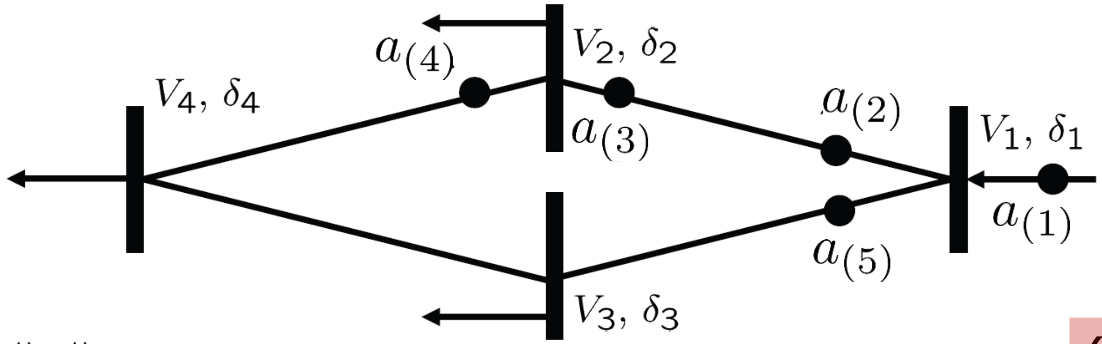
Attack Space





[Teixeira *et al.*, 2012]

Minimum-Resource Attack: The Static Case



$$\min_{\mathbf{a}} \|\mathbf{a}\|_0$$

(disrupt minimum number of channels)

such that

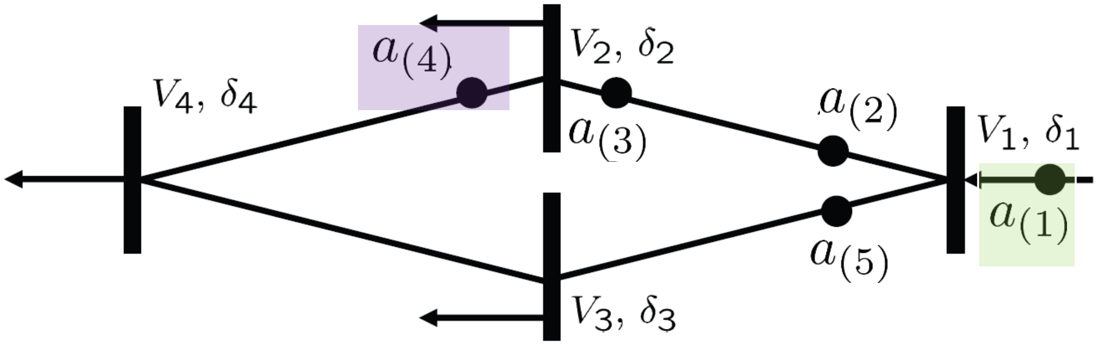
$$\underbrace{\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}}_{\mathbf{r}} = \underbrace{\begin{pmatrix} 0.40 & -0.20 & 0.20 & 0 & -0.40 \\ -0.20 & 0.60 & 0.40 & 0 & 0.20 \\ 0.20 & 0.40 & 0.60 & 0 & -0.20 \\ 0 & 0 & 0 & 0 & 0 \\ -0.40 & 0.20 & -0.20 & 0 & 0.40 \end{pmatrix}}_{\mathcal{T}_r} \underbrace{\begin{pmatrix} a_{(1)} \\ a_{(2)} \\ a_{(3)} \\ a_{(4)} \\ a_{(5)} \end{pmatrix}}_{\mathbf{a}}$$

(no alarms)

$$a_{(k)} = 1, \quad k \in \{1, 2, 3, 4, 5\}$$

(reach attack goals)

Minimum-Resource Attack: The Static Case



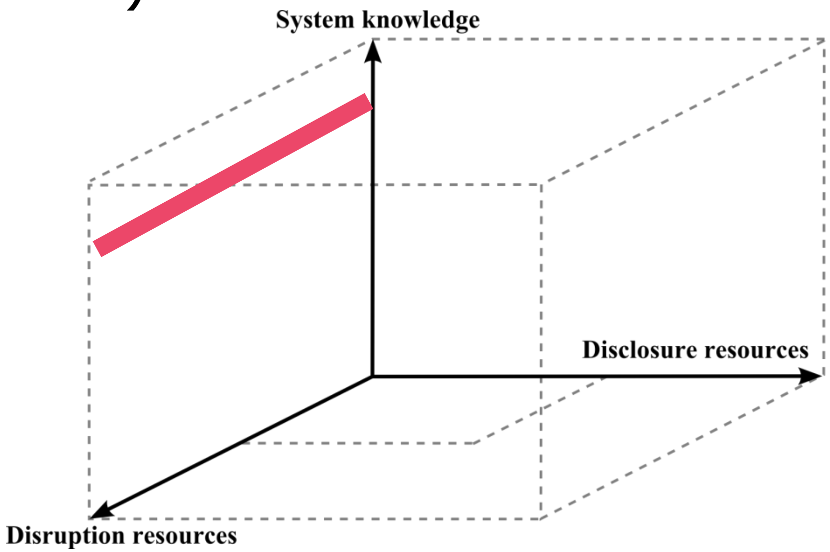
$$\begin{pmatrix}\mathbf{a}_1^\star & \mathbf{a}_2^\star & \mathbf{a}_3^\star & \mathbf{a}_4^\star & \mathbf{a}_5^\star\end{pmatrix} = \begin{pmatrix} 1 & 1 & -1 & 0 & 1 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

[Sandberg *et al*., 2010; Sou *et al*., 2011]

Extensions to Dynamical Systems

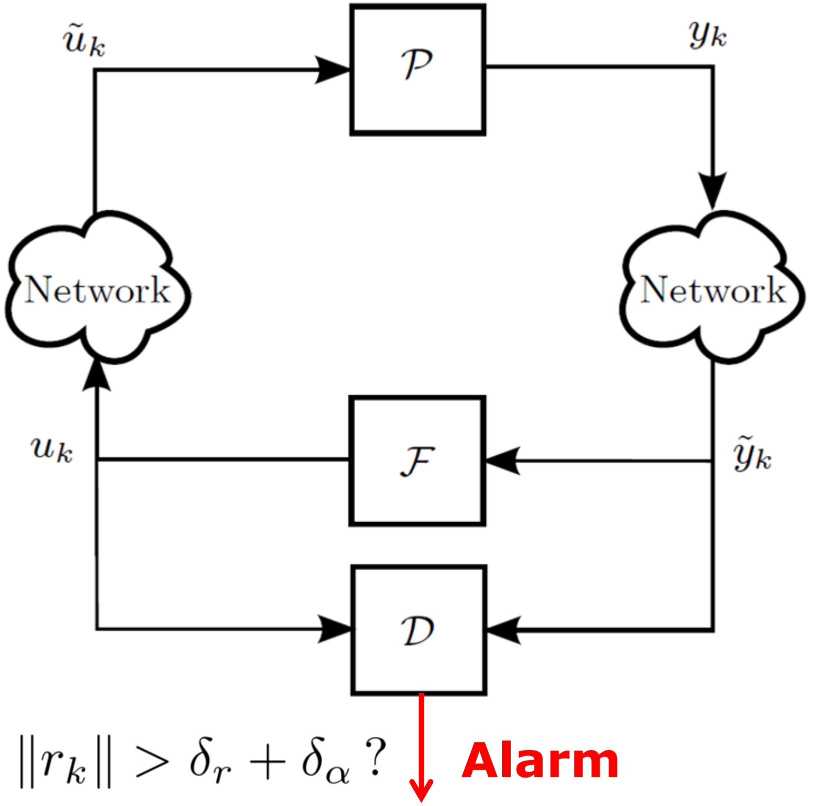
- Attacker needs to satisfy constraints not only across channels (*spatial dimension*) but also constraints across time (*temporal dimension*)

- Cases considered here:
  1. Minimum resource attacks
  2. Maximum impact attacks
  3. Maximum impact bounded resource attacks



- Considered attacks are in *open loop*. No disclosure resources explicitly used (works due to linearity of systems)

Networked Control System



- Physical Plant

$$\mathcal{P} : \begin{cases} x_{k+1} = Ax_k + B\tilde{u}_k + Gw_k \\ y_k = Cx_k + v_k \end{cases}$$

- Feedback Controller

$$\mathcal{F} : \begin{cases} z_{k+1} = A_c z_k + B_c \tilde{y}_k \\ u_k = C_c z_k + D_c \tilde{y}_k \end{cases}$$

- Anomaly Detector

$$\mathcal{D} : \begin{cases} \hat{x}_{k|k} = A\hat{x}_{k-1|k-1} + Bu_{k-1} + K(\tilde{y}_k - \hat{y}_{k|k-1}) \\ r_k = V(\tilde{y}_k - \hat{y}_{k|k}) \end{cases}$$

  - Alarm triggered if $\|r_k\| > \delta_r + \delta_\alpha$

1. Minimum Resource Attack: Dynamical Case

Dynamical anomaly detector for closed-loop system:

$$\begin{aligned}\xi_{k+1} &= \mathbf{A_e}\xi_k + \mathbf{B_e}a_k + \mathbf{G_e}w_k \\ r_k &= \mathbf{C_e}\xi_k + \mathbf{D_e}a_k + \mathbf{H_e}v_k\end{aligned}$$

Lift to time interval $[0, N]$
with zero-initial conditions and no noise:

$$\underbrace{\begin{bmatrix} r_0 \\ r_1 \\ r_2 \\ \vdots \\ r_N \end{bmatrix}}_{\mathbf{r}} = \underbrace{\begin{bmatrix} \mathbf{D_e} & 0 & \dots & 0 \\ \mathbf{C_e}\mathbf{B_e} & \mathbf{D_e} & \dots & 0 \\ \mathbf{C_e}\mathbf{A_e}\mathbf{B_e} & \mathbf{C_e}\mathbf{B_e} & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ \mathbf{C_e}\mathbf{A_e}^{N-1}\mathbf{B_e} & \mathbf{C_e}\mathbf{A_e}^{N-2}\mathbf{B_e} & \dots & \mathbf{D_e} \end{bmatrix}}_{\mathcal{T}_r} \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix}}_{\mathbf{a}}$$

1. Minimum Resource Attack: Dynamical Case

Signal strength channel i

$$\min_{\mathbf{a}} \|h_p(\mathbf{a})\|_0$$

such that

$$h_p(\mathbf{a}) = [\|\mathbf{a}_{(1)}\|_{\ell_p}, \ldots, \|\mathbf{a}_{(i)}\|_{\ell_p}, \ldots, \|\mathbf{a}_{(q_a)}\|_{\ell_p}]$$

$$\|\mathbf{r}\|_q = \|\mathcal{T}_r \mathbf{a}\|_q \leq \delta_\alpha$$

$$\mathbf{a} \in \mathcal{G}$$

- Minimize disruption resources (#channels attacked)
- No alarms (threshold δ_α)
- Reach attack goals $\mathcal{G}$ (compare static case)

A Problem with 1-Norm Relaxation

$$\xcancel{\min_{\mathbf{a}} \|h_p(\mathbf{a})\|_1 = \|\mathbf{a}_{(1)}\|_{\ell_p} + \cdots + \|\mathbf{a}_{(i)}\|_{\ell_p} + \cdots + \|\mathbf{a}_{(q_a)}\|_{\ell_p}}$$

such that

$$\|\mathbf{r}\|_q = \|\mathcal{T}_r \mathbf{a}\|_q \leq \delta_\alpha$$

$$\mathbf{a} \in \mathcal{G}$$

- Mixes the temporal and spatial dimensions!
- Attacking a new channel should be more expensive than accessing an already attacked channel over again
- Compare with [Fawzi *et al.*, 2012]

Formulate as MILP Instead

Note that

$$\|h_p(\mathbf{a})\|_0 \le \epsilon$$

can equivalently be formulated as

$$\begin{aligned} \mathbf{a}_{(i)} &\le M_h \mathbf{z}_i \mathbf{1} && \forall\, i = 1, \ldots,\ q_a \\ -\mathbf{a}_{(i)} &\le M_h \mathbf{z}_i \mathbf{1} && \forall\, i = 1, \ldots,\ q_a \\ \sum_{i=1}^{q_a} \mathbf{z}_i &\le \epsilon \\ \mathbf{z}_i &\in \{0, 1\} && \forall\, i = 1, \ldots,\ q_a. \end{aligned}$$

where M_h is a large constant ("infinity")

ROYAL INSTITUTE OF TECHNOLOGY

1. Minimum Resource Attack: Dynamical Case

$$\min_{\mathbf{a},\, \epsilon} \epsilon$$

such that

$$h_p(\mathbf{a}) = [\|\mathbf{a}_{(1)}\|_{\ell_p}, \ldots, \|\mathbf{a}_{(i)}\|_{\ell_p}, \ldots, \|\mathbf{a}_{(q_a)}\|_{\ell_p}]$$

$$\|h_p(\mathbf{a})\|_0 \leq \epsilon$$

$$\|\mathbf{r}\|_q = \|\mathcal{T}_r \mathbf{a}\|_q \leq \delta_\alpha$$

$$\mathbf{a} \in \mathcal{G}$$

- Minimize disruption resources (#channels attacked)
- No alarms (threshold δ_α)
- Reach attack goals $\mathcal{G}$ (compare static case)

- MILP if $p = q = \infty$

2. Maximum Impact Attack: Dynamical Case

Dynamics of plant and controller:

$$\eta_{k+1} = \mathbf{A}\eta_k + \mathbf{B}a_k + \mathbf{G}w_k$$
$$x_k = \mathbf{C}\eta_k + \mathbf{D}a_k + \mathbf{H}v_k$$

Lifting to time interval $[0, N]$
with zero-initial conditions and no noise:

$$\underbrace{\begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}}_{\mathbf{x}} = \underbrace{\begin{bmatrix} \mathbf{D} & 0 & \dots & 0 \\ \mathbf{CB} & \mathbf{D} & \dots & 0 \\ \mathbf{CAB} & \mathbf{CB} & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ \mathbf{CA}^{N-1}\mathbf{B} & \mathbf{CA}^{N-2}\mathbf{B} & \dots & \mathbf{D} \end{bmatrix}}_{\mathcal{T}_x} \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix}}_{\mathbf{a}}$$

2. Maximum Impact Attack: Dynamical Case

$$\max_{\mathbf{a}} \|\mathcal{T}_x \mathbf{a}\|_p$$

such that

$$\|\mathbf{r}\|_q = \|\mathcal{T}_r \mathbf{a}\|_q \leq \delta_\alpha$$

- Maximize impact (push $\|\mathbf{x}\|_p$ far away from equilibrium)
- No alarms (threshold δ_α)

- Not a convex optimization problem!
- Closed-form solution when $p = q = 2$ (use Courant-Fischer)

2. Maximum Impact Attack: Dynamical Case

$$\max_{\mathbf{a}} \|\mathcal{T}_x \mathbf{a}\|_p$$

such that

$$\|\mathbf{r}\|_q = \|\mathcal{T}_r \mathbf{a}\|_q \leq \delta_\alpha$$

Theorem: Bounded solution iff $\ker(\mathcal{T}_r) \subseteq \ker(\mathcal{T}_x)$

Theorem ($p = q = 2$): Assume bounded solution, then

$$\mathbf{a}^\star = \frac{\delta_\alpha}{\|\mathcal{T}_r \mathbf{v}_{\max}\|_\mathbf{2}} \mathbf{v}_{\max}, \qquad \|\mathcal{T}_\mathbf{x} \mathbf{a}^\star\|_\mathbf{2} = \sqrt{\lambda_{\max}} \delta_\alpha$$

$0 = (\lambda_{\max} \mathcal{T}_r^\top \mathcal{T}_r - \mathcal{T}_x^\top \mathcal{T}_x) \mathbf{v}_{\max}$ ($\lambda_{\max} / \mathbf{v}_{\max}$ max generalized eigenpair)

ROYAL INSTITUTE OF TECHNOLOGY

3. Maximum Impact Bounded Resource Attack

$$\max_{\mathbf{a}} \|\mathcal{T}_x \mathbf{a}\|_p$$

such that

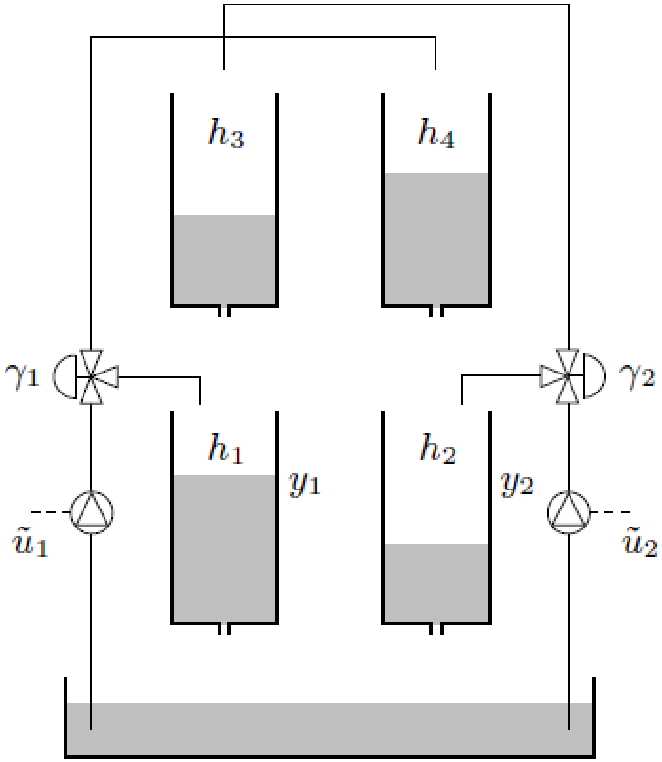
$$\|\mathbf{r}\|_q = \|\mathcal{T}_r \mathbf{a}\|_q \leq \delta_\alpha$$

$$\|h_p(\mathbf{a})\|_0 \leq \epsilon$$

- Maximize impact (push $\|\mathbf{x}\|_p$ far away from equilibrium)
- No alarms (threshold δ_α)
- Use no more than ϵ channels

- $p = q = \infty$ yields Mixed Integer Linear Program (MILP)

Numerical Example



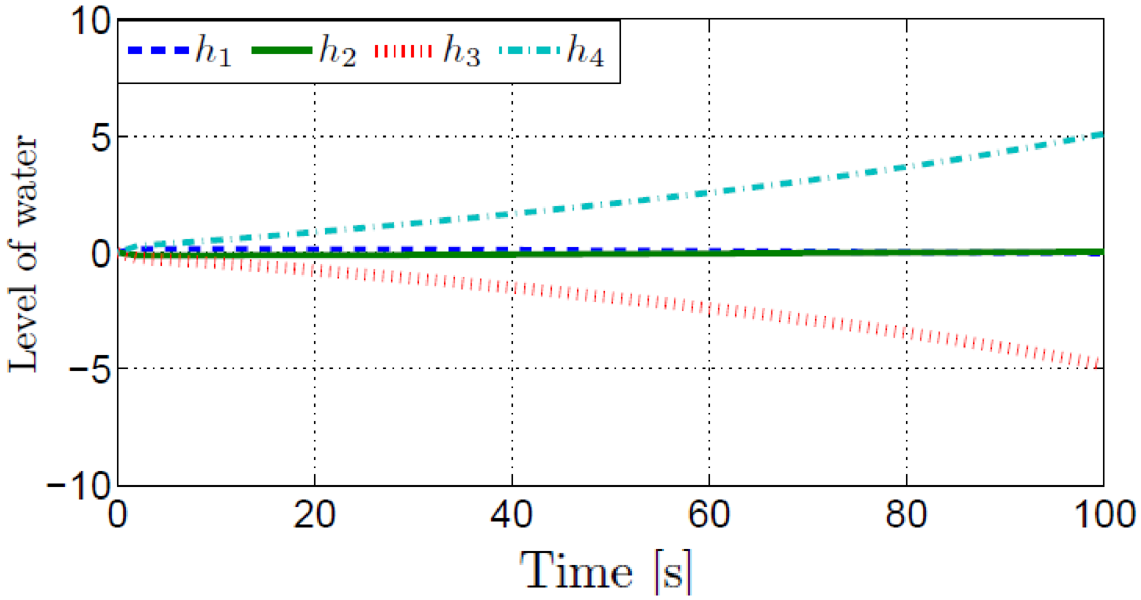
$$
\begin{aligned}
\dot{h}_1 &= -\frac{a_1}{A_1}\sqrt{2gh_1} + \frac{a_3}{A_1}\sqrt{2gh_3} + \frac{\gamma_1 k_1}{A_1}u_1,\\
\dot{h}_2 &= -\frac{a_2}{A_2}\sqrt{2gh_2} + \frac{a_4}{A_2}\sqrt{2gh_4} + \frac{\gamma_2 k_2}{A_2}u_2,\\
\dot{h}_3 &= -\frac{a_3}{A_3}\sqrt{2gh_3} + \frac{(1-\gamma_2)k_2}{A_3}u_2,\\
\dot{h}_4 &= -\frac{a_4}{A_4}\sqrt{2gh_4} + \frac{(1-\gamma_1)k_1}{A_4}u_1,
\end{aligned}
$$

- Wireless LQG controller
- 4 channels: 2 actuators and 2 measurements
- Minimum phase or non-minimum phase depending on γ_1, γ_2

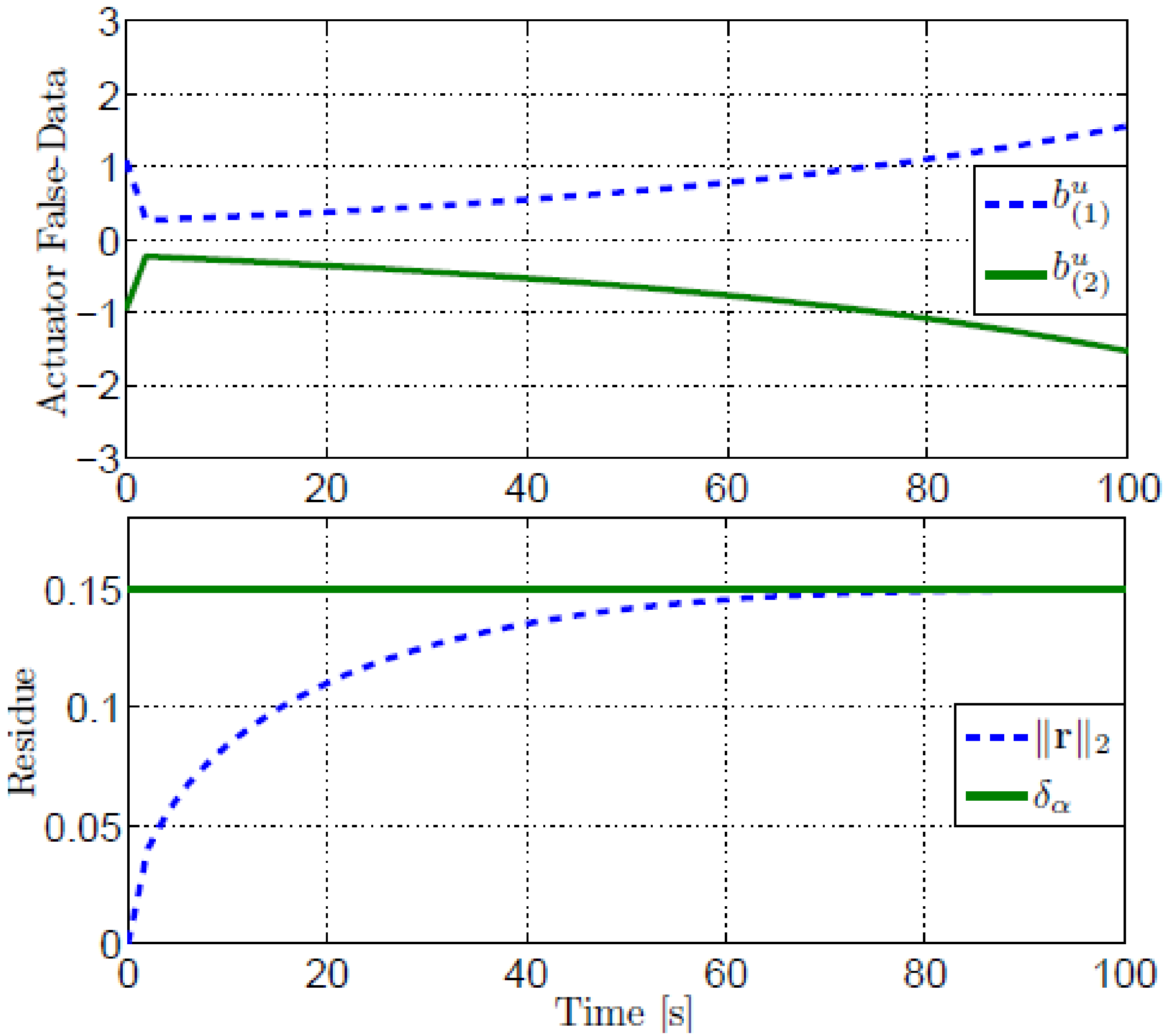
Numerical Example (Non-Min Phase)

Values of $\|\mathbf{x}\|_p$ for maximum impact formulation with $p = q = 2,\ \delta_\alpha = 0.15$

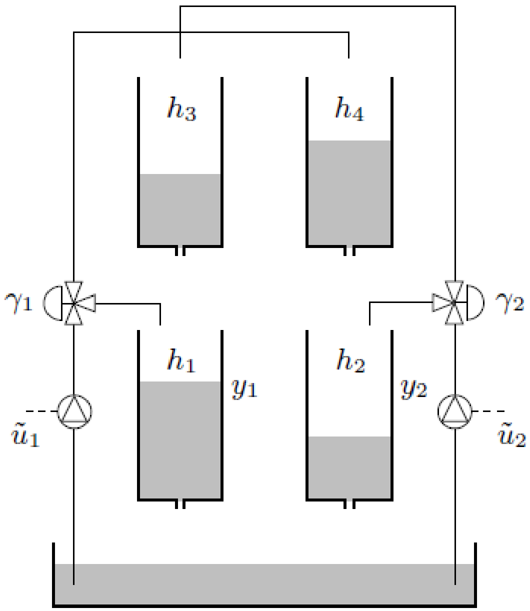
	$\|h_p(\mathbf{a})\|_0$			
	1	**2**	**3**	**4**
Minimum phase	1.15	140.39	∞	∞
Non-minimum phase	2.80	689.43	∞	∞

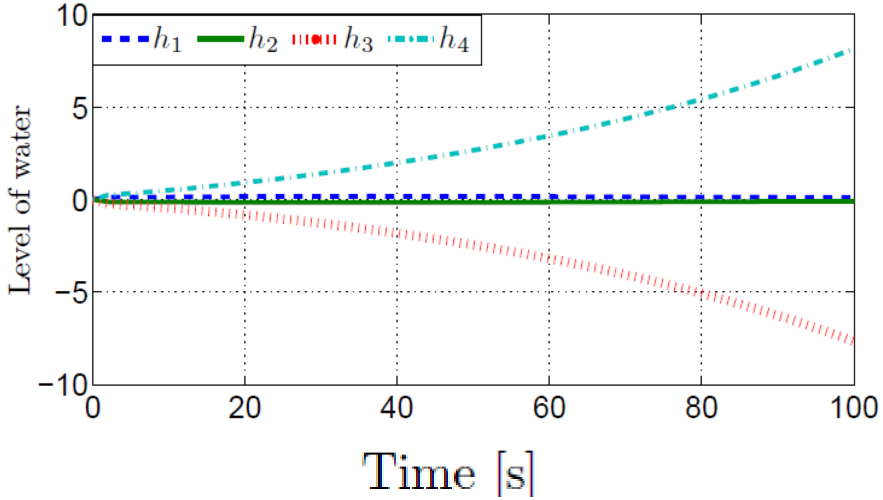


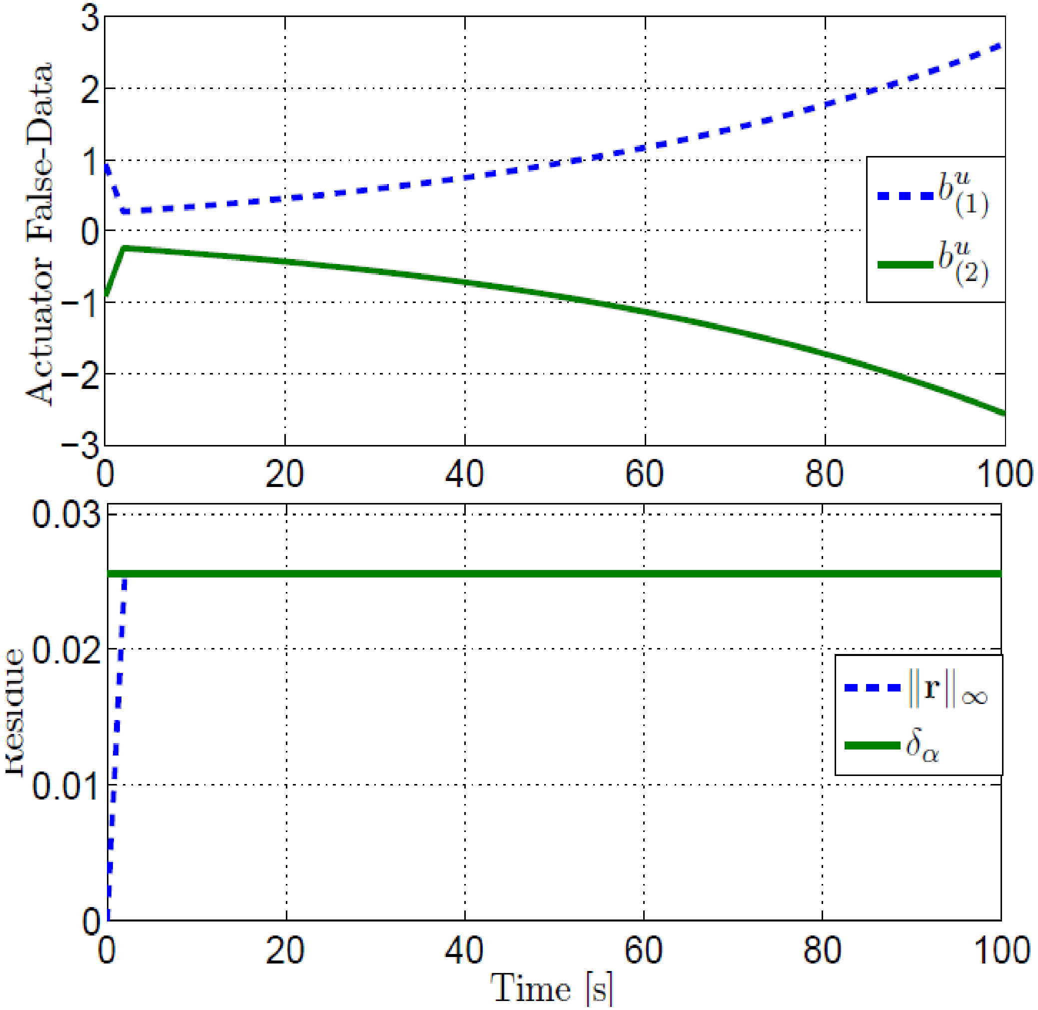
Numerical Example (Non-Min Phase)



Numerical Example (MILP)







Numerical Example

- Maximum Impact Bounded Resource attack illustrated
- 2 channels allowed: MILP selects the actuators
- 3-4 channels allowed: Unbounded impact (any attack on actuators can be hidden by corrupting 2 measurements)
- Infinity norm criteria ($p = q = \infty$) yields more aggressive attack (bounds saturated)
- Not surprisingly, non-min phase plant more sensitive

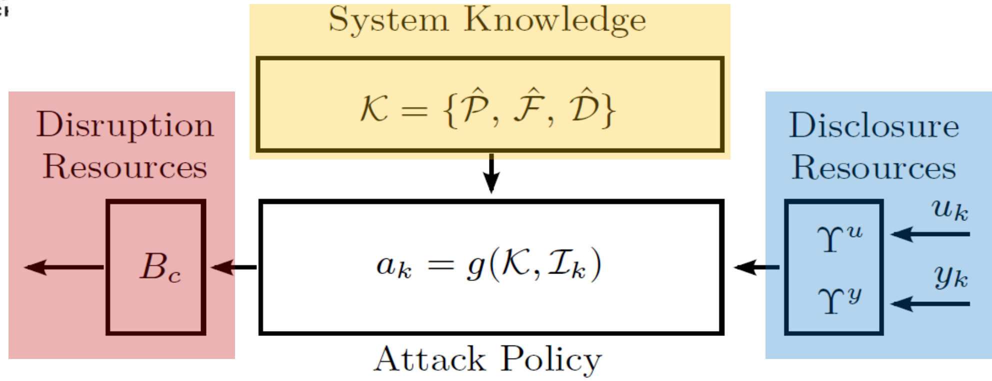
Steady-State Attacks

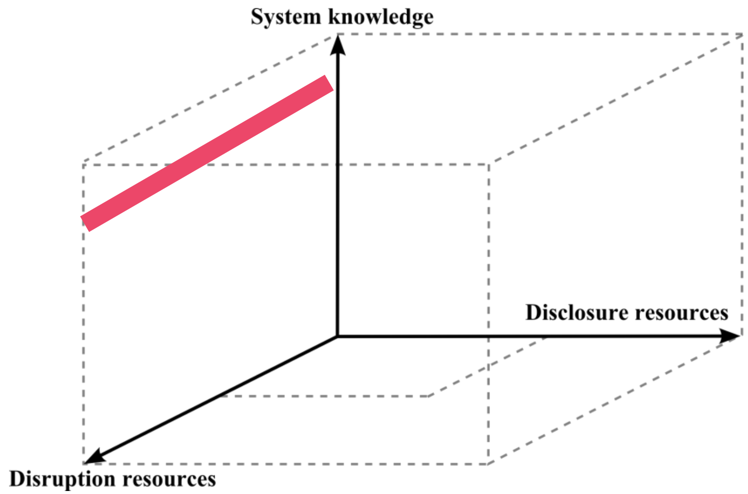
- Consider attacks over $[0, N]$ where
  - $N \to \infty$
  - $a_k = g e^{i\omega k}, \quad \omega \in \mathbb{R},\ g \in \mathbb{C}^{q_a}$ (sinusoidal attacks)

- Similar analysis carries through but make substitutions
  - $\mathcal{T}_r \to G_r(e^{i\omega})$
  - $\mathcal{T}_x \to G_x(e^{i\omega})$

- Yields worst-case attack frequency ω etc. Details in paper

Summary



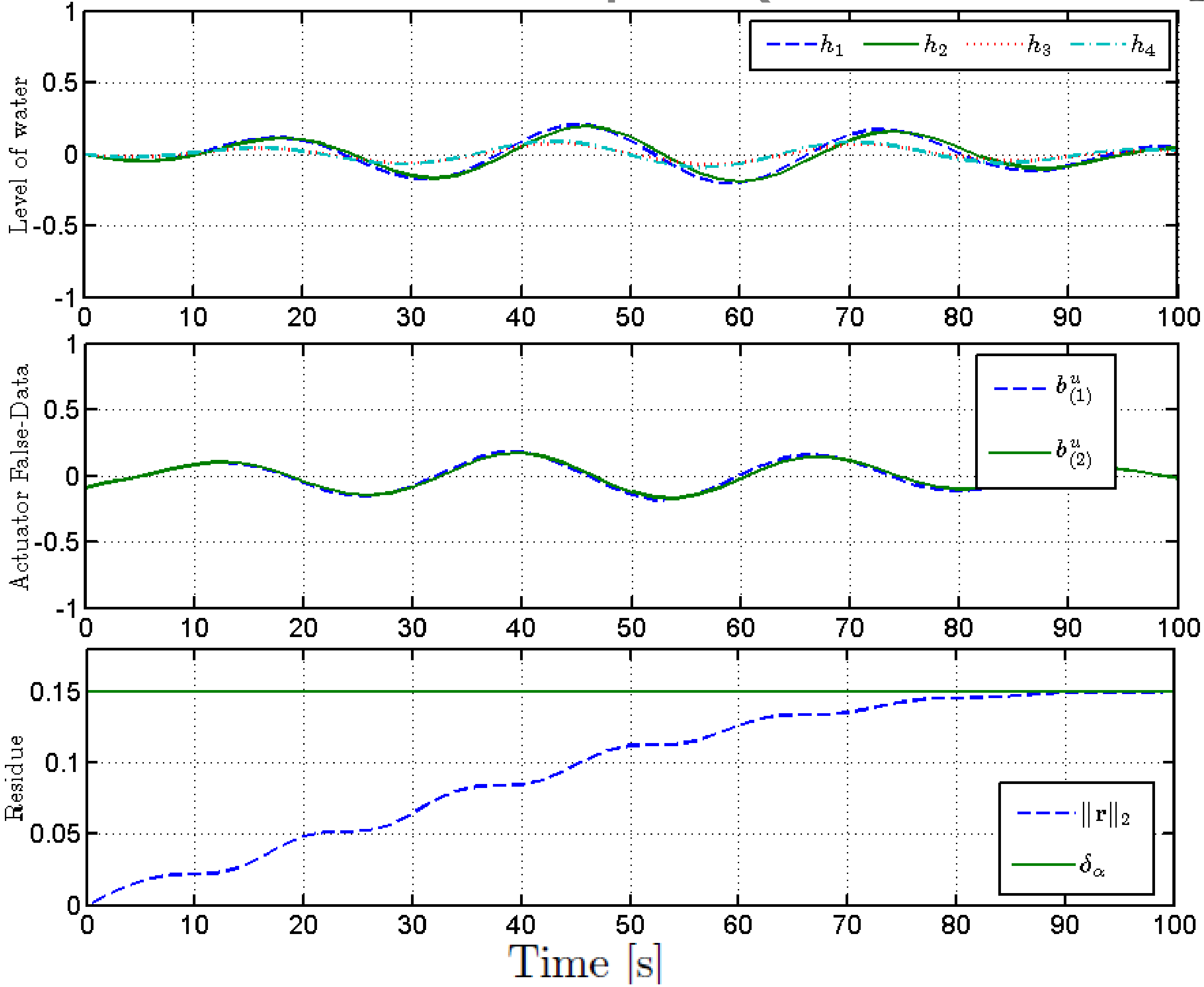


- Tools for quantitative trade-off analysis between attacker’s impact and resources: Important for defense prioritization

- For dynamical systems there are *temporal* as well as *spatial (channel) constraints* for attacker to fulfill
  - Enforced through lifting and frequency-response models

- Closed-form solutions and mixed integer linear programming formulations

Numerical Example (Min Phase $p = q = 2$)



Numerical Example (Min Phase $p = q = \infty$)