Toward AI-enhanced Design of Resilient Cyber-Physical Systems: a Journey from Inception to Present Times

Bruno Sinopoli
Das Family Distinguished Professor and Chair
Department of ESE
Washington University
Saint Louis, MO

CPS week
May 20, 2021
Nashville, TN

Acknowledge my students and postdocs

- Yilin Mo, *faculty at Tsinghua University*
- Sean Weerakkody, *Applied Physics Labs, JHU*
- Xiaofei Liu, *LinkedIn*
- Nicola Forti, NATO *Centre for Maritime Research and Experimentation (CMRE), La Spezia, Italy*
- Walter Lucia, *faculty at Concordia University, Montreal*
- Omur Ozel, *faculty at George Washington University*
- Raffaele Romagnoli*, Postdoc, Carnegie Mellon University*
- Paul Griffioen*, Ph.D. student, Carnegie Mellon University*
- Carmel Fiscko*, Ph.D. student, Carnegie Mellon University*
- Rohan Chabukshwar, *UTRC, Ireland*
- Mehdi Hosseinzadeh, *Postdoc, Washington U. in STL*
- Bahram Yaghooti, *Ph.D. Student, Washington U. in STL*
- Jonathan Gornet, *Ph.D. Student, Washington U. in STL*

And many many collaborators around the world…

Once upon a time, when Math reigned…

Beautiful theories were developed…

$$\int_0^\infty \ln|S(j\omega)|d\omega = \int_0^\infty \ln\left|\frac{1}{1+L(j\omega)}\right|d\omega = \pi\sum Re(p_k) - \frac{\pi}{2}\lim_{s\to\infty} sL(s)$$

4. EV normieren: u=r/|r| bzw. Gram-Schmid
5. Matrix U bilden: $U=[u_1\ u_2\ \dots]$
6. Singular values σ_i = sqrt(Eigenwerte)

$A=[U_1|U_2]\begin{bmatrix}\Sigma r & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}V_1\\ V_2\end{bmatrix}$ and $x_{ln,ls}=V_1\Sigma_r^{-1}U_1^T y$

initial condition
► iff $A^T P + PA = -Q$ with P, Q positive defi
► Asymptotic S. implies exponential stability
► observability Gramian converges to the uni

$$C = \sup_{p_X} I(X;Y)$$

$$f_p \le 2B$$

$$R(p_b) = \frac{C}{1-H_2(p_b)}$$

Boolean Expression	Description	Equivalent Switching Circuit	Boolean Algebra Law or Rule
A + 1 = 1	A in parallel with closed = "CLOSED"		Annulment
A + 0 = A	A in parallel with open = "A"		Identity
A . 1 = A	A in series with closed = "A"		Identity
A . 0 = 0	A in series with open = "OPEN"		Annulment
A + A = A	A in parallel with A = "A"		Idempotent
A . A = A	A in series with A = "A"		Idempotent
NOT $\bar{A}$ = A	NOT NOT A (double negative) = "A"		Double Negation
A + $\bar{A}$ = 1	A in parallel with NOT A = "CLOSED"		Complement
A . $\bar{A}$ = 0	A in series with NOT A = "OPEN"		Complement
A+B = B+A	A in parallel with B = B in parallel with A		Commutative
A.B = B.A	A in series with B = B in series with A		Commutative
$\overline{A+B} = \bar{A}.\bar{B}$	invert and replace OR with AND		de Morgan's Theorem
$\overline{A.B} = \bar{A}+\bar{B}$	invert and replace AND with OR		de Morgan's Theorem

rix with Jordan diagonal, N-matrix

and $e^{Nt} = I + \frac{1}{1!}Nt + \frac{1}{2!}N^2t^2 \dots$

e^At=matrixofeigenvectors * erse(matrixofeigenvectors)

of parallelpiped = det(e^{At})=$e^{tr(A)t}$. Can or grow (e^t)

Controllability: x' = Ax + Bu y=Cx
controllability Matrix C: [B AB A²B
controllable if det(C) ≠0 ≙ rank(C)
controllable if $G_C(t_1)>0 \leftrightarrow det$ (

where $G_C(t_1)=\int_0^{t1} e^{A\tau}BB^Te^{A^T\tau}$

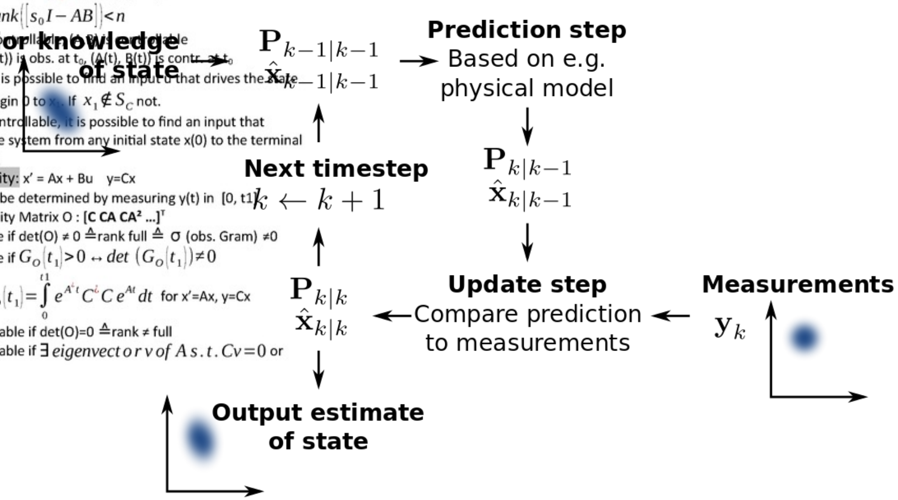
not controllable if det(C)=0 ≙ rank(C) ≠full
not controllable if $\exists w\ s.t.\ wB=0$ or
$\exists s_0\ s.t.\ rank([s_0I-AB])<n$

If (A,B) is controllable, it is possible to find an input that transfers the system from any initial state x(0) to the terminal state x_1 at t_1

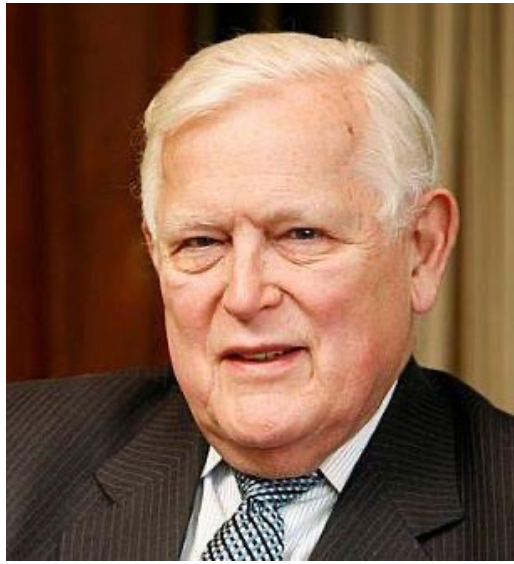
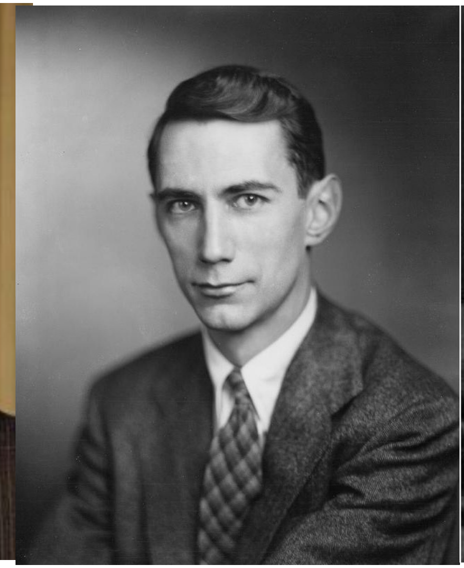
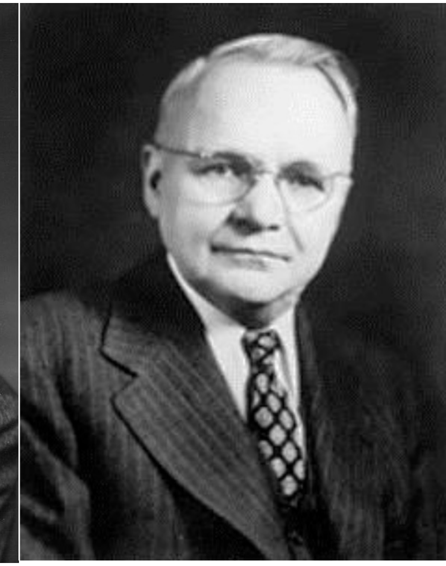
Observability: x' = Ax + Bu y=Cx
↔ x(0) can be determined by measuring y(t) in [0, t1]
observability Matrix O : [C CA CA² ...]^T
observable if det(O) ≠ 0 ≙rank full ≙ σ (obs. Gram) ≠0
observable if $G_O(t_1)>0 \leftrightarrow det(G_O(t_1))\neq 0$

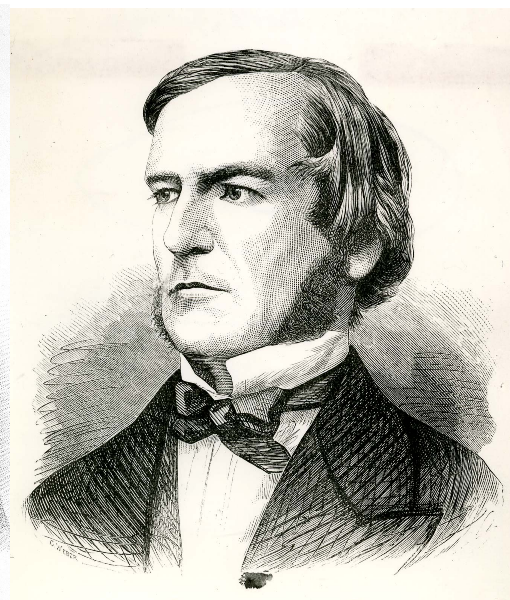
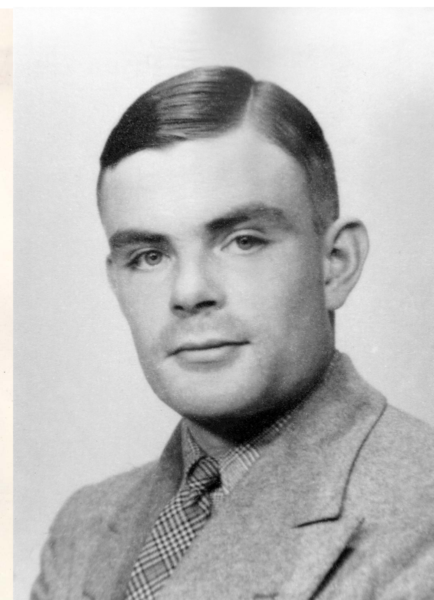
where $G_O(t_1)=\int_0^{t1} e^{A^T\tau}C^TCe^{At}dt$ for x'=Ax, y=Cx

not observable if det(O)=0 ≙rank ≠ full
not observable if ∃ eigenvector v of A s.t. Cv=0 or



by eminent minds


Some connections were made…

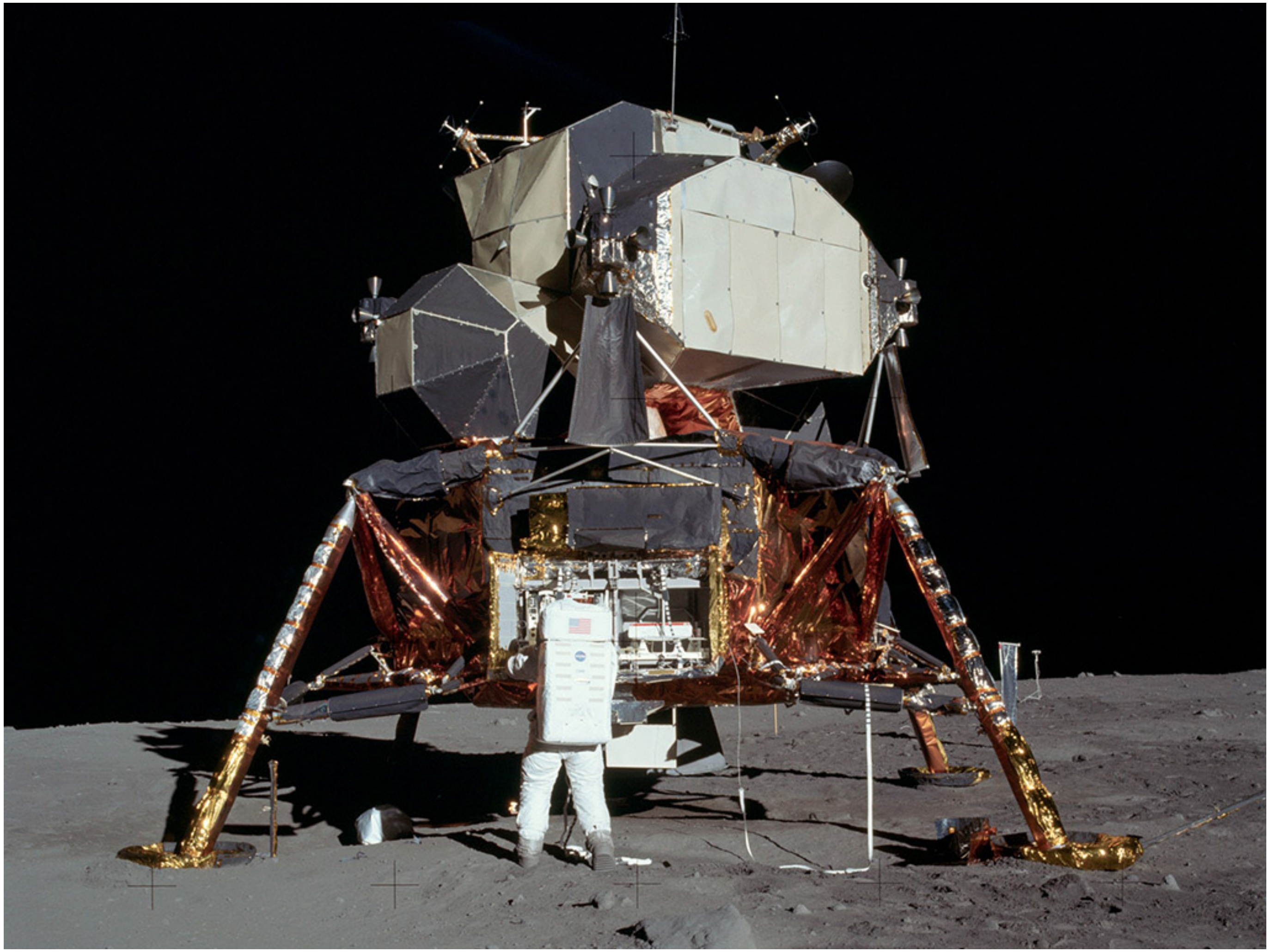
Norbert Wiener

Cybernetics:
The science of communications and automatic control systems in both machines and living things.

Hans Witsenhausen

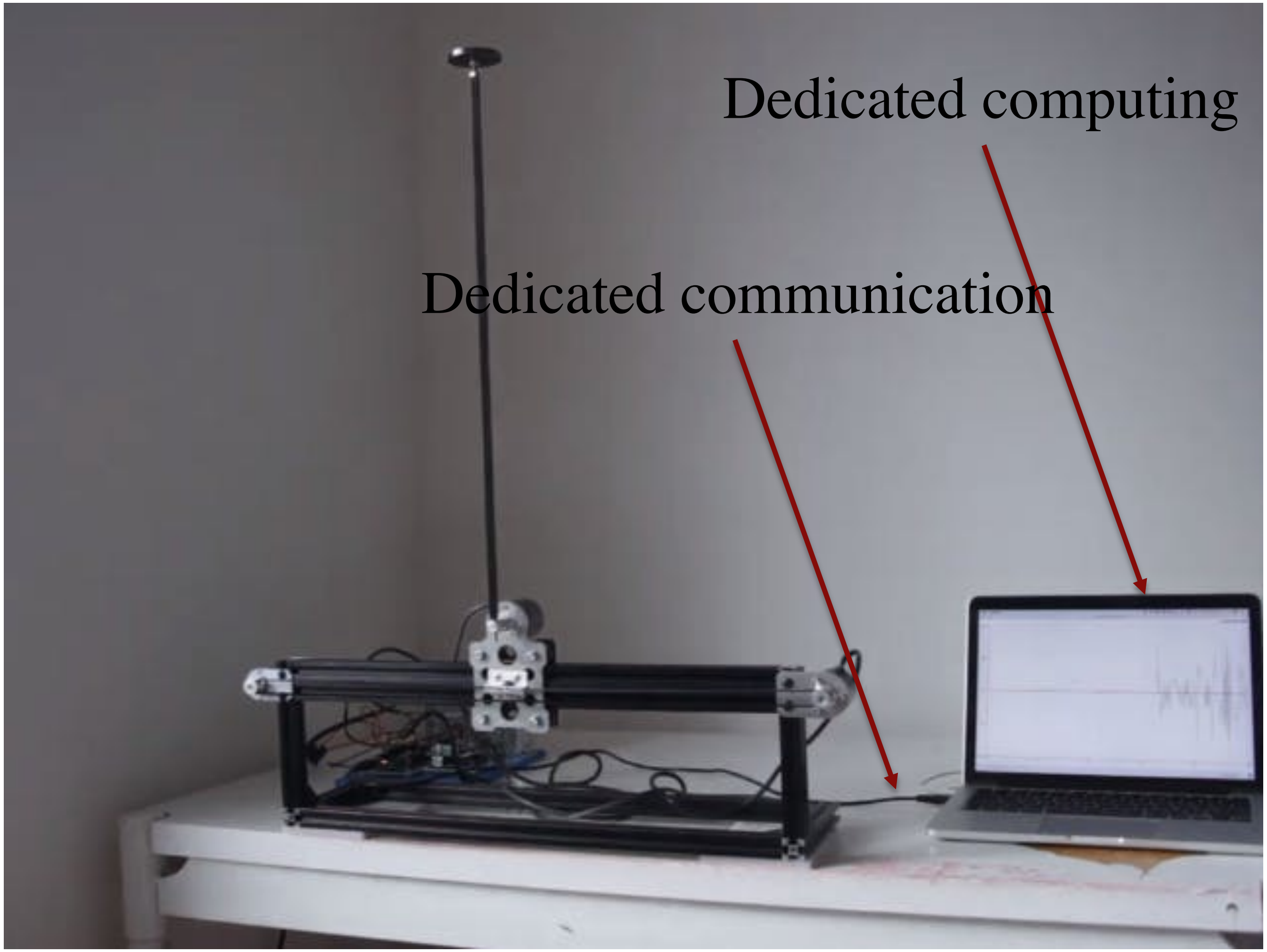
Information Patterns:
Who knows what and when

and systems built without CPS



How? Via separation of concerns…

$T_{compute} + T_{comm} < T_{sampling}$

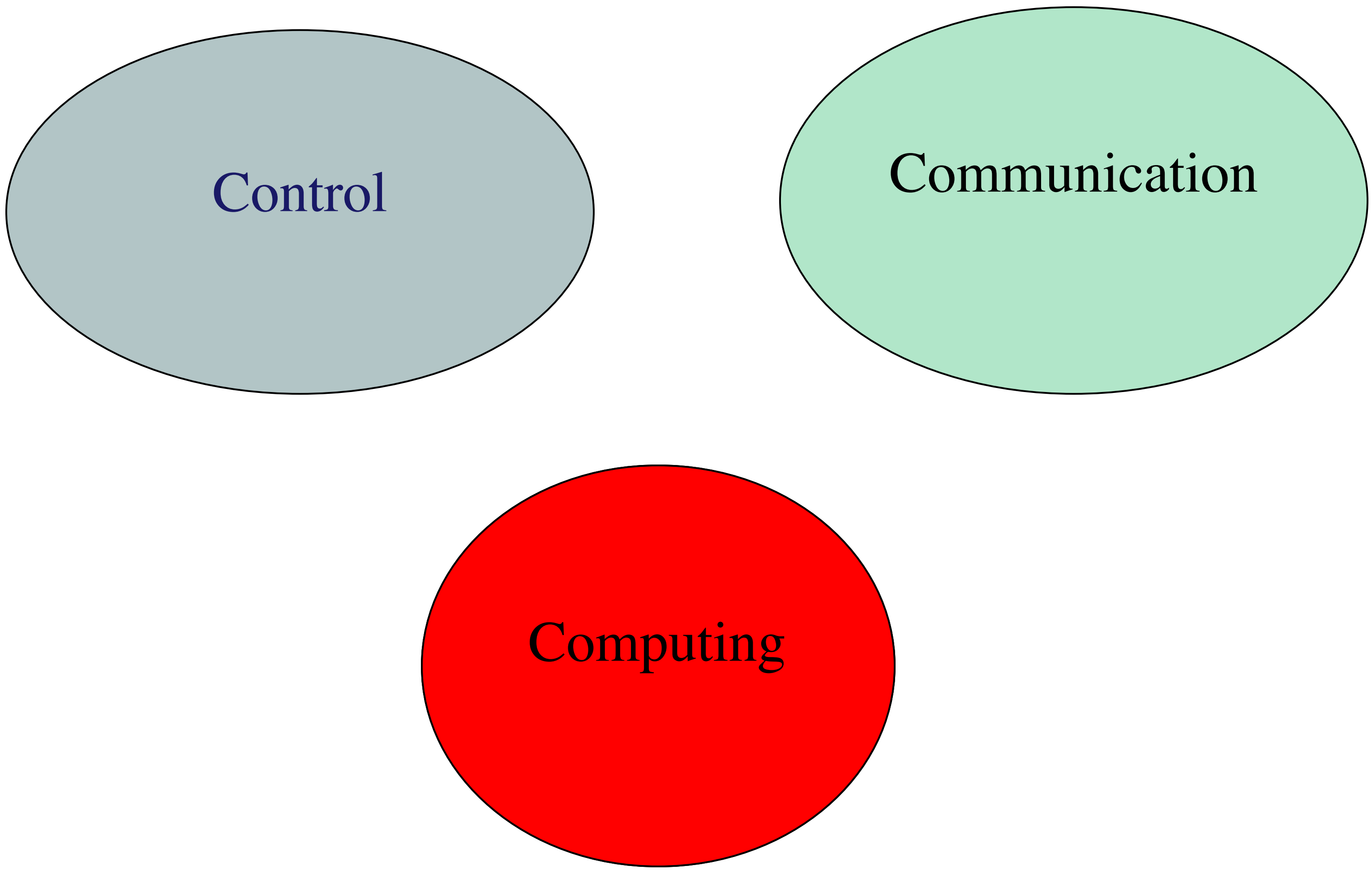


and via great investments

"At its peak, the Apollo program employed 400,000 people and required the support of over 20,000 industrial firms and universities"

https://www.nasa.gov/centers/langley/news/factsheets/Apollo.html

Largely Independent disciplines



Application Pull

Transportation

https://www.asmag.com/showpost/22236.aspx

Smart Buildings

https://www.forbes.com/sites/honeywell/2016/10/28/why-we-need-smart-buildings/#32499e5977d9

Manufacturing

https://medium.com/@julienmthn/5-questions-and-answers-about-smart-manufacturing-da9f600627a

Smart Grid

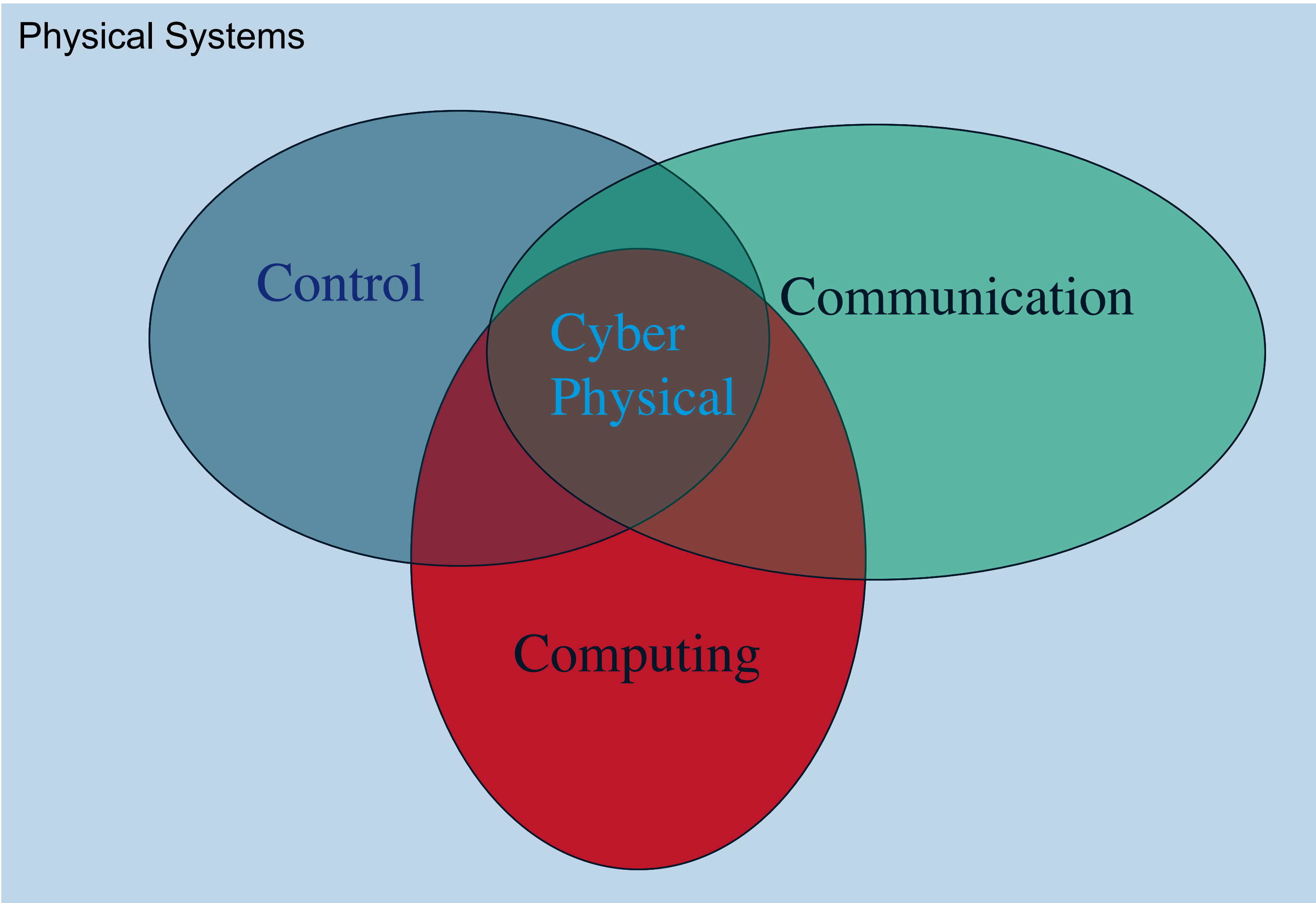
https://www.nts.com/industries/energy/smart-grid/

Technology Push

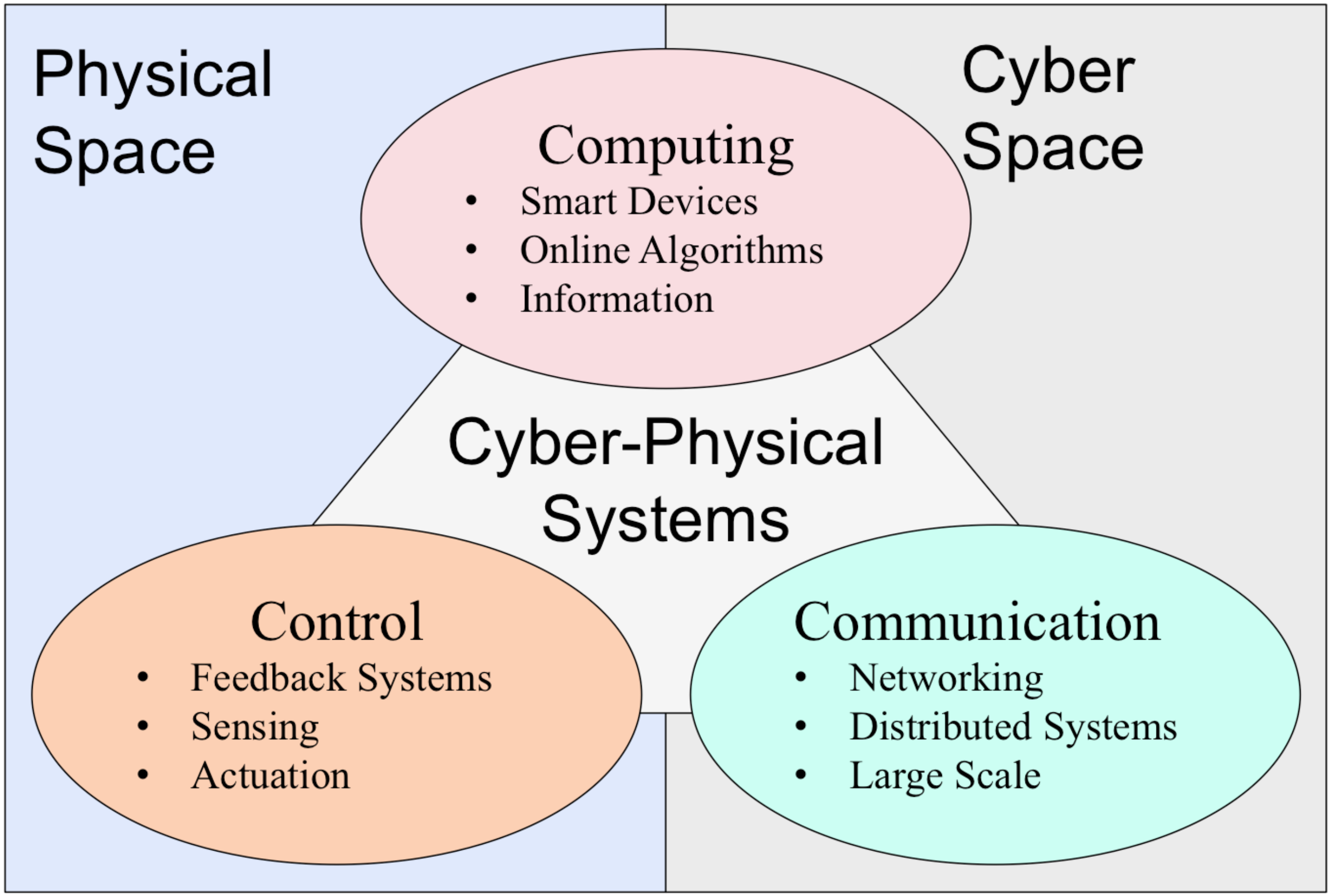
- Widespread networking, wireless, ubiquitous computing
- Off-the-shelf HW/SW

- No more dedicated computing/comm
- No more air gaps

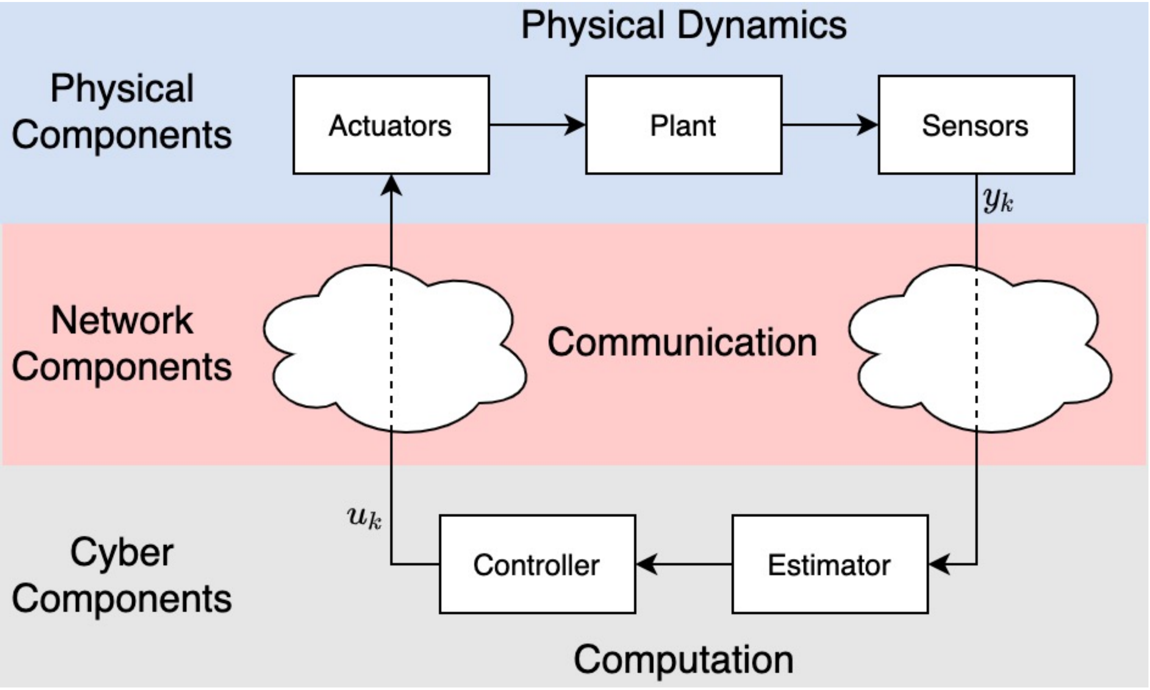
Things became less clean



Focus on the intersection of domains



Modeling Cyber-Physical Systems



Nonlinear

$$x_{k+1} = f_k(x_k, u_k, w_k)$$
$$y_k = h_k(x_k, u_k, v_k)$$

Linear

$$x_{k+1} = Ax_k + Bu_k + w_k$$
$$y_k = Cx_k + v_k$$

state vector: $x_k \in \mathbb{R}^n$

control inputs: $u_k \in \mathbb{R}^p$

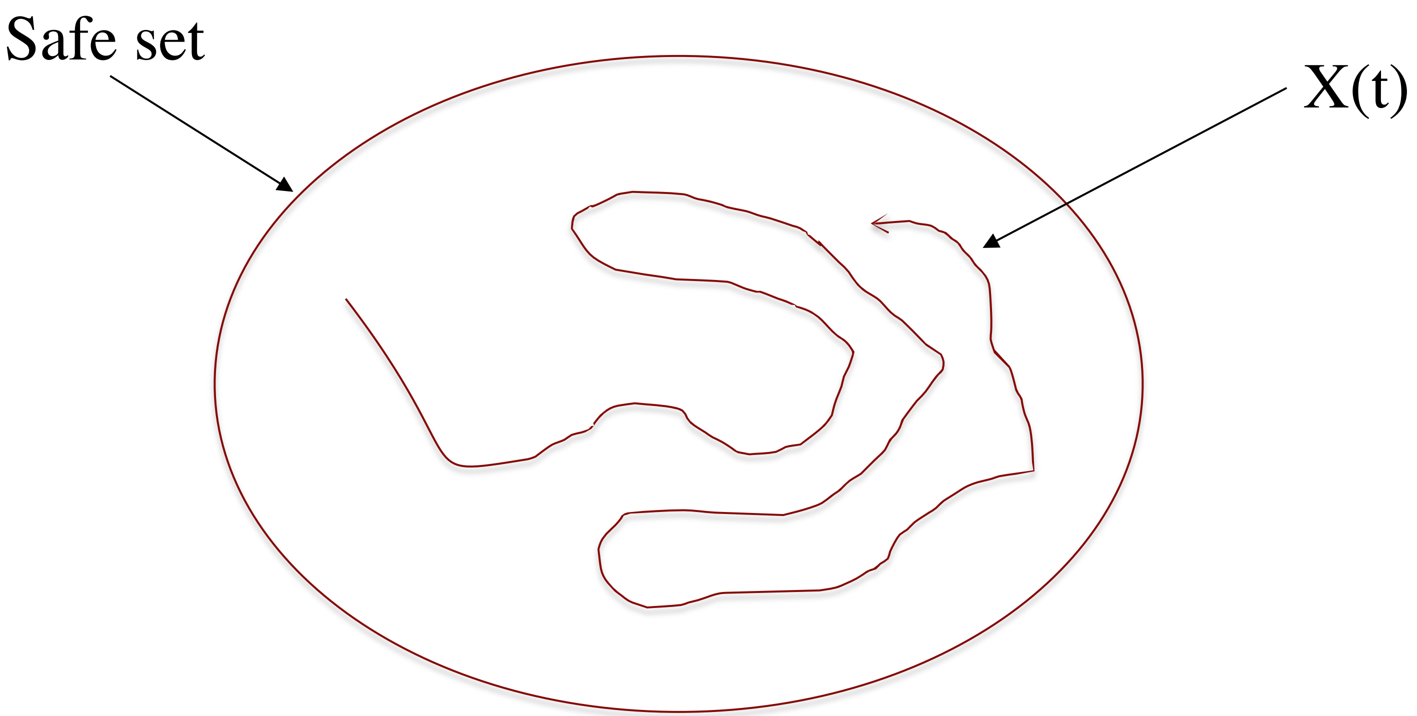
sensor measurements: $y_k \in \mathbb{R}^m$

process disturbance/noise: $w_k \in \mathbb{R}^n$

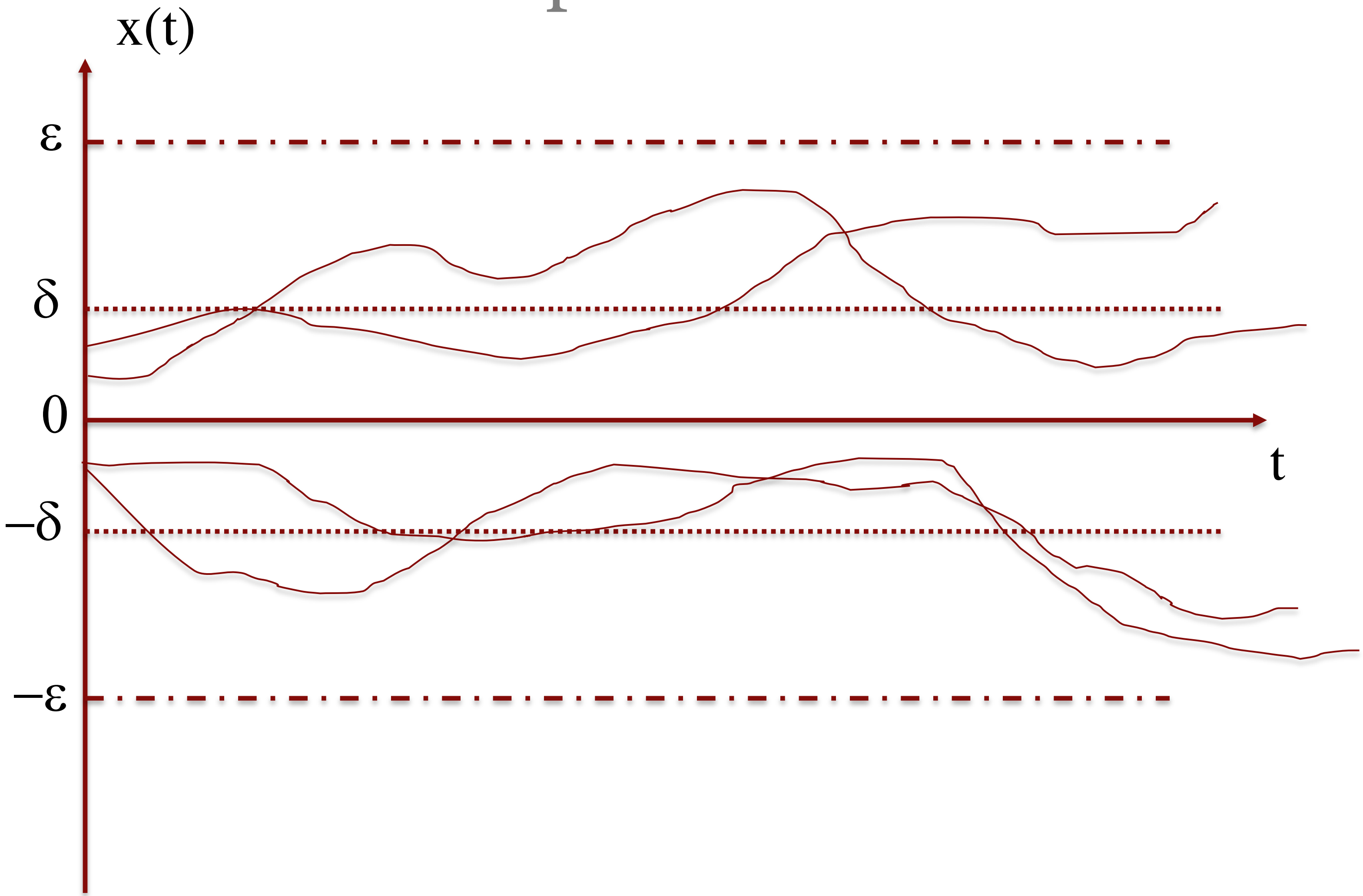
measurement disturbance/noise: $v_k \in \mathbb{R}^m$

From stability to safety

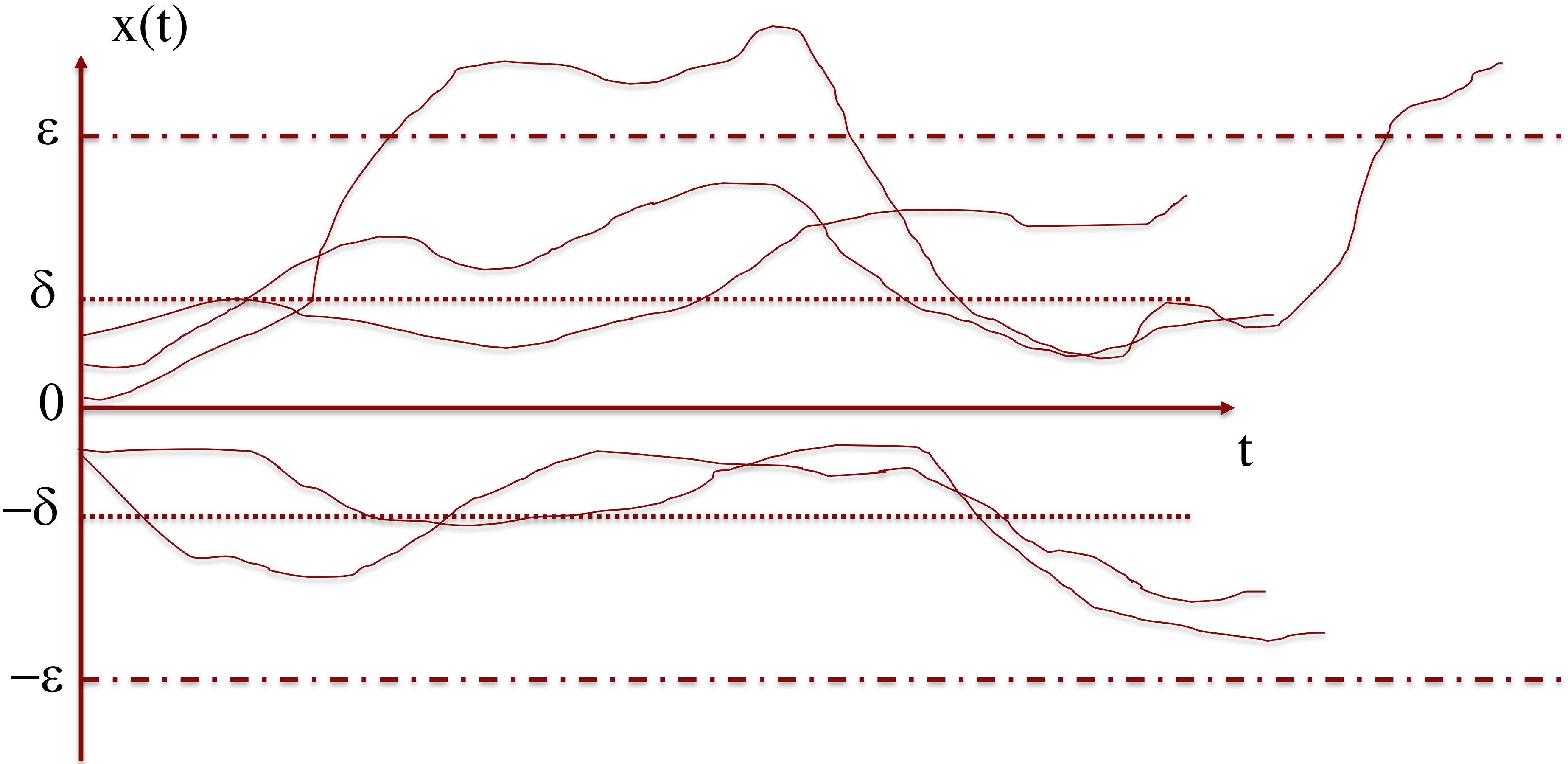
- Preserving safe operation of the CPS is the main goal…



Is (Asymptotic) Lyapunov Stability still a relevant concept?



What happens if trajectories occasionally exit the ε-ball?



Probably nothing as long as the set of states reached are safe

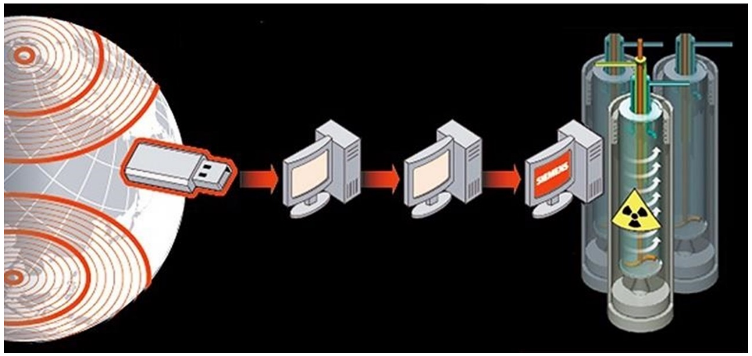
From robustness to resilience

- **Robustness**
  - Ability of the system to withstand perturbation without the need for adaptation
    - Pros: no need for adaptation
    - Cons: conservative design solutions, reduced performance

- **Resilience**
  - Ability of the system to respond to perturbation and restore a certain level of functionality
    - Pros: ability to restore full functionality, can be less conservative in design
    - Cons: added complexity

CPS security is a major issue

Stuxnet Malware (2010)



Colonial Pipeline CEO admits to authorizing $4.4 million ransomware payment

By Geneva Sands, CNN
Updated 5:15 PM ET, Wed May 19, 2021

Jeep wireless hack (2015)

WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY SIGN IN SUBSCRIB

ANDY GREENBERG SECURITY 07.21.2015 06:00 AM

Hackers Remotely Kill a Jeep on the Highway—With Me in It

I was driving 70 mph on the edge of downtown St. Louis when the exploit began to take hold.

Ukraine Power System Attack (2015)

CPS security is a major issue

- **There is strong evidence that the next wave of cyber attacks will target physical infrastructures.**
  - CPS are often a composition of various heterogeneous systems and components
  - CPS are increasingly connected, e.g can be accessed via the internet
  - The insider threat
- **Motivation**
  - Cyber warfare (disrupt key infrastructure, induce strategic damage)
  - Commercial advantage (espionage, reduce competitor's performance)
  - Ransom (just like Spectre in 007 movies)
- **It is a matter of national interest**
  - It is not just a technological problem
  - Public/private partnership may be needed

Cyber vs Cyber-Physical Security

- Inertia
- Continue operating under attack via graceful degradation
- Cultural issue
- Patches may be expensive

- Use predictive power of accurate models
- Sensor data and control inputs can be used as active monitors
- Physical channels can be used for authentication of cyber systems
- Prove security properties

Vision for CPS resilience

Goal: Design the system and the associated security countermeasures so that graceful degradation is achieved when the system is under attack

CPS Design

1) System Design
Design controller and system for performance and security

Detection

2) Detect Attacks
Leverage system knowledge to recognize attacks

Identification

3) Isolate Attacks
Use system knowledge to isolate malicious/faulty components

Time-Triggered Response

4) Attack Resilience I
Deploy time-triggered prevention mechanisms to guarantee security

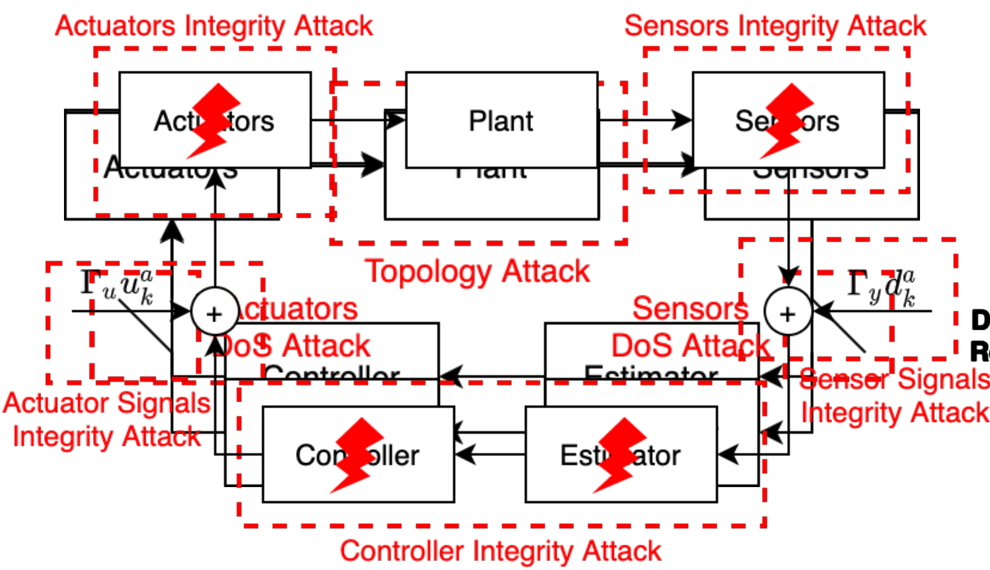
Detection-Triggered Response

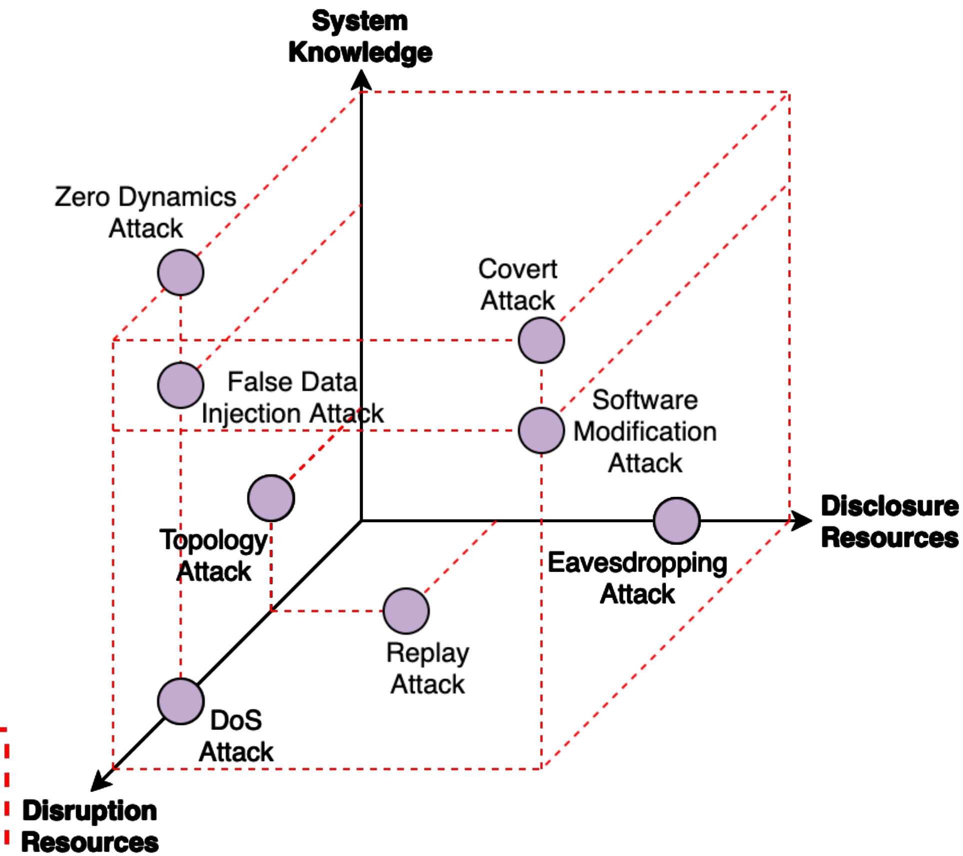
5) Attack Resilience II
Deploy detection-triggered mechanisms to maximize security and performance

Our Focus

Attacker Capabilities[1]

- System knowledge
- Disclosure resources
  - Eavesdropping attack
- Disruption resources
  - Topology attack
  - Denial of service attack
  - Integrity attack

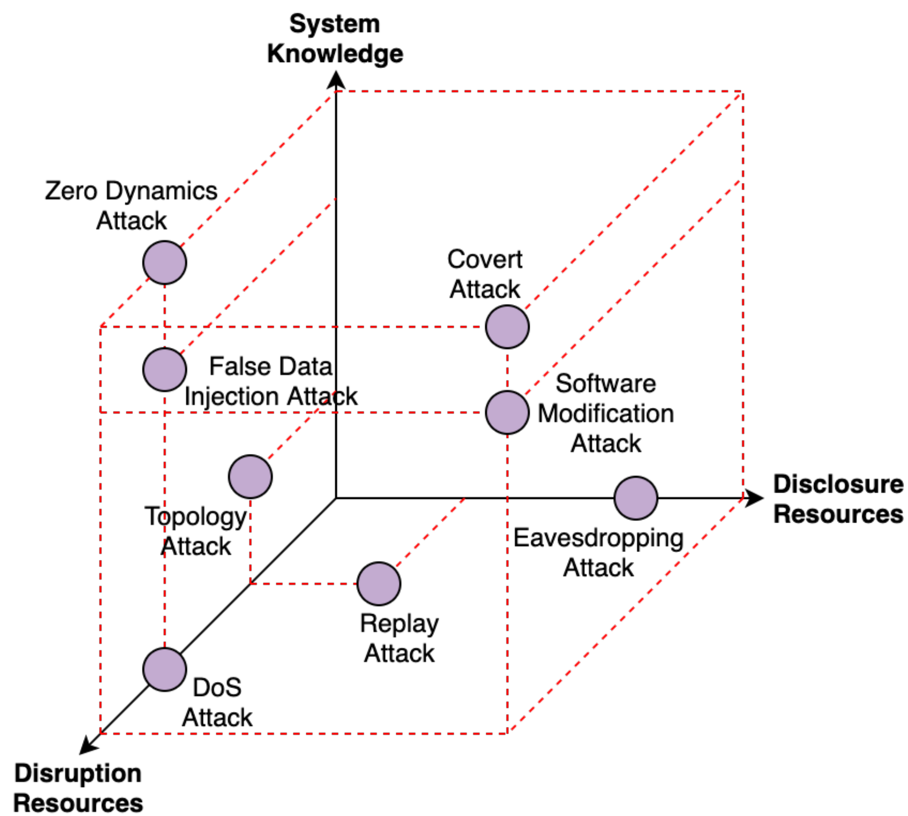




[1] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "A secure control framework for resource-limited adversaries," *Automatica*, vol. 51, pp. 135–148, 2015.

Attack Strategies[1]

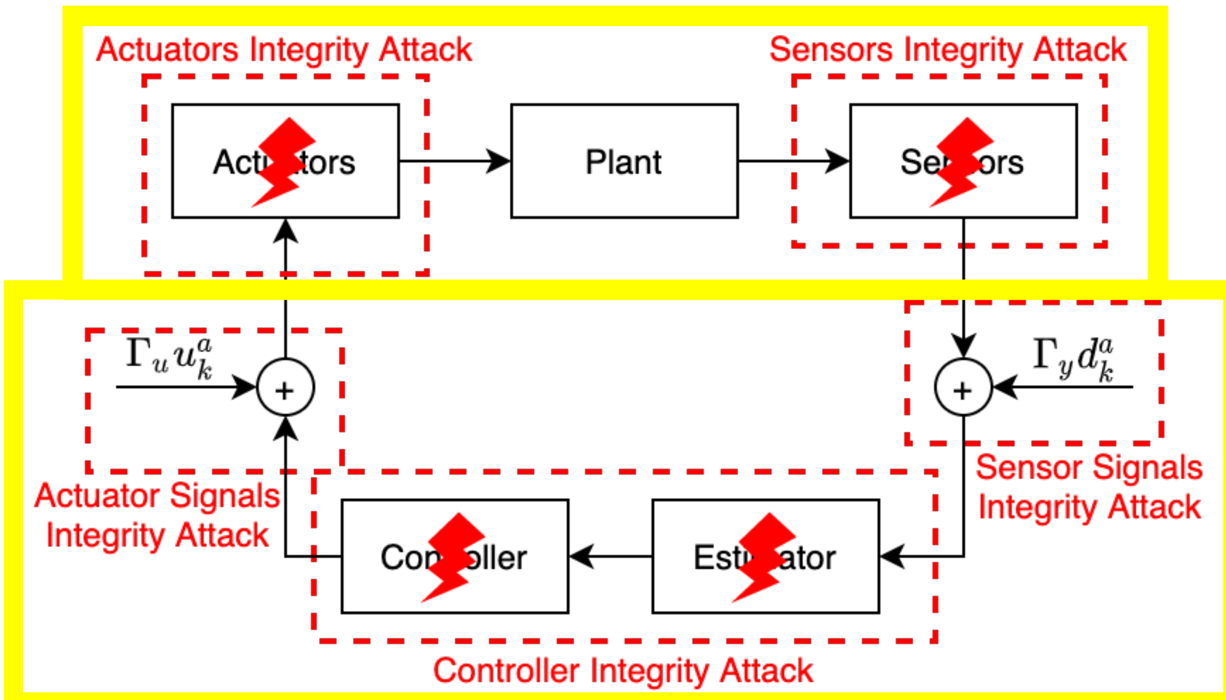
- Compromise confidentiality
  - Eavesdropping attack
- Compromise availability
  - Denial of service attack
- Compromise integrity
  - Topology attack
  - Integrity attack
    - Replay attack
    - False data injection attack
    - Zero dynamics attack
    - Covert attack
    - Software modification attack



[1] A. A. Cardenas, S. Amin, and S. Sastry, "Secure control: Towards survivable cyber-physical systems," in *2008 The 28th International Conference on Distributed Computing Systems Workshops*. IEEE, 2008, pp. 495–500.

Integrity Attacks

- Can be performed in both the cyber and physical realms
- Cyber realm: attacks on the controller, actuator signals, or sensor signals
- Physical realm: attacks on the actuators or sensors

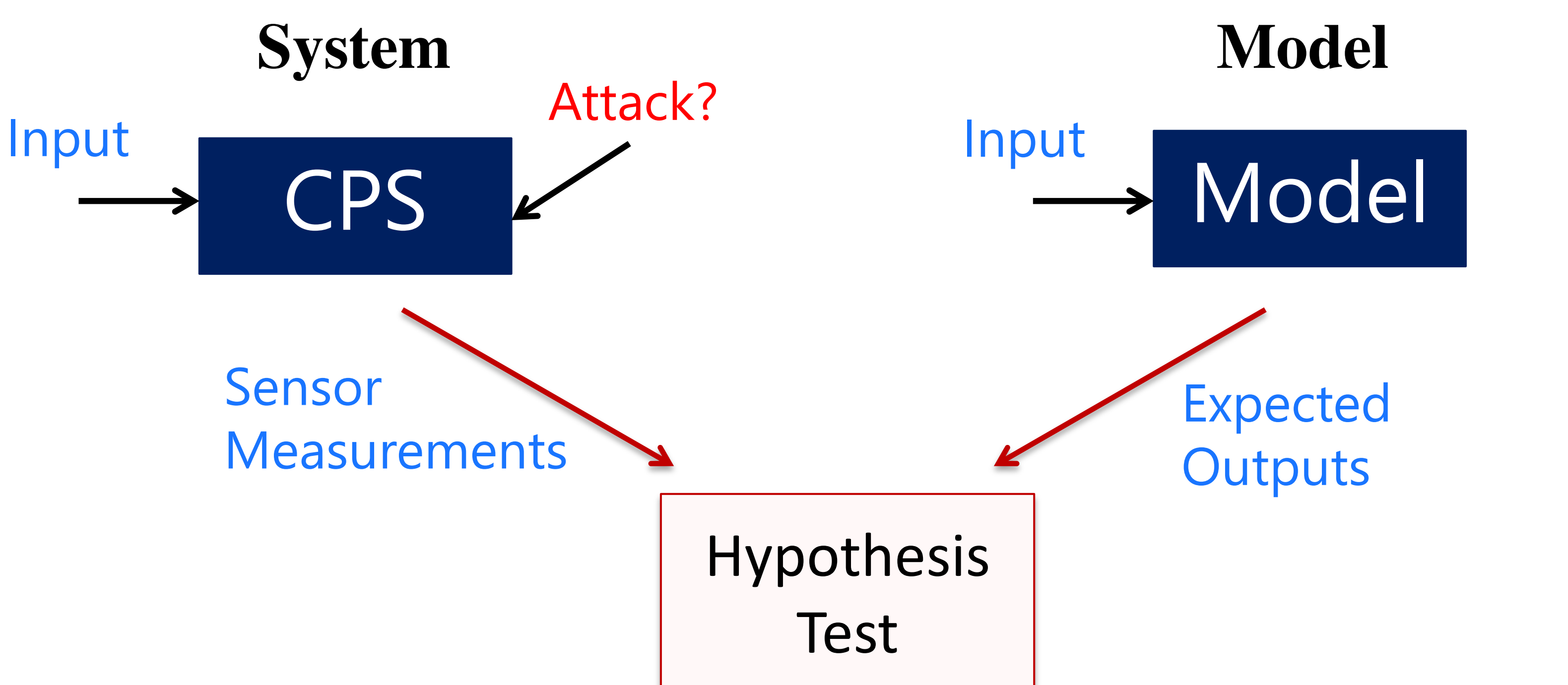


$$x_{k+1} = Ax_k + B(u_k + \Gamma_u u_k^a) + w_k$$

$$y_k^a = Cx_k + \Gamma_y d_k^a + v_k$$

Passive Detection

- Detect interference from an attacker using standard detection techniques
- Assuming that the dynamical model is known, leverage existing detection theory to detect attacks
- Utilize data from passive observation of sensor measurements

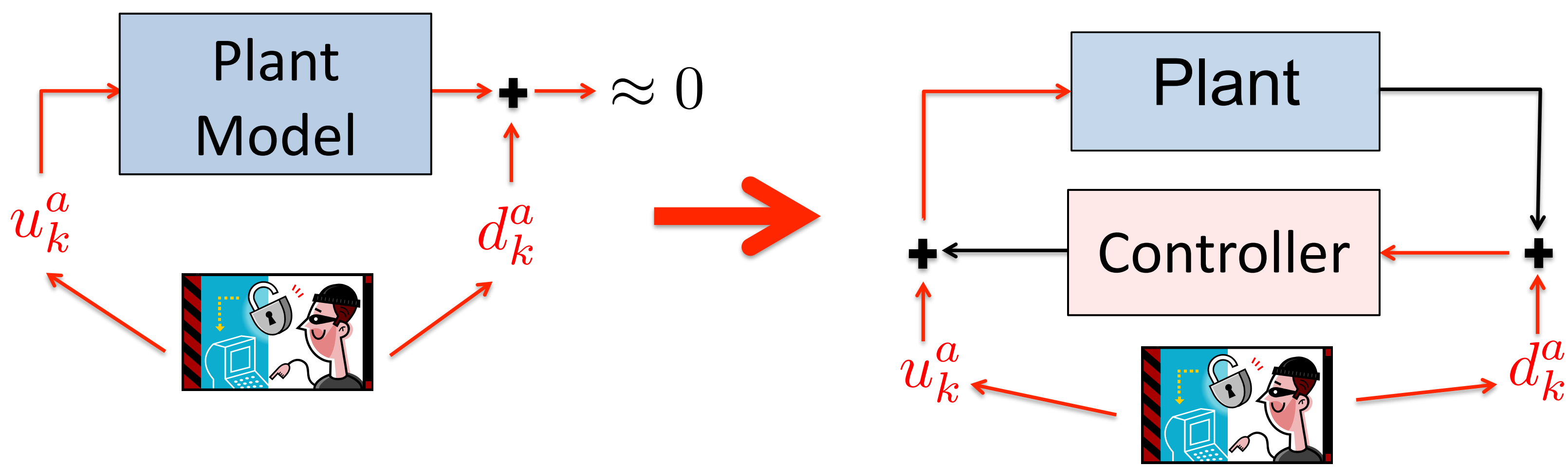


Limitations of Passive Detection[1]

- Highly knowledgeable and powerful adversaries can bypass passive detection techniques
- Attacks can be designed so that the outputs received by a system operator are statistically consistent with expected output behavior

$$x_{k+1} = Ax_k + B(u_k + u_k^a) + w_k$$
$$y_k = Cx_k + d_k^a + v_k$$

Covert attack: $d_k^a = -Cx_k^a, \quad x_{k+1}^a = Ax_k^a + Bu_k^a, \quad x_0^a = 0$



[1] R. S. Smith, "Covert misappropriation of networked control systems: Presenting a feedback structure," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 82–92, 2015.

The value of analysis: illustrative example

- **We consider a vehicle moving along the x -axis.**

$$\dot{x}_{k+1} = \dot{x}_k + w_{k,1},$$

$$x_{k+1} = x_k + \dot{x}_k + w_{k,2}$$

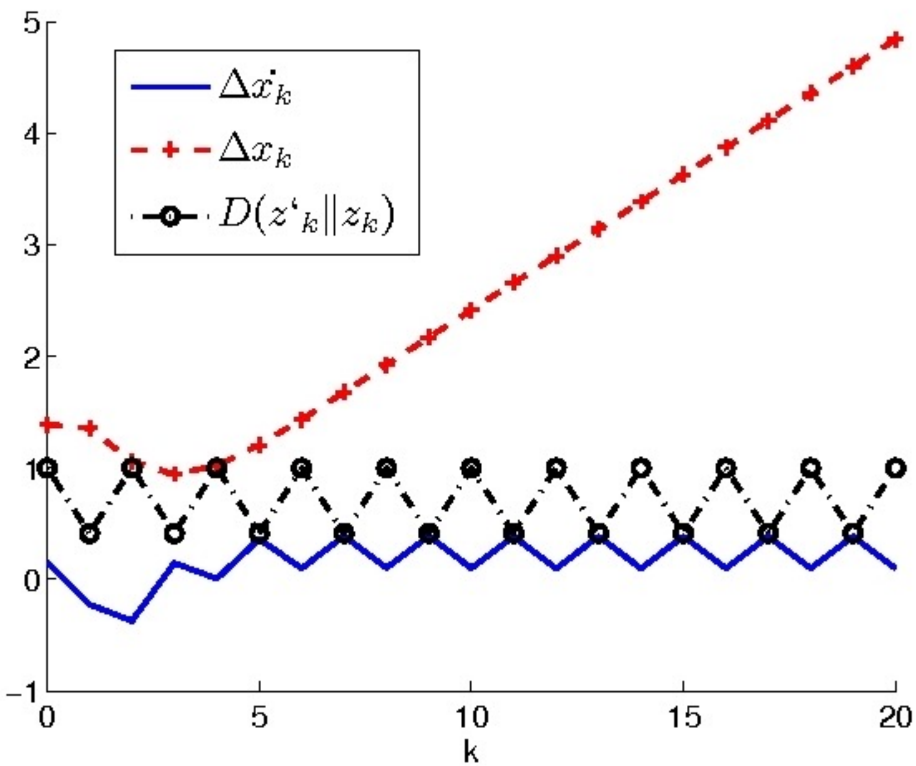
- **Two sensors are used to measure position and velocity respectively.**

$$y_{k,1} = \dot{x}_k + v_{k,1},$$

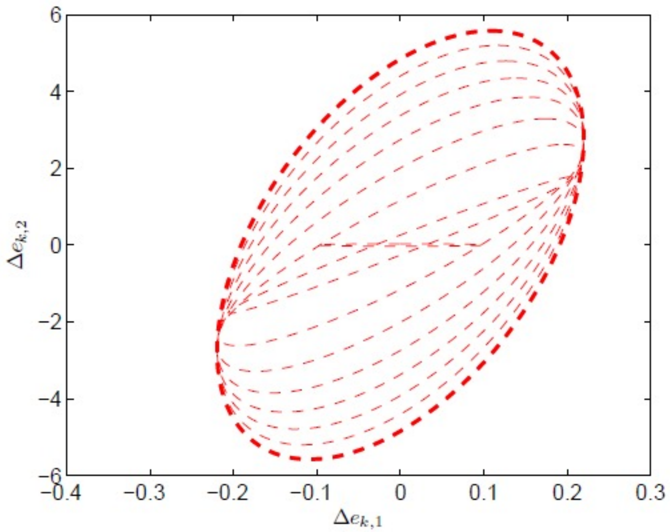
$$y_{k,2} = x_k + v_{k,2}.$$

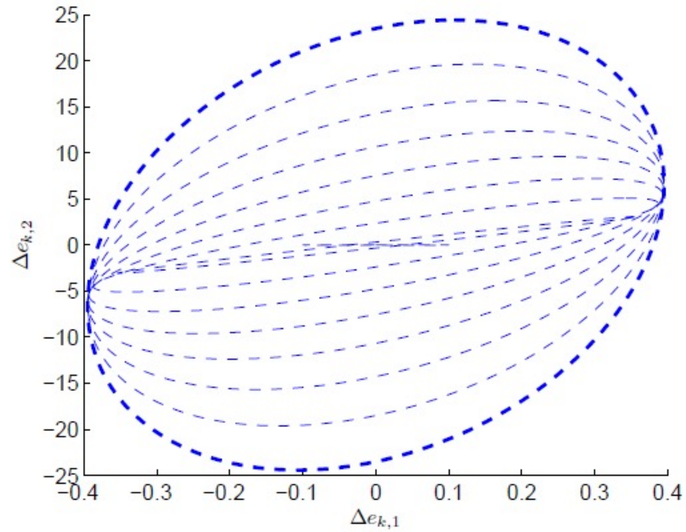
- **We assume that** $Q = R = I_2$ **.**

Position sensor is compromised: the system can be destabilized

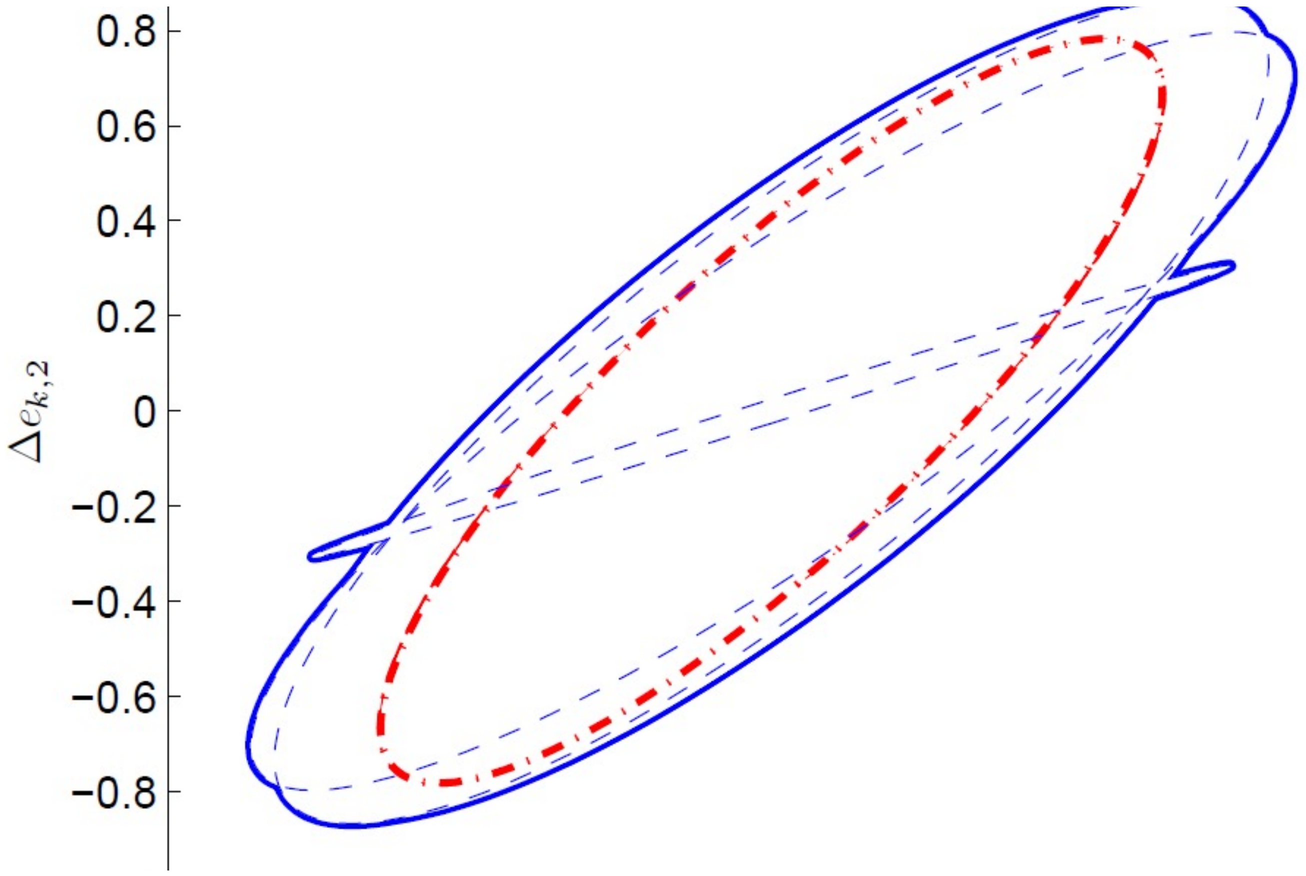


Simulation Result: Compromising the Position Sensor



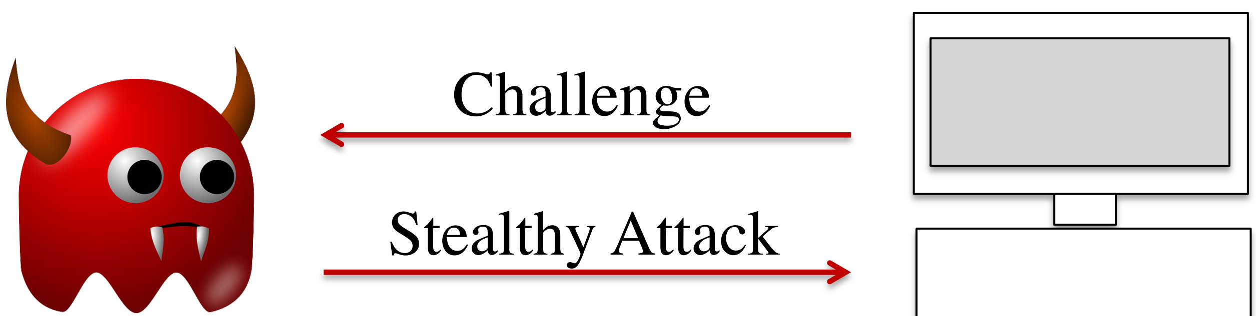


Velocity Sensor is compromised: Maximum Perturbation is bounded



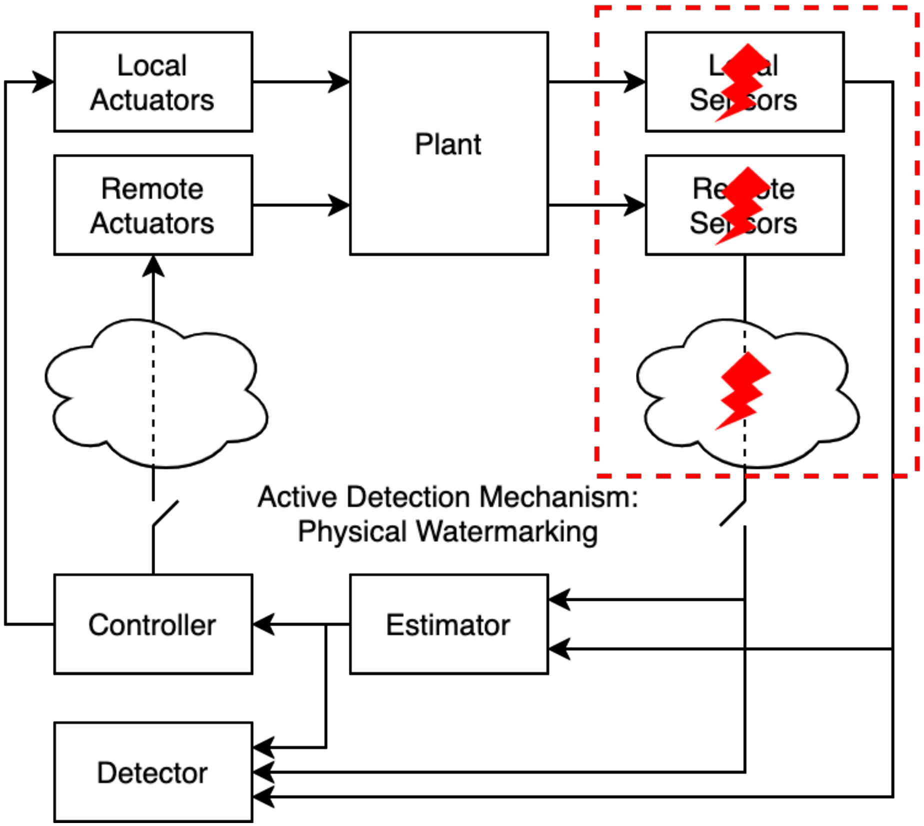
Active Detection

- Actively perturb the system, leveraging the system's available degrees of freedom to detect attacks
- Introduce a challenge response physical authentication into the system
  - The challenge is based on a secret unknown to the adversary
  - The secret is embedded in the physical dynamics using degrees of freedom in the control system/parameters

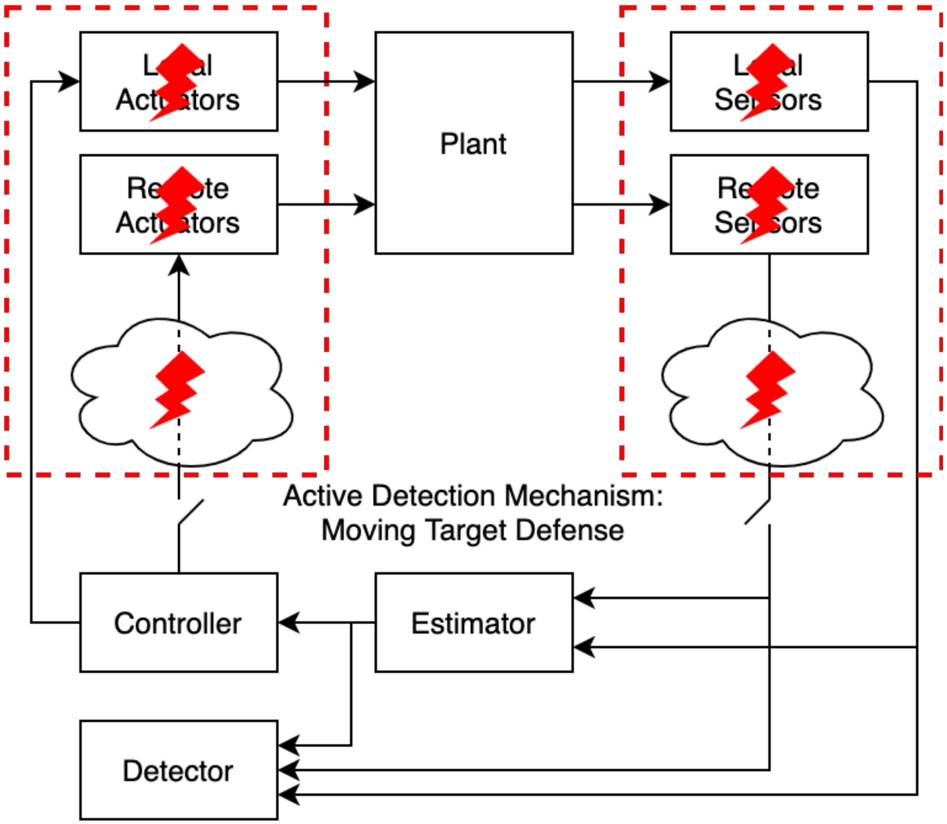


  - Poor responses provide proof of attacker's presence due to inconsistencies with modeling

Overview of Active Detection Mechanisms



Active Detection Mechanism for Attacks on the Sensor Measurements



Active Detection Mechanism for Attacks on the Control Inputs and Sensor Measurements

Physical watermarking as an active detection scheme

Mo et al.,

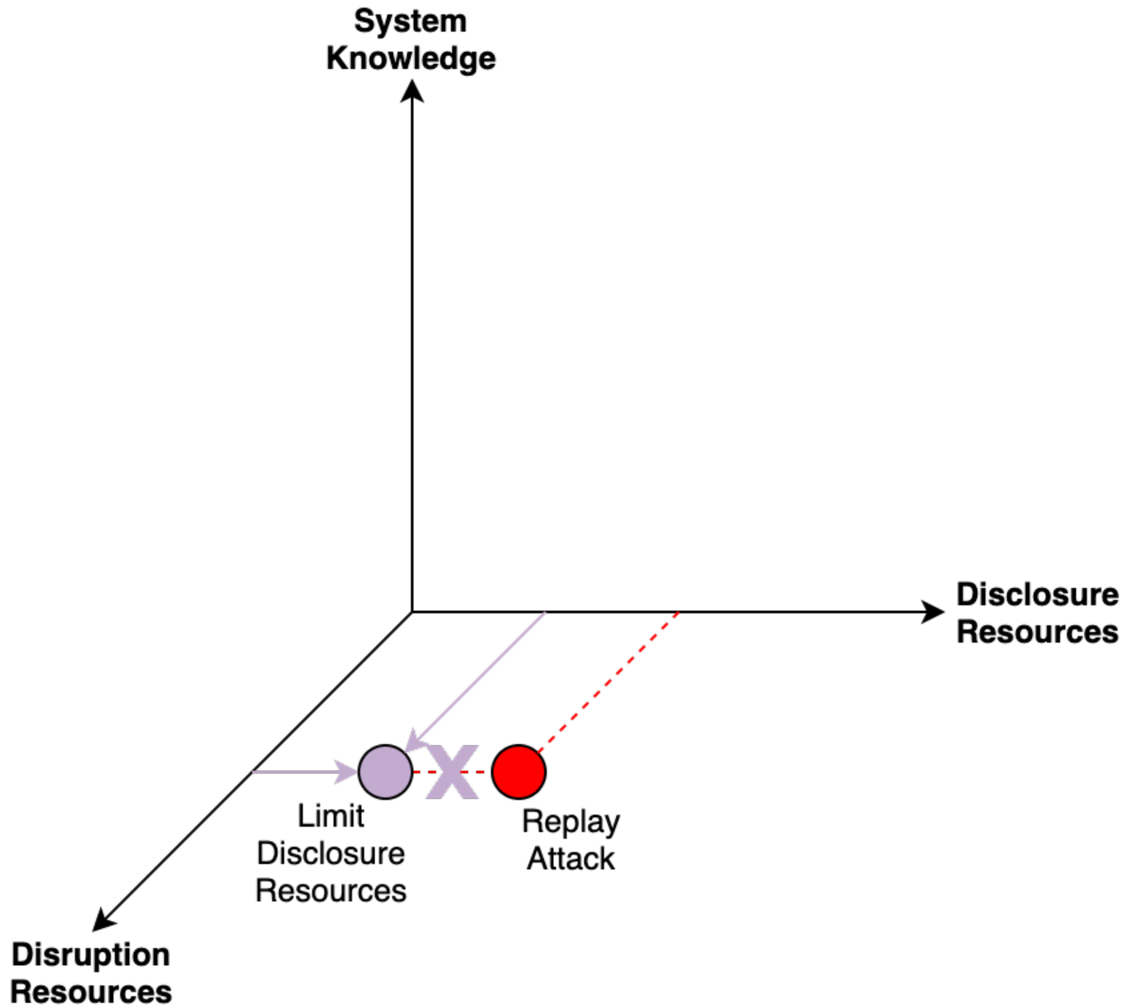
Allerton 2009, IEEE TCST 2014, IEEE CSM 2015

Replay Attack Model

- The attacker can
  - Record and modify the sensors' readings y_k
  - Inject malicious control input
- Replay Attack
  - Record sufficient number of y_k without adding control inputs.
  - Inject malicious control input to the system and replay the previous y_k. We denote the replayed measurements to be y_k'.
- When replay begins, there is no information from the systems to the controller. As a result, the controller cannot guarantee any close-loop control performance. The only chance is to detect the replay.

Physical Watermarking

Goal: limit the adversary's disclosure resources

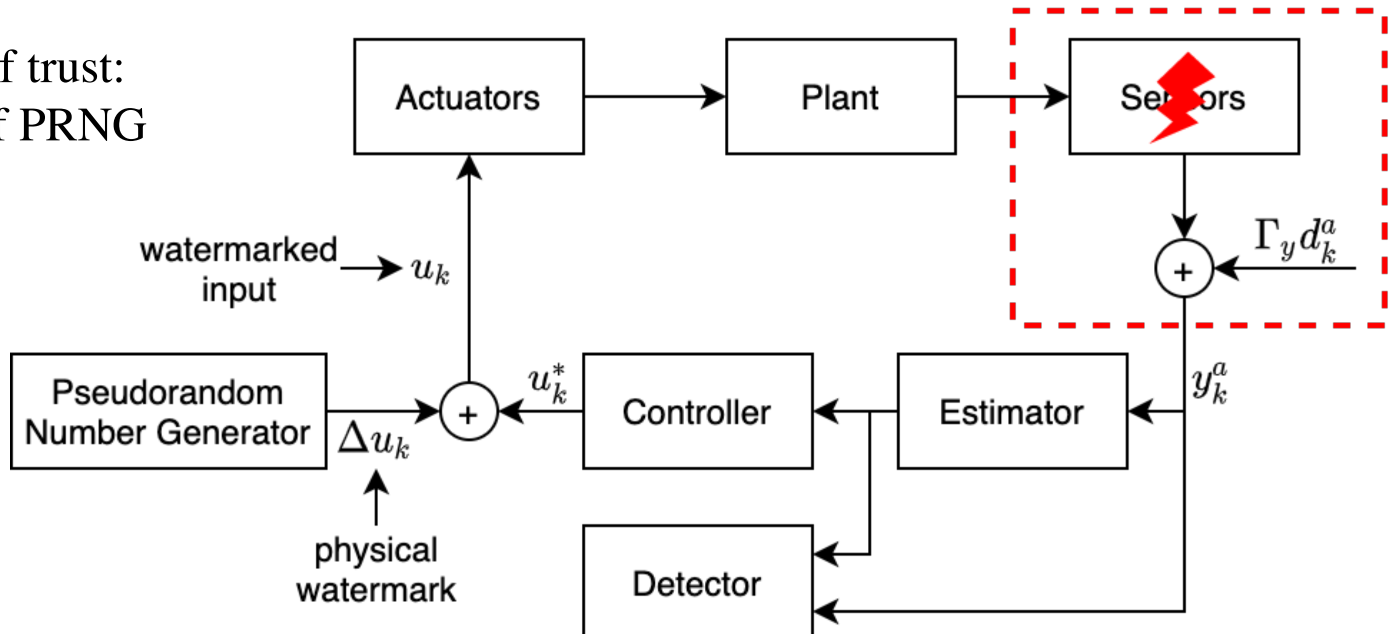


Physical Watermarking

- A cyber-physical “nonce” or small perturbation introduced in the control input
- Is effective in detecting replay attacks
- Introduces a tradeoff between detection and system performance

$$x_{k+1} = Ax_k + B(u_k^* + \Delta u_k) + w_k$$
$$y_k^a = Cx_k + \Gamma_y d_k^a + v_k$$

Root of trust:
seed of PRNG



The System Model

Suppose we have system dynamics as follows:

$$x_{k+1} = Ax_k + Bu_k + w_k \qquad x_k \in \mathbb{R}^n,\ u_k \in \mathbb{R}^p,\ w_k \sim \mathcal{N}(0, Q)$$

$$y_k = Cx_k + v_k \qquad y_k \in \mathbb{R}^m,\ v_k \sim \mathcal{N}(0, R)$$

A Linear Quadratic Gaussian controller is implemented.

Linear Quadratic Regulator

$$J = \lim_{T\to\infty} \frac{1}{2T+1} \mathbb{E}\left[\sum_{k=-T}^{T} x_k^T W x_k + u_k^T U u_k\right]$$

$$u = u_k^* = L\hat{x}_{k|k} \qquad L = -\left(B^T S B + U\right)^{-1} B^T S A$$

Kalman Filter

$$\hat{x}_{k+1|k} = A\hat{x}_{k|k} + Bu_k \qquad \hat{x}_{k|k} = \hat{x}_{k|k-1} + Kz_k$$

$$z_k = y_k - C\hat{x}_{k|k-1} \qquad K = PC^T(CPC^T + R)^{-1}$$

Failure Detector

- **A failure detector is used to detect abnormality in the system, which triggers an alarm based on the following condition:**

$$g_k > threshold$$

where $$g_k = g(y_k, \hat{x}_k, \ldots, y_{k-\mathcal{T}}, \hat{x}_{k-\mathcal{T}}),$$

and the function g **is continuous.**

Failure Detector

- **For example,** g_k **for a chi-square detector takes the following form:**

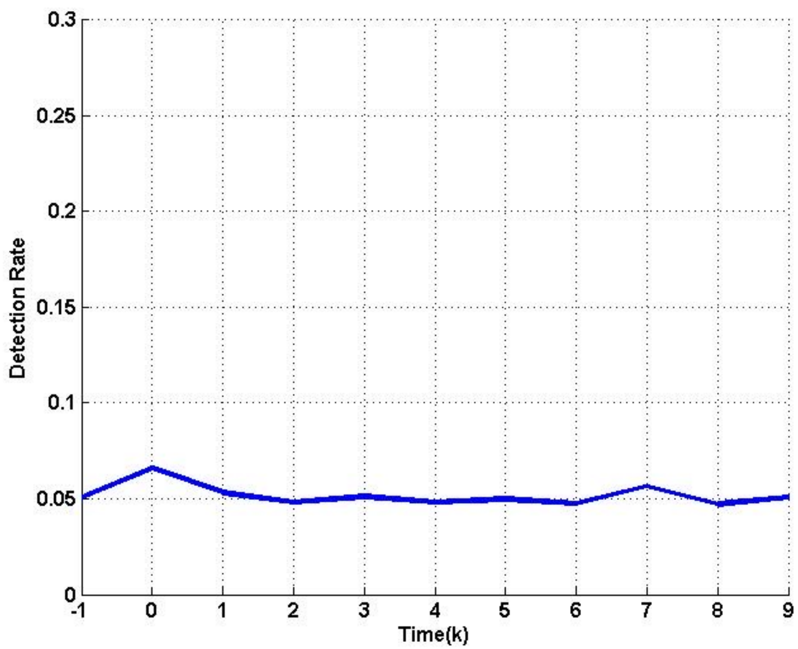
$$g_k = z_k^T \mathcal{P}^{-1} z_k$$

where $$z_k = y_k - CA\hat{x}_{k-1},$$

and $\mathcal{P}$ **is the covariance of** z_k **.**

A X^2 detector may not detect the attack

- **Suppose the attacker records from time –T and replay begins at time 0.**



- **Detection rate is equal to false alarm rate… no detection**

Detection of Replay Attack

- **Manipulating equations:**

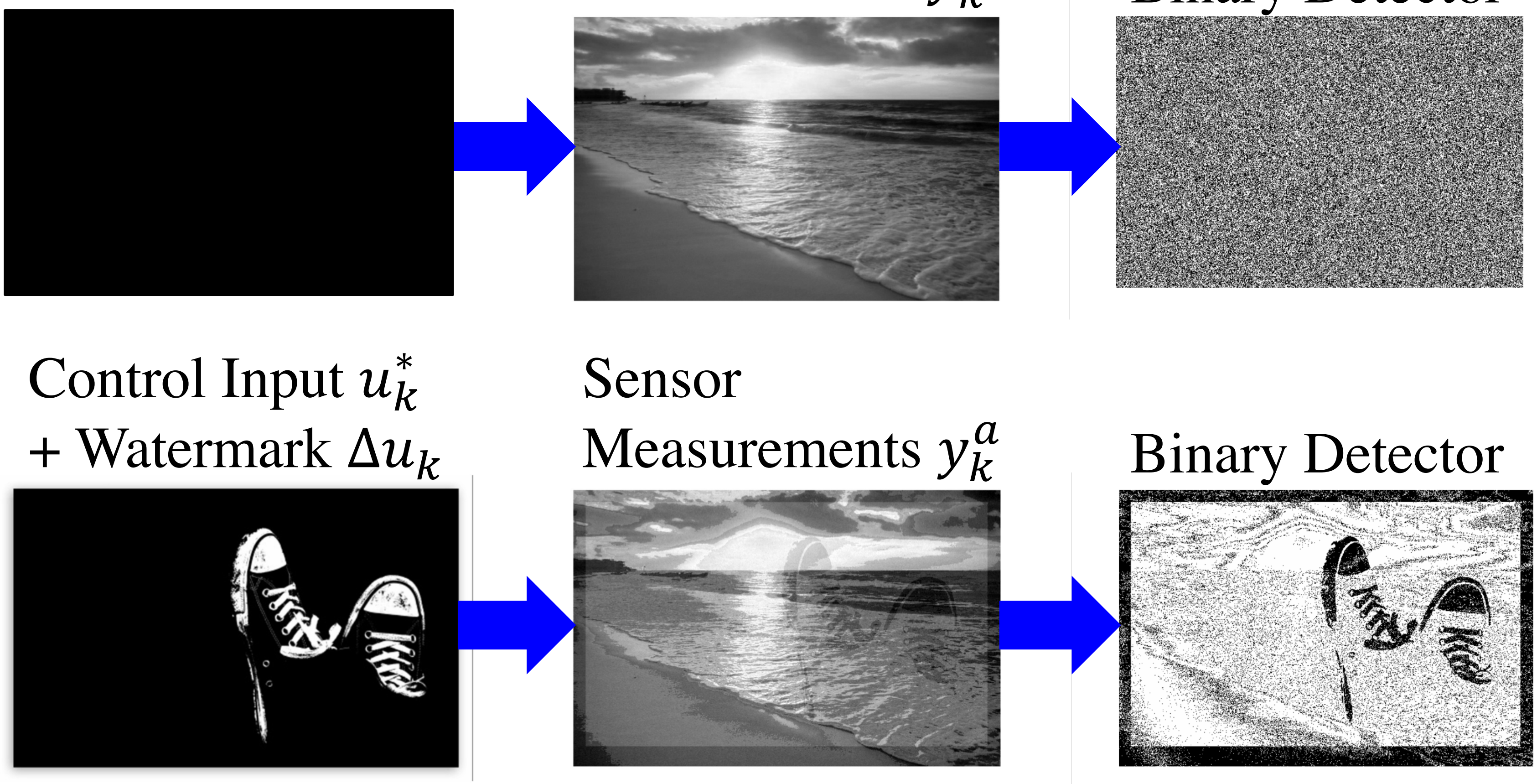
$$\underbrace{\boxed{y'_k - C\hat{x}_{k|k-1}}}_{\uparrow \atop \text{innovation under replay}} = \underbrace{\boxed{\left(y_{k-T} - C\hat{x}_{k-T|k-T-1}\right)}}_{\uparrow \atop \text{innovation without replay}} + \underbrace{\boxed{C\mathcal{A}^k\left(\hat{x}_{0|-1} - \hat{x}_{-T|-T-1}\right)}}_{\uparrow \atop \text{converges to 0 if } \|\mathcal{A}\| < 1},$$

- **If $\mathcal{A}^k$ converges to 0 very fast, then there is no way to distinguish the compromised system and healthy system.**

Physical Watermarking

Control Input u_k^* | Sensor Measurements y_k^a | Binary Detector

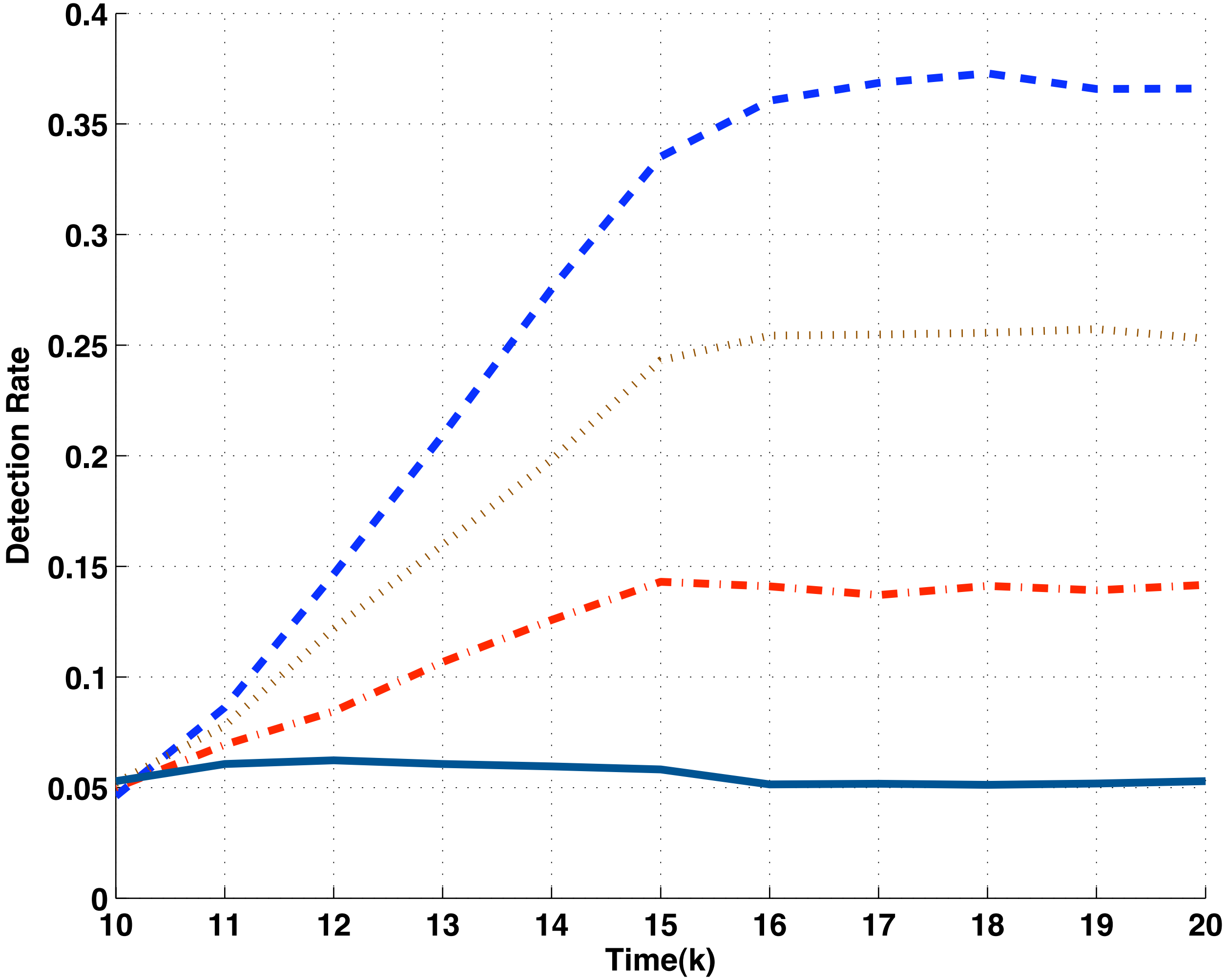
Control Input u_k^* + Watermark Δu_k | Sensor Measurements y_k^a | Binary Detector



Counter Measure

- **Innovation with random input:**

$$y'_k - C\hat{x}_{k|k-1} = y_{k-T} - C\hat{x}_{k-T|k-T-1} + C\mathcal{A}^k(\hat{x}_{0|-1} - \hat{x}_{-T|-T-1})$$
$$+ \boxed{C\sum_{i=0}^{k-1}\mathcal{A}^{k-i-1}B(\Delta u_i - \Delta u_{-T+i})} \leftarrow \text{Can be detected!.}$$



Blue: $\mathcal{Q} = 0.6$

Brown: $\mathcal{Q} = 0.4$

Red: $\mathcal{Q} = 0.2$

Dark Blue: $\mathcal{Q} = 0$

Detection Rate of Different Random Signal Strength

Effect of Authentication Signal

- **Expectation of residuals increases under attack, which triggers detector**

where $E\left[g_k\right] = m\mathcal{T} + 2\mathcal{T}\,\mathrm{tr}\left(\mathbf{C}\mathcal{P}^{-1}\mathbf{C}\mathcal{U}\right)$

- **Performance cost increases**

$$\mathcal{U} = \mathcal{A}\mathcal{U}\mathcal{A}^T + \mathbf{B}\mathcal{Q}\mathbf{B}^T$$

$$J = J^* + \mathrm{tr}\left[\left(\mathbf{U} + \mathbf{B}^T\mathbf{S}\mathbf{B}\right)\mathcal{Q}\right]$$

Optimization Goals

- Constrain performance loss to be below certain ΔJ value and maximize Δg_k

OR

- Constrain increase in expectation of to be above certain value g_k , while minimizing loss of performance ΔJ

[1] Under attack, the residuals follow a generalized distribution, and an analytical form for detection rate does not exist. We thus maximize the increase Δg_k hoping for maximum detection rate χ^2

Optimize for Q

$$\underset{Q}{\text{maximize}} \quad trace(C^T P^{-1} C U)$$

$$\text{subject to} \quad U - BQB^T = AUA^T$$

$$trace[(U + B^T SB)Q] \leq \Delta J$$

OR

$$\underset{Q}{\text{minimize}} \quad trace[(U + B^T SB)Q]$$

$$\text{subject to} \quad U - BQB^T = AUA^T$$

$$trace(C^T P^{-1} C U) \geq E[\Delta g_k]$$

Some Remarks

- **Solving either optimization problem guarantees same performance.**
- **An intuitive way to see this, is that *Q* measures sensitivity of system to different forms of authentication signal**
- **Form of Q* should be a property of the system.**

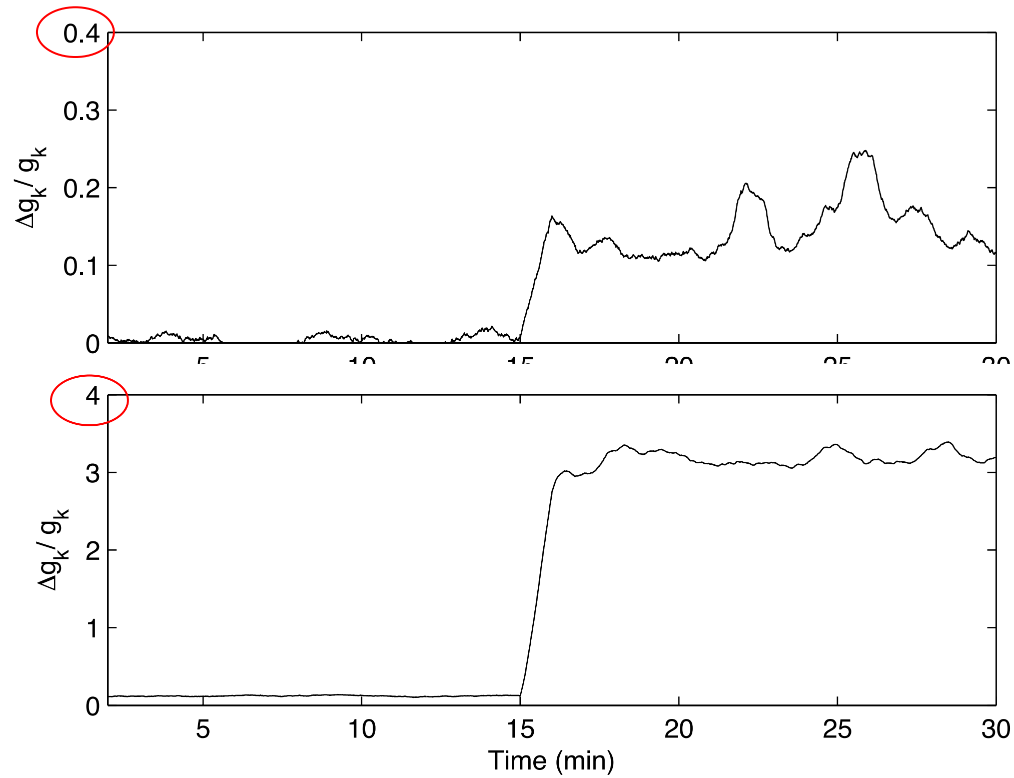
Decoupling

- **Linear programming enables us to decouple the control problem into two steps:**
  - First find the direction of $Q^* = vv'$
  - Then decide upon the norm of Q^*
- **Equivalent to deciding the vector direction of the signal, then the vector magnitude**

Decoupling

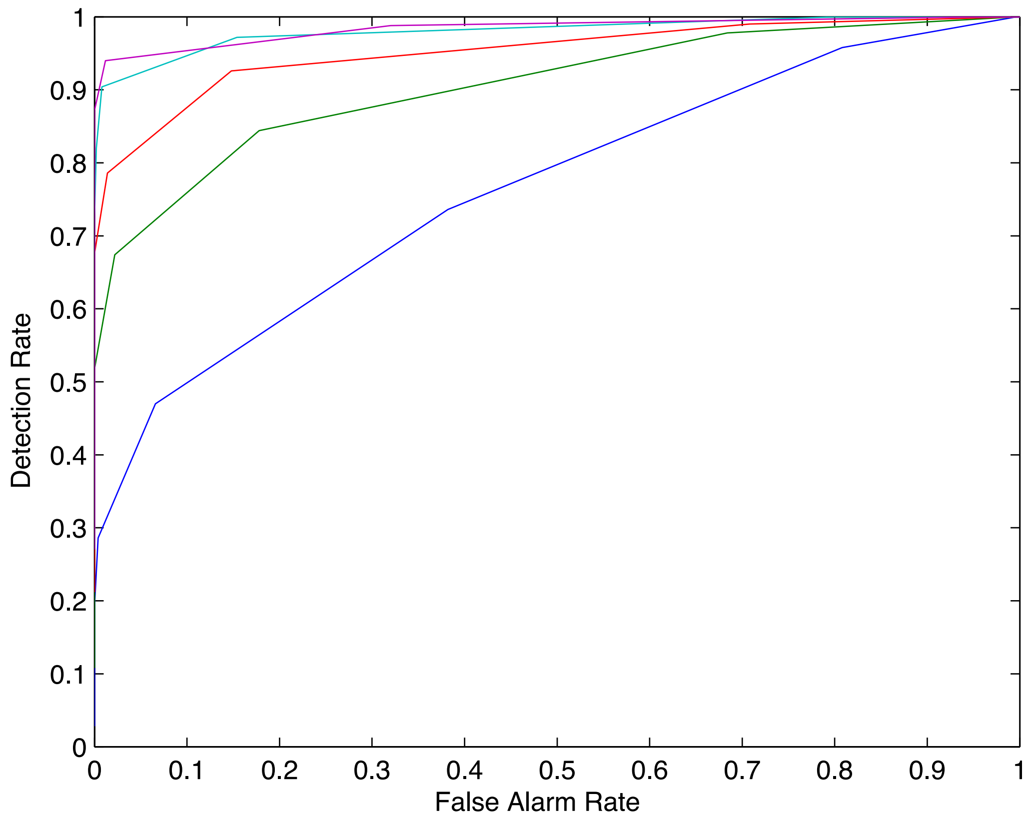
- Linear programming enables us to decouple the control problem into two steps:
  - First find the form of Q^*
  - Then decide upon the norm of Q^*
- Equivalent to deciding the vector direction of the signal, then the vector magnitude

Direction of Q^*



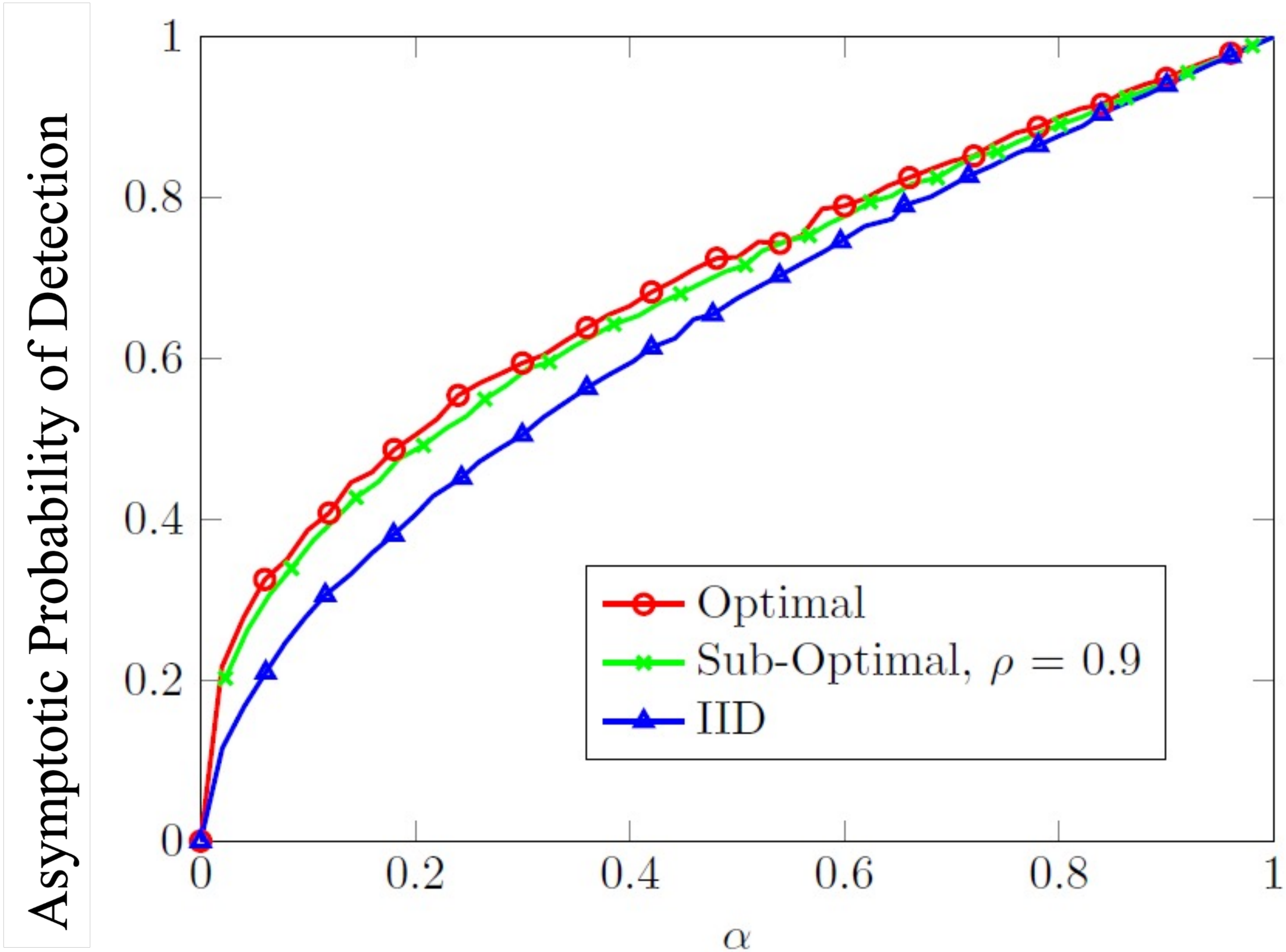
- **Comparison of the two detectors over time. The importance of optimization can be seen by performance improvement (note the change of scale by a factor of 10)**

Norm of Q^*



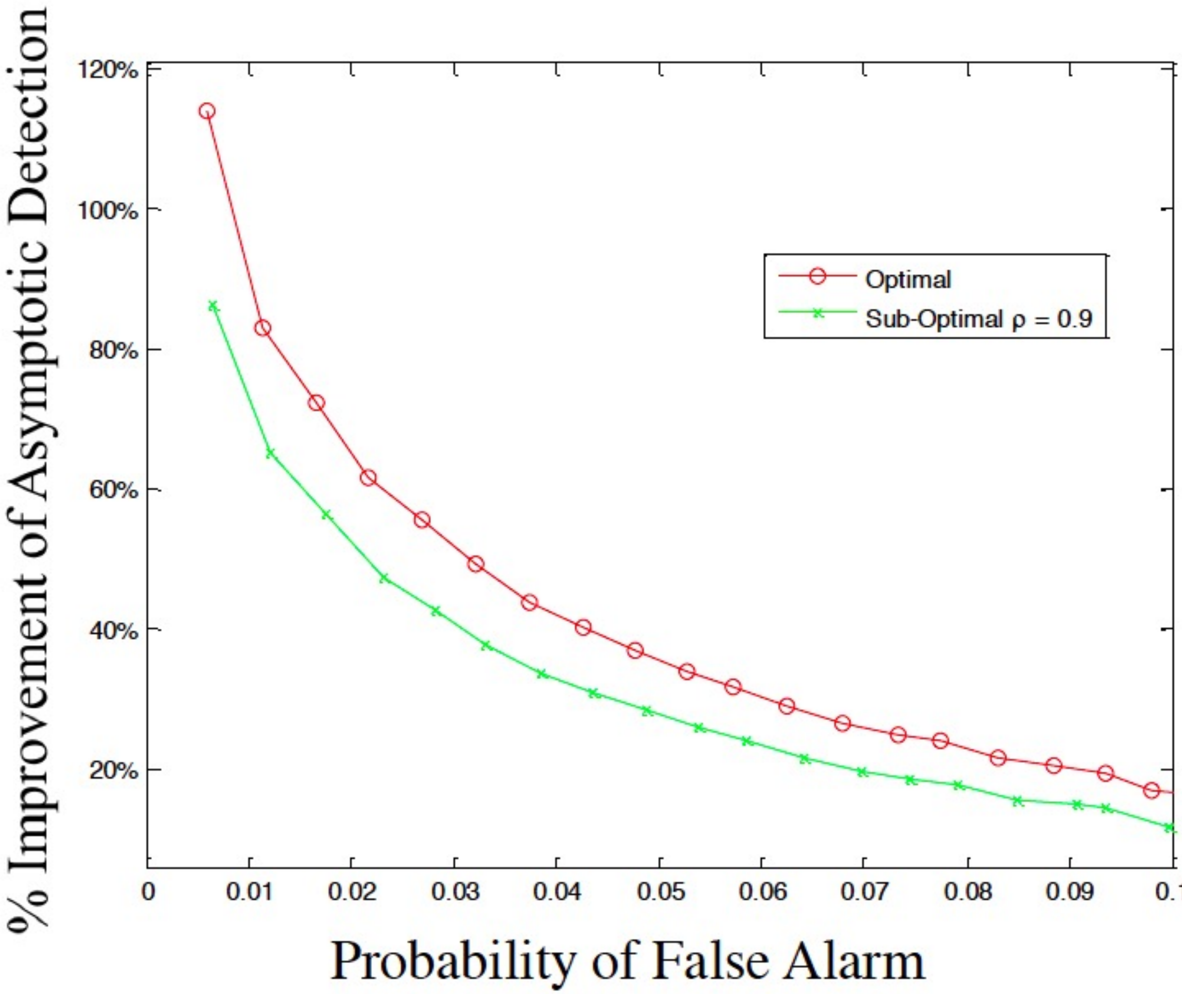
- **ROC Curve for detector, with Q increasing linearly from 0.2 to 1 times the maximum value**

Non I.I.D. case



Probability of False Alarm

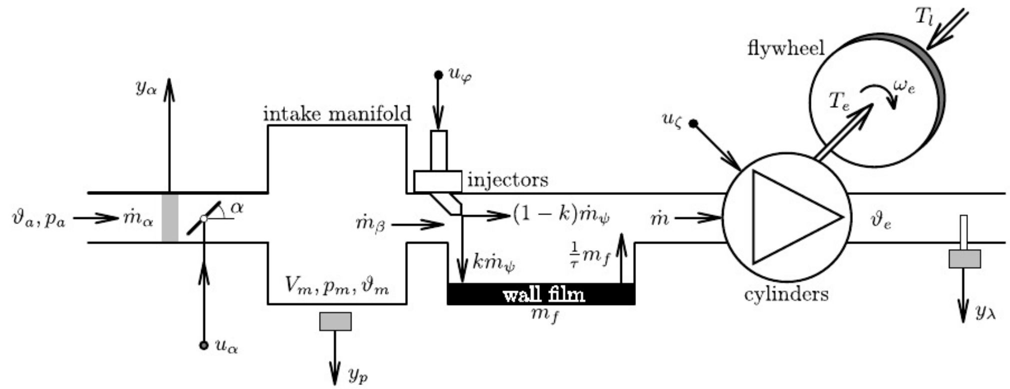
Improvement over IID is actually sizeable at low false alarm rates

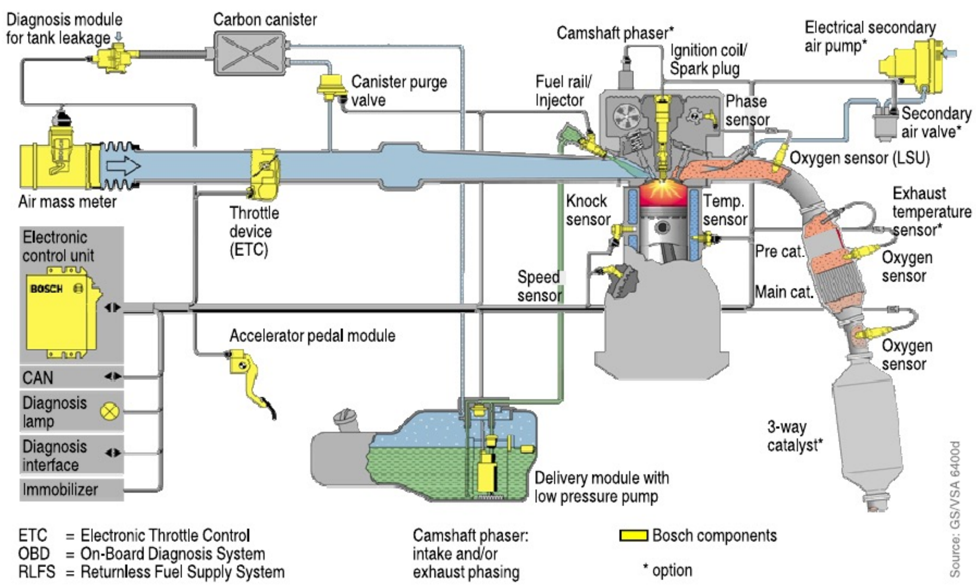


Internal Combustion (IC) Engine

$$\frac{dp_m(t)}{dt} = \frac{R\theta_m}{V_d}\left(\dot{m}_\alpha(t) - \dot{m}_\beta(t)\right)$$

$$\frac{d\omega_e(t)}{dt} = \frac{1}{\theta_e}\left(T_e(t) - T_l(t)\right)$$



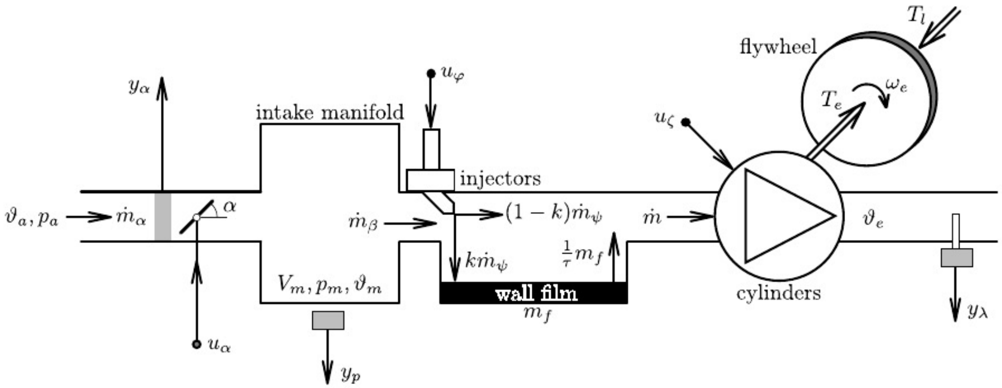


1. **Throttle body block**
2. **Intake manifold block**
3. Injection block
4. wall-wetting block
5. Gas exchange block
6. **Combustion and torque generation**
7. Engine inertia block
8. Gas transport block

Cruise control problem

Guzzella, L., & Onder, C. (2009). *Introduction to modeling and control of internal combustion engine systems*. Springer Science & Business Media.

Nonlinear Model



$$\frac{dp_m(t)}{dt} = \frac{R\theta_m}{V_d}\Bigg(A_\alpha(t)\frac{p_a}{\sqrt{2R\theta_a}} - \Big(\frac{p_m(t)}{R\theta_m}(\gamma_0 + \gamma_1\omega_e(t) + \gamma_2\omega_e^2(t))$$

$$\Big(\frac{V_c + V_d}{V_d} - \frac{V_c}{V_d}\big(\frac{p_{out}}{p_m}\big)^{\frac{1}{\kappa}}\Big)\frac{V_d\omega_e(t)}{4\pi}\frac{\alpha}{\alpha+1}\Big)\Bigg)$$

$$\frac{d\omega_e(t)}{dt} = \frac{1}{\theta_e}\Bigg[\Bigg(\Big(\eta_0 + \eta_1\omega_e(t)\Big)\frac{H_f . p_m(t)}{R\theta_m}(\gamma_0 + \gamma_1\omega_e(t) + \gamma_2\omega_e^2(t))$$

$$\Big(\frac{V_c + V_d}{V_d} - \frac{V_c}{V_d}\big(\frac{p_{out}}{p_m}\big)^{\frac{1}{\kappa}}\Big).\frac{V_d}{\alpha+1}$$

$$-\left(\beta_0 + \beta_2\omega_e^2(t) + (p_{out} - p_m(t))\right)\frac{V_d}{4\pi}\Bigg) - T_l(t)\Bigg]$$

- $R = 287\ [J/kgK]$: Gas constant air
- $T_a = 298\ [K]$: Ambient Temperature
- $P_a = 10^5\ [Pa]$: Ambient Pressure
- $V_m = 5.8 \times 10^{-3}\ [m^3]$: Volume Intake manifold
- $T_m = 340\ [K]$: Temperature air in manifold
- $\gamma_0 = 0.45$: Coefficient
- $\gamma_1 = 3.42 \times 10^{-3}\ [s]$: Coefficient
- $\gamma_2 = -7.7 \times 10^{-6}\ [s^2]$: Coefficient
- $V_c = 0.277 \times 10^{-3}\ [m^3]$: Compression volume
- $V_d = 2.77 \times 10^{-3}\ [m^3]$: Displacement
- $\kappa = 1.35$: isentropic exponent air
- $P_e = 1e5\ [Pa]$ Back pressure exhaust mixture
- $\eta_0 = 0.16\ [kJ/kg]$ and $\eta_1 = 2.21 \times 10^{-3}\ [J/kg]$: Willians parameters
- $\beta_0 = 15.6\ [Nm]$ and $\beta_2 = 0.175 \times 10^{-3}\ [Nms^2]$
- $\theta_e = 0.2\ [kg/m^2]$: Engine inertia
- $\alpha = 14.70$
- $H_l = 45.8 \times 10^6$: Heating value

Linearized Model

Equilibrium Point

$$x_{eq} = \begin{bmatrix} p_{m_{eq}} & \omega_{e_{eq}} \end{bmatrix}^T = \begin{bmatrix} 6303 & 440 \end{bmatrix}^T \longrightarrow \dot{x} = Ax + Bu + w$$

$$A = \begin{bmatrix} -91257.9 & -23.0 \\ 628.9 & -2.3 \end{bmatrix} \quad B = \begin{bmatrix} 4.139 \times 10^9 \\ 0 \end{bmatrix} \quad C = \begin{bmatrix} 0 & 1 \end{bmatrix}$$

Discretization Ts=0.01s

$$\begin{aligned} x_{k+1} &= A_d x_k + B_d u_k + w_k \\ y_k &= C x_k + v_k \end{aligned}$$

LQG Control

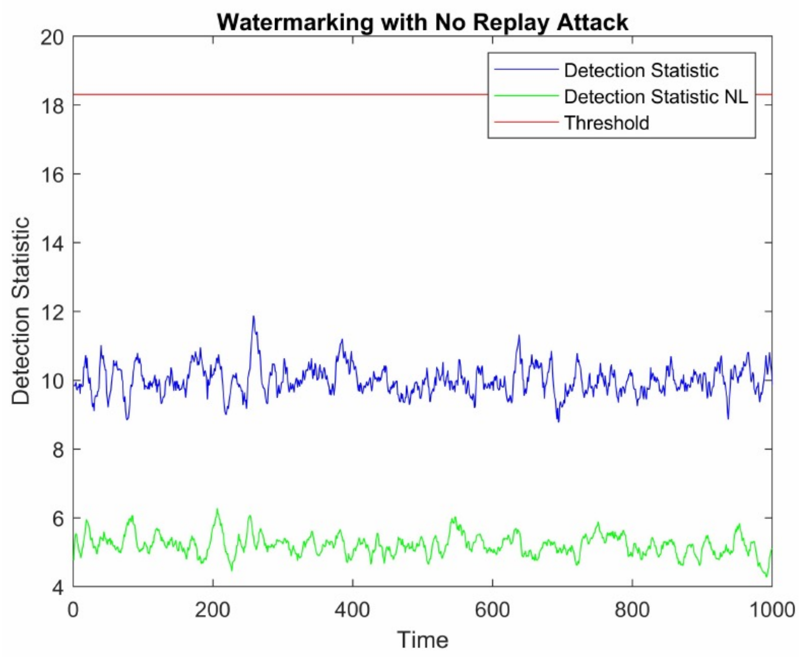
$$J = \lim_{N\to\infty} \mathbb{E} \frac{1}{N} \left[\sum_{k=0}^{N-1} (x_k^\top W x_k + u_k^\top U u_k) \right], \qquad u_k = L \hat{x}_{k|k}$$

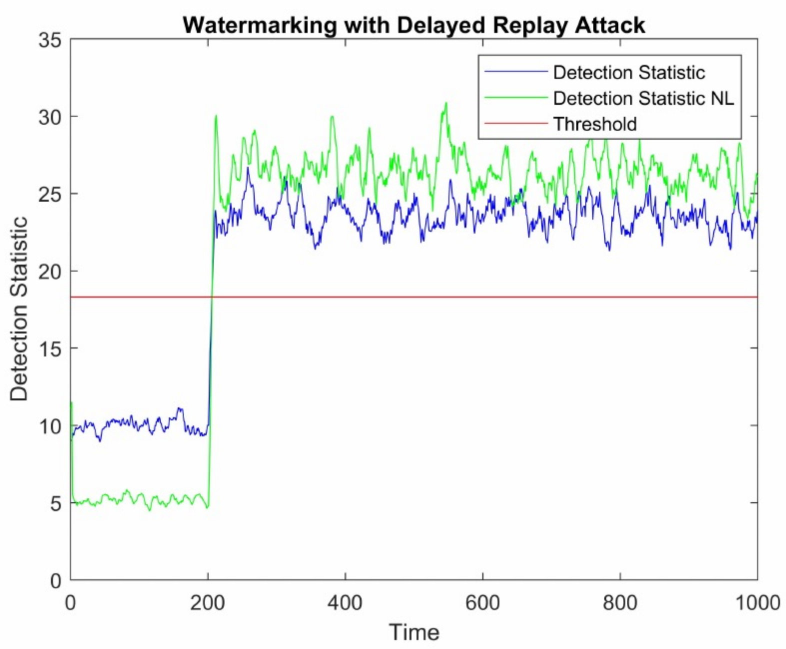
Kalman Filter

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K \left(y_k - C \hat{x}_{k|k-1} \right)$$

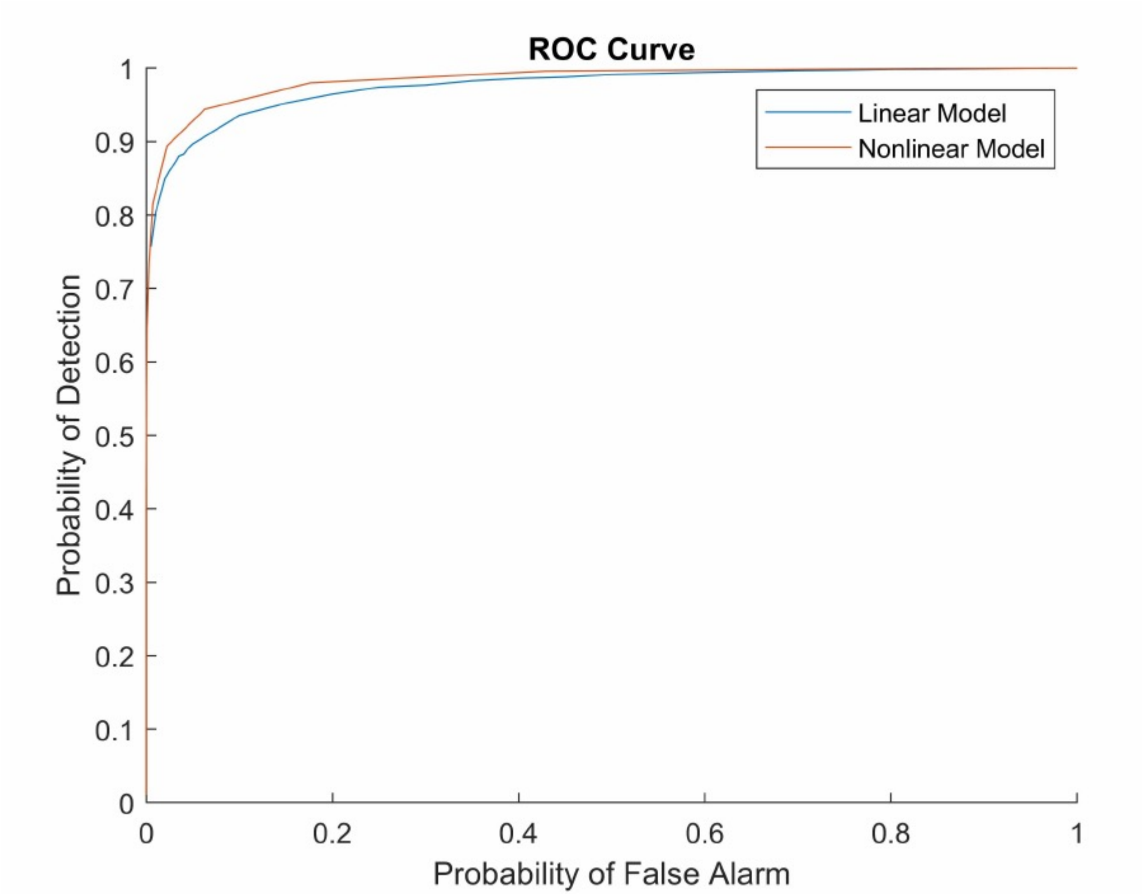
Simulations

Simulating the IC engine as linear system (blue),
Simulating the IC considering the nonlinear dynamics (green)

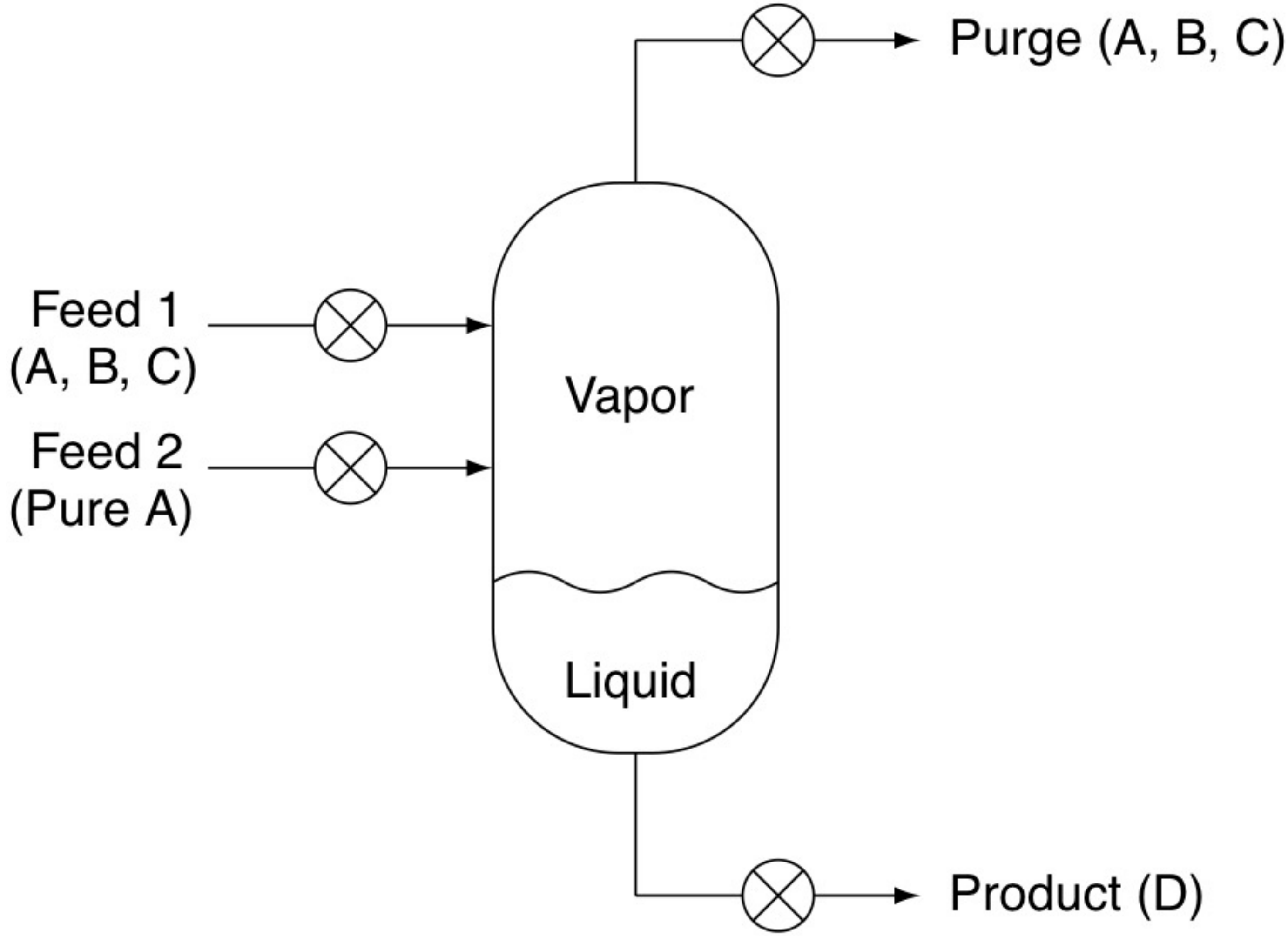




Simulations

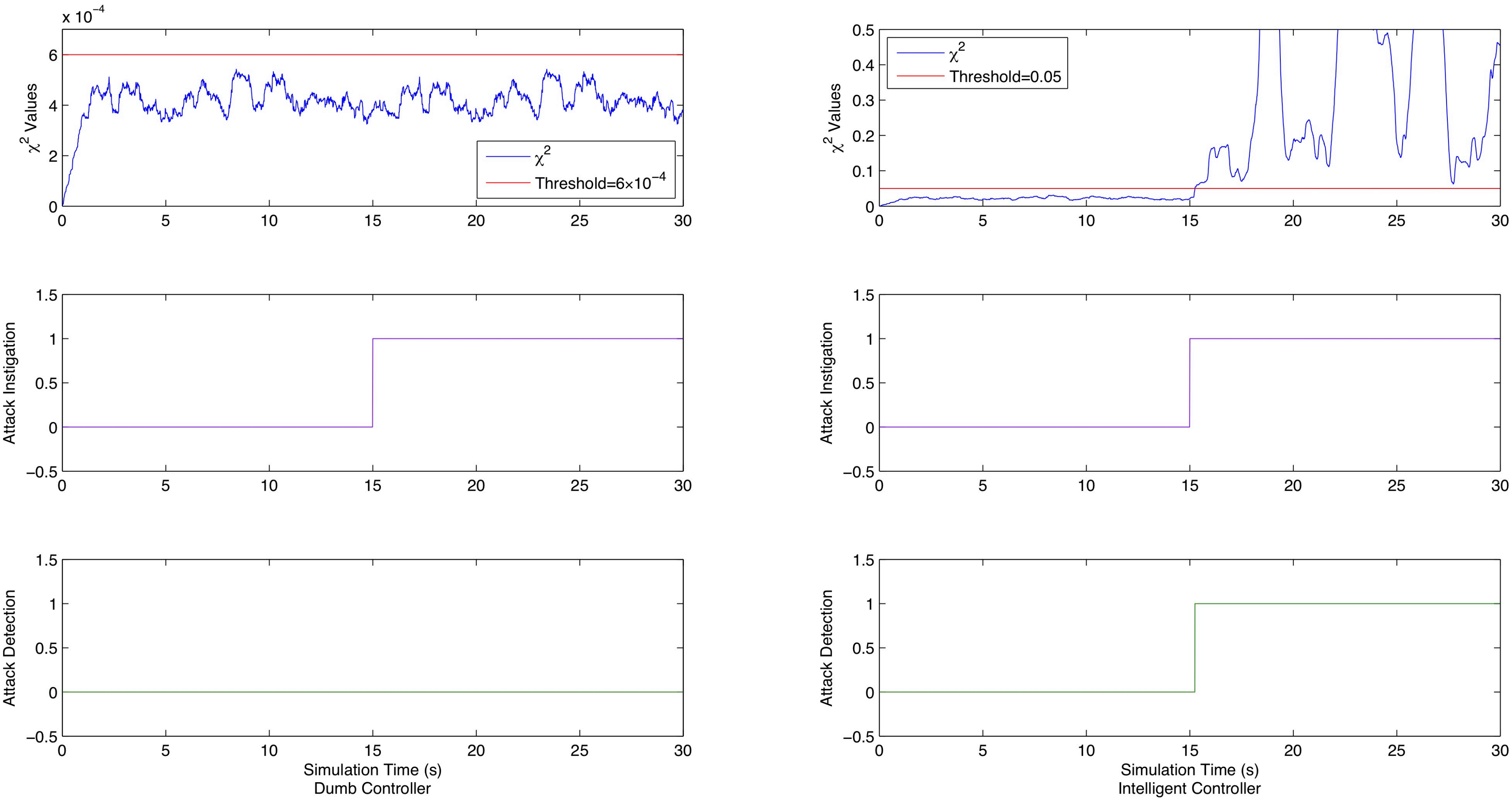


Chemical Plant (A + C → D)



Objectives: Maintain production rate by controlling valves
Minimize operating cost (function of purge loss of A and C)

Regular vs. Secure controller



Time for detection = 25 ms

The attack

The New York Times Reprints

January 15, 2011

Israeli Test on Worm Called Crucial in Iran Nuclear Delay

By WILLIAM J. BROAD, JOHN MARKOFF and DAVID E. SANGER

This article is by ***William J. Broad, John Markoff*** *and* ***David E. Sanger.***

The biggest single factor in putting time on the nuclear clock appears to be Stuxnet, the most sophisticated cyberweapon ever deployed.

The worm itself now appears to have included two major components. One was designed to send Iran's nuclear centrifuges spinning wildly out of control. Another seems right out of the movies: The computer program also secretly recorded what normal operations at the nuclear plant looked like, then played those readings back to plant operators, like a pre-recorded security tape in a bank heist, so that it would appear that everything was operating normally while the centrifuges were actually tearing themselves apart.

The counterattack

The New York Times Reprints

This copy is for your personal, noncommercial use only. You can order pr
distribution to your colleagues, clients or customers here or use the "Repr
article. Visit www.nytreprints.com for samples and additional information. C

January 15, 2011

Israeli Test on Worn
Iran Nuclear Delay

By WILLIAM J. BROAD, JOHN MARKOFF and DAVID

This article is by ***William J. Broad, John M***

The biggest single factor in putting time on the nu
sophisticated cyberweapon ever deployed.
The worm itself now appears to have included t
Iran's nuclear centrifuges spinning wildly out of
The computer program also secretly recorded w
like, then played those readings back to plant o
bank heist, so that it would appear that everythi
were actually tearing themselves apart.

Forty-Seventh Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
September 30 - October 2, 2009

Secure Control Against Replay Attacks

Yilin Mo, Bruno Sinopoli *†

This paper analyzes the effect of replay attacks on a control system. We assume an attacker wishes to disrupt the operation of a control system in steady state. In order to inject an exogenous control input without being detected the attacker will hijack the sensors, observe and record their readings for a certain amount of time and repeat them afterwards while carrying out his attack. This is a very common and natural attack (we have seen numerous times intruders recording and replaying security videos while performing their attack ···

Watermarking Challenges

- Can we extend watermarking approach to other attack models where the system model is known.

- **Challenge 1**
  - The inputs (not just the watermark), must be kept secret.
  - Attacker could observe u_k and simulate output to system

- **Challenge 2**
  - The attacker can subtract his influence on the system

$$x_{k+1} = Ax_k + B(u_k^* + \Delta u_k + u_k^a) + w_k$$
$$y_k = Cx_k + v_k$$

$$x_{k+1}^a = Ax_k^a + Bu_k^a$$
$$y_k^a = Cx_k^a$$

$$\longrightarrow$$

$$x_{k+1} = Ax_k + B(u_k^* + \Delta u_k) + w_k$$
$$y_k = Cx_k + v_k$$

System ID Attack

- Idea: Use the system model as our secret.

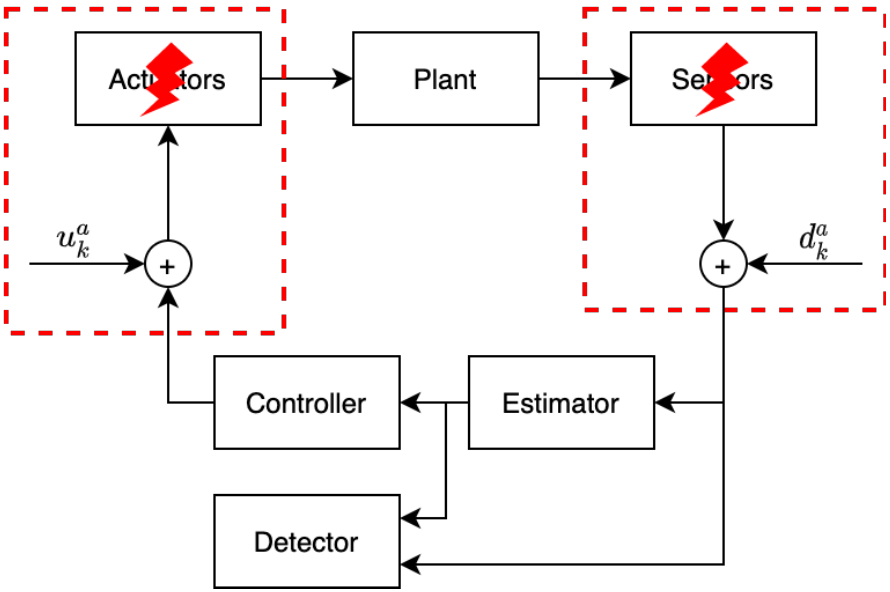
- **Attacker Capabilities**
  - Attacker can read all sensor and actuator channels.
  - Attacker can violate the integrity of all sensor and actuator channels.

- **Attack Strategy**

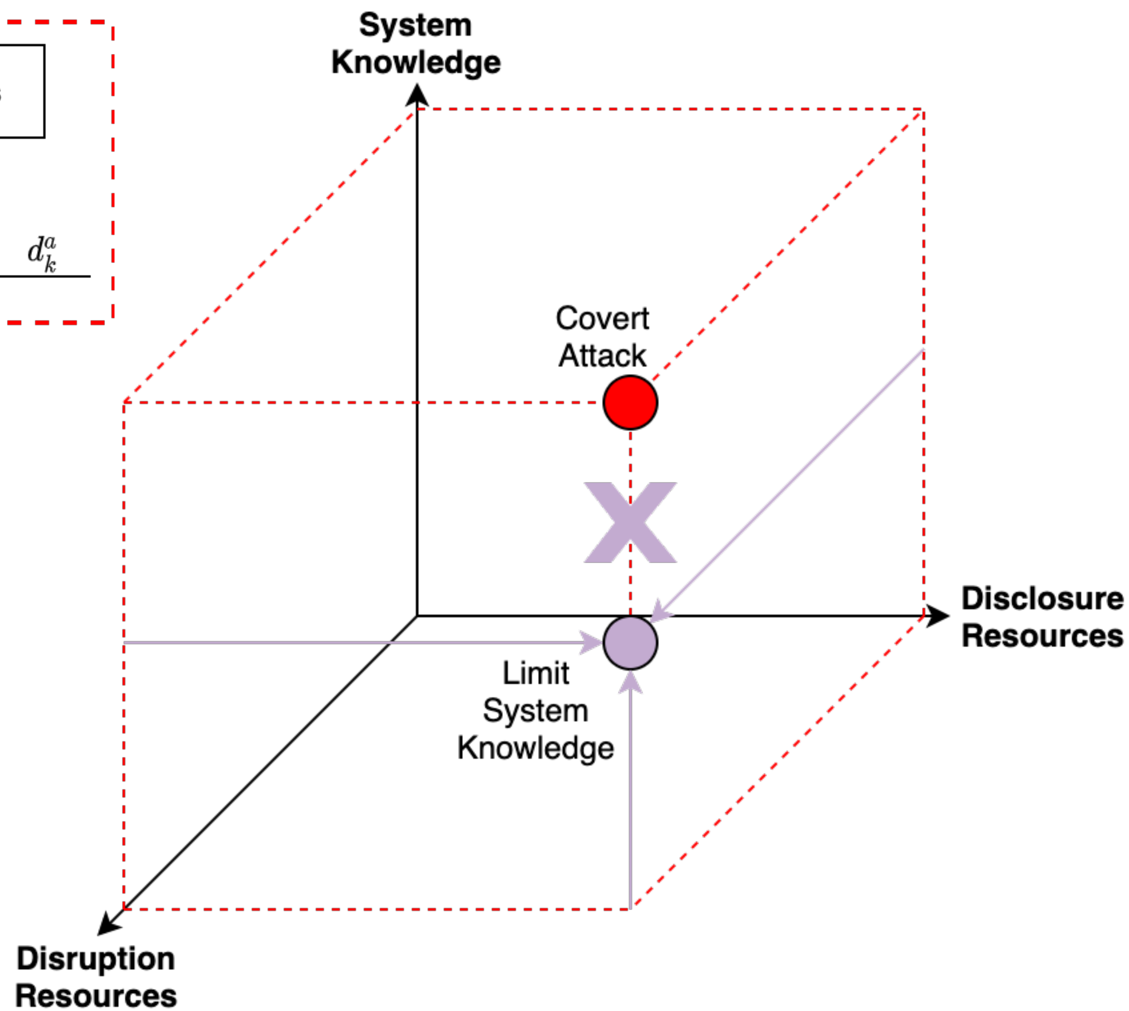
1) Use knowledge of inputs and outputs to identify the system model.
2) Violate the integrity of sensors with "convincing" measurements.
3) Insert harmful inputs into system.

Moving Target Defense

Goal: limit the adversary's system knowledge



Covert Attack



Moving Target Approach

(Weerakkody and Sinopoli, 2015)

Goal: Design system to prevent identification

Challenge: Many existing methods for identifying systems

- Prediction Error Method
- Instrumental Variable Methods
- Subspace Based Approaches

Attacker does not need an exact working model of system

Approach: The Moving Target

Design system to be time varying so that the model changes before the attacker can perform adequate identification

Hybrid Moving Target Defense

- A cyber-physical "message authentication code" or perturbation introduced in the system dynamics
- Is effective in detecting more powerful covert attacks
- Introduces a tradeoff between detection and system performance

$$x_{k+1} = A_k x_k + B_k u_k + w_k$$
$$y_k^a = C_k x_k + d_k^a + v_k$$

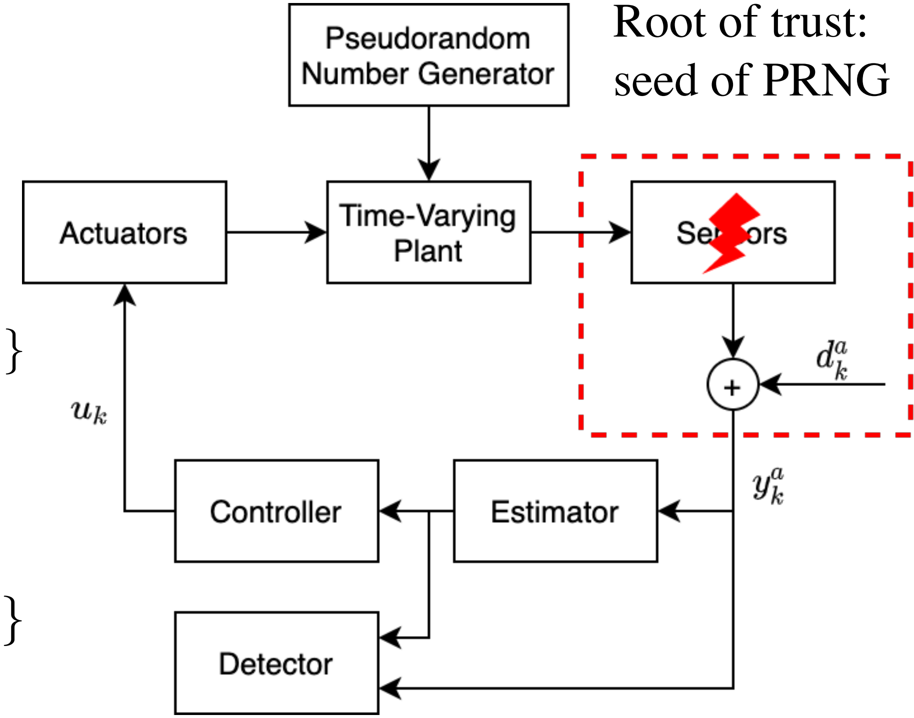
$(A_k, B_k, C_k) \in \Upsilon$

$\Upsilon \triangleq \{(A(1), B(1), C(1)), \cdots, (A(l), B(l), C(l))\}$

$\mathcal{I}_k^D$: defender's information
$\mathcal{I}_k^A$: attacker's information

$$\mathcal{I}_k^D \triangleq \{A_{0:k}, B_{0:k}, C_{0:k}, u_{0:k}, y_{0:k}^a, f(w_k, v_k)\}$$
$$\mathcal{I}_k^A \triangleq \{\Upsilon, u_{0:k}, y_{0:k}, d_{0:k}^a, f(w_k, v_k)\}$$

Root of trust: seed of PRNG



Extended Moving Target Defense

- Motivation: watermarking is ineffective against model-aware attackers
- Goal: design the system in a way that prevents system identification
- Approach: add an auxiliary system with time-varying dynamics to authenticate the original system

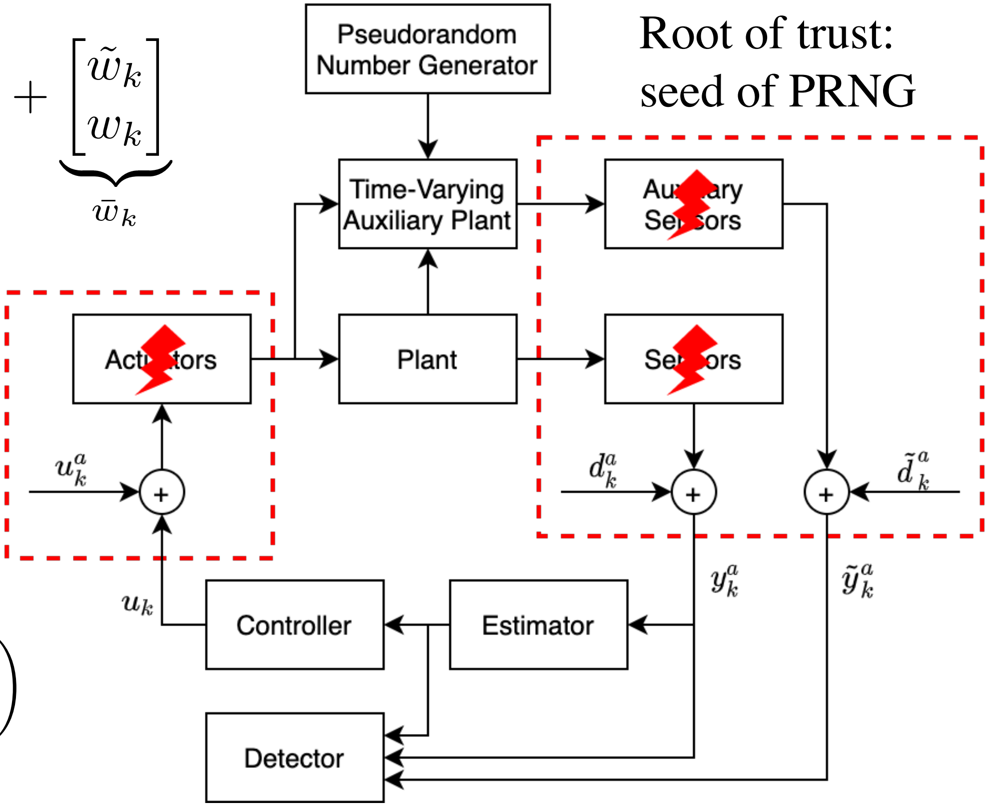
$$\underbrace{\begin{bmatrix}\tilde{x}_{k+1}\\ x_{k+1}\end{bmatrix}}_{\bar{x}_{k+1}} = \underbrace{\begin{bmatrix}\tilde{A} & \bar{A}_k\\ 0 & A\end{bmatrix}}_{\mathcal{A}_k} \underbrace{\begin{bmatrix}\tilde{x}_k\\ x_k\end{bmatrix}}_{\bar{x}_k} + \underbrace{\begin{bmatrix}\tilde{B}_k\\ B\end{bmatrix}}_{\mathcal{B}_k}(u_k + u_k^a) + \underbrace{\begin{bmatrix}\tilde{w}_k\\ w_k\end{bmatrix}}_{\bar{w}_k}$$

$$\underbrace{\begin{bmatrix}\tilde{y}_k^a\\ y_k^a\end{bmatrix}}_{\bar{y}_k^a} = \underbrace{\begin{bmatrix}\tilde{C} & \bar{C}_k\\ 0 & C\end{bmatrix}}_{\mathcal{C}_k} \underbrace{\begin{bmatrix}\tilde{x}_k\\ x_k\end{bmatrix}}_{\bar{x}_k} + \underbrace{\begin{bmatrix}\tilde{d}_k^a\\ d_k^a\end{bmatrix}}_{\bar{d}_k^a} + \underbrace{\begin{bmatrix}\tilde{v}_k\\ v_k\end{bmatrix}}_{\bar{v}_k}$$

auxiliary states: $\tilde{x}_k \in \mathbb{R}^{\tilde{n}}$
auxiliary sensors: $\tilde{y} \in \mathbb{R}^{\tilde{m}}$

$$\bar{w}_k \sim \mathcal{N}\left(0, \underbrace{\begin{bmatrix}\tilde{Q} & 0\\ 0 & Q\end{bmatrix}}_{\mathcal{Q}\succ 0}\right) \qquad \bar{v}_k \sim \mathcal{N}\left(0, \underbrace{\begin{bmatrix}\tilde{R} & \tilde{R}_{12}\\ \tilde{R}_{12}^T & R\end{bmatrix}}_{\mathcal{R}\succ 0}\right)$$

Root of trust: seed of PRNG



$$\mathcal{I}_k^D \triangleq \{A, B, C, \tilde{A}, \bar{A}_{0:k}, \tilde{B}_{0:k}, \tilde{C}, \bar{C}_{0:k}, u_{0:k}, \bar{y}_{0:k}^a, f(\bar{w}_k, \bar{v}_k)\}$$

$$\mathcal{I}_k^A \triangleq \{A, B, C, \tilde{A}, \tilde{C}, f(\bar{A}, \tilde{B}, \bar{C}), u_{0:k}, u_{0:k}^a, \bar{y}_{0:k}, \bar{d}_{0:k}^a, f(\bar{w}_k, \bar{v}_k)\}$$

Nonlinear Moving Target Defense

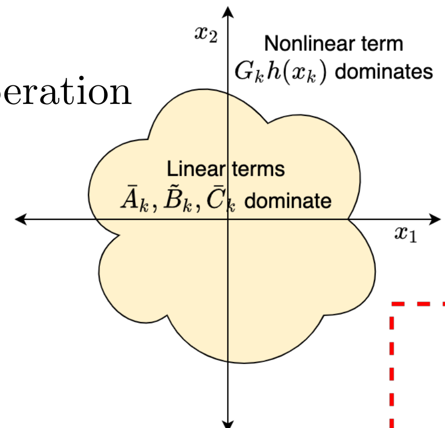
- Motivation: the sensor measurements of the extended moving target still reveal some information about the system dynamics
- Goal: limit this information available to an attacker
- Approach: introduce nonlinearities into the auxiliary sensor measurements

$G_k h(x_k) \to 0$ under normal operation

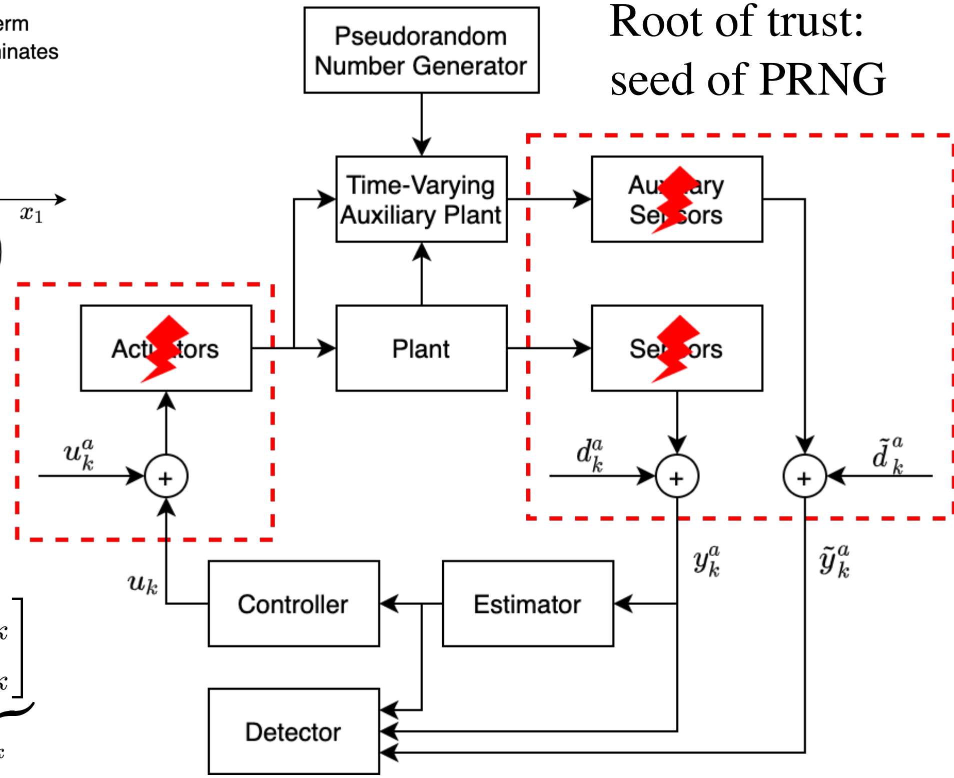
$G_k h(x_k) \to \infty$ under attack

$G_k \in \mathbb{R}^{\tilde{m} \times n}$

$h(x_k)$ is an element-wise mapping from $\mathbb{R}^n \to \mathbb{R}^n$



Root of trust: seed of PRNG



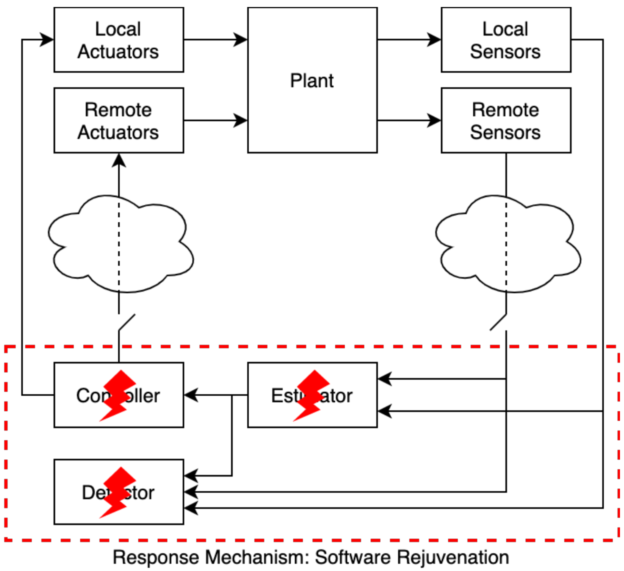
$$\underbrace{\begin{bmatrix} \tilde{x}_{k+1} \\ x_{k+1} \end{bmatrix}}_{\bar{x}_{k+1}} = \underbrace{\begin{bmatrix} \tilde{A} & \bar{A}_k \\ 0 & A \end{bmatrix}}_{\mathcal{A}_k} \underbrace{\begin{bmatrix} \tilde{x}_k \\ x_k \end{bmatrix}}_{\bar{x}_k} + \underbrace{\begin{bmatrix} \tilde{B}_k \\ B \end{bmatrix}}_{\mathcal{B}_k} (u_k + u_k^a) + \underbrace{\begin{bmatrix} \tilde{w}_k \\ w_k \end{bmatrix}}_{\bar{w}_k}$$

$$\underbrace{\begin{bmatrix} \tilde{y}_k^a \\ y_k^a \end{bmatrix}}_{\bar{y}_k^a} = \underbrace{\begin{bmatrix} \tilde{C} & \bar{C}_k \\ 0 & C \end{bmatrix}}_{\mathcal{C}_k} \underbrace{\begin{bmatrix} \tilde{x}_k \\ x_k \end{bmatrix}}_{\bar{x}_k} + \begin{bmatrix} G_k h(x_k) \\ 0 \end{bmatrix} + \underbrace{\begin{bmatrix} \tilde{d}_k^a \\ d_k^a \end{bmatrix}}_{\bar{d}_k^a} + \underbrace{\begin{bmatrix} \tilde{v}_k \\ v_k \end{bmatrix}}_{\bar{v}_k}$$

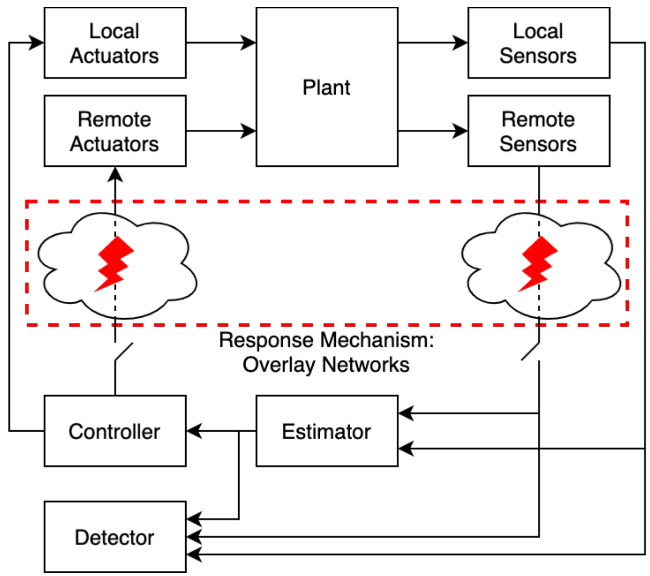
$$\mathcal{I}_k^D \triangleq \{A, B, C, \tilde{A}, \bar{A}_{0:k}, \tilde{B}_{0:k}, \tilde{C}, \bar{C}_{0:k}, G_{0:k}, \text{nonlinear function } h, u_{0:k}, \bar{y}_{0:k}^a, f(\bar{w}_k, \bar{v}_k)\}$$

$$\mathcal{I}_k^A \triangleq \{A, B, C, \tilde{A}, \tilde{C}, f(\bar{A}, \tilde{B}, \bar{C}), f(G), \text{nonlinear function } h, u_{0:k}, u_{0:k}^a, \bar{y}_{0:k}, \bar{d}_{0:k}^a, f(\bar{w}_k, \bar{v}_k)\}$$

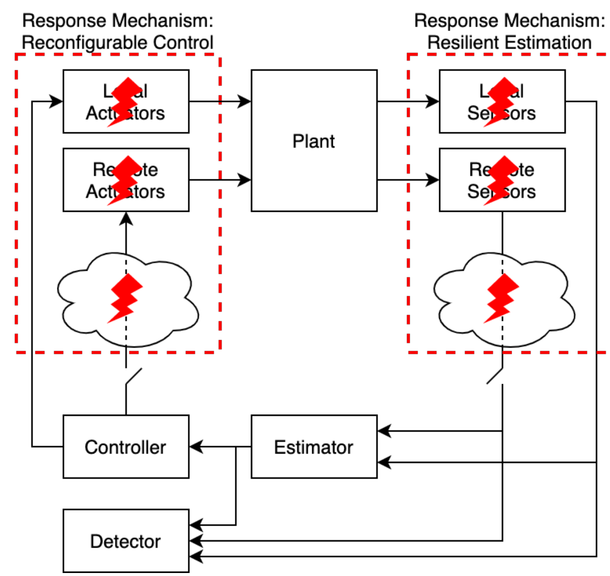
Overview of Resilience Strategies



Response Mechanisms for Control Software Attacks



Response Mechanisms for Communication Attacks

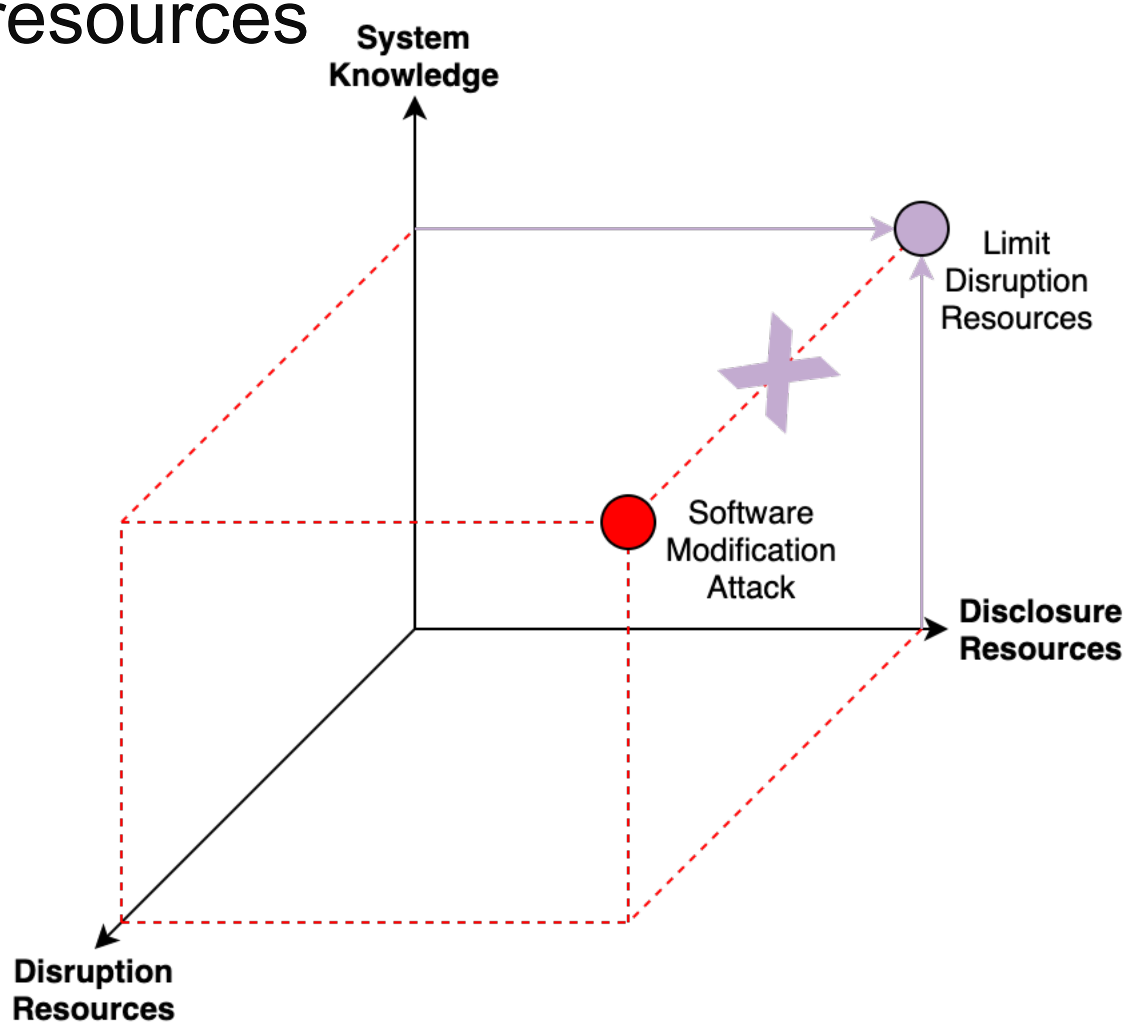


Response Mechanisms for Physical and Communication Attacks

- Each scenario includes components that can:
  - Constantly be trusted for all time
  - Occasionally be trusted for certain periods of time
- Goal: leverage the periods of time when the occasionally trusted components are secure to recover the system from attacks

Software Rejuvenation

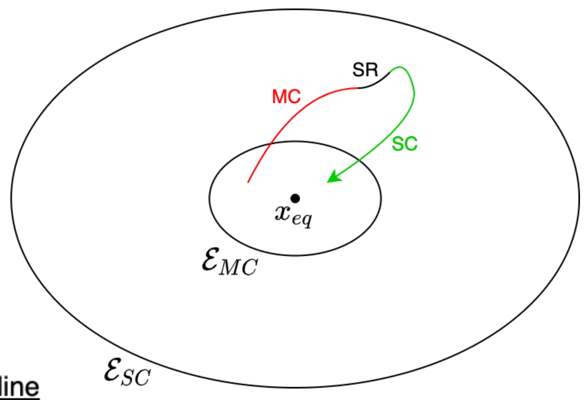
Goal: periodically limit the adversary's disruption resources



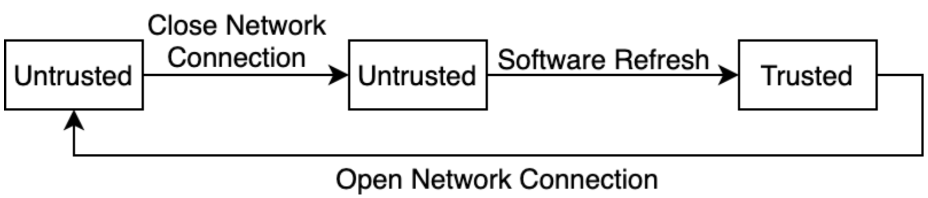
CPS Software Rejuvenation

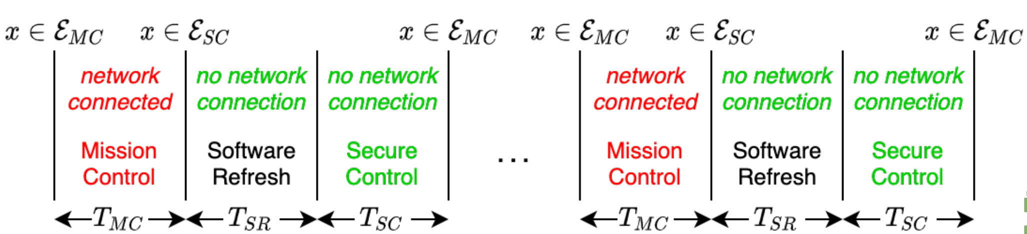
- The system is normally connected to the network to receive and transmit critical mission data
- Local information is sufficient for recovery

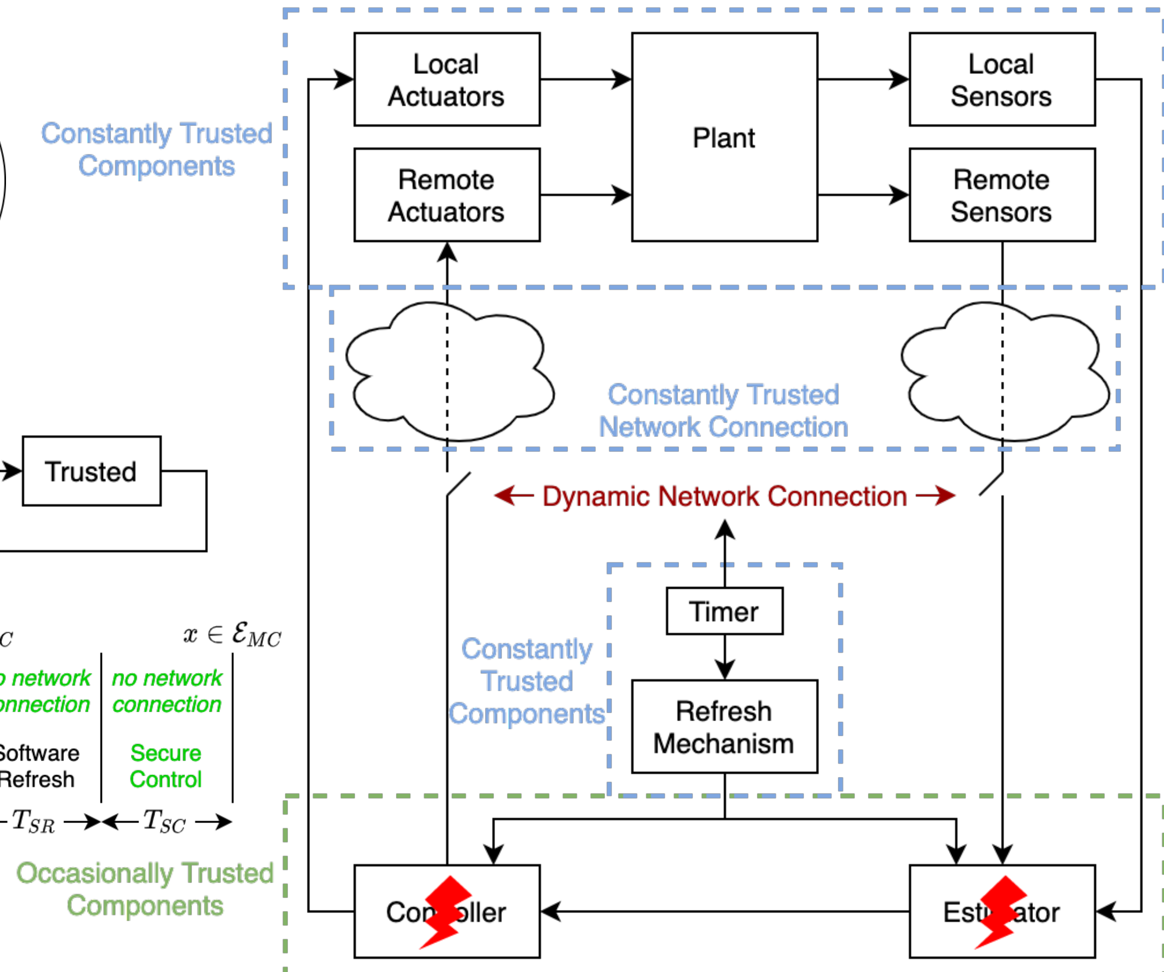
Root of trust: secure onboard hardware module



Trust Timeline



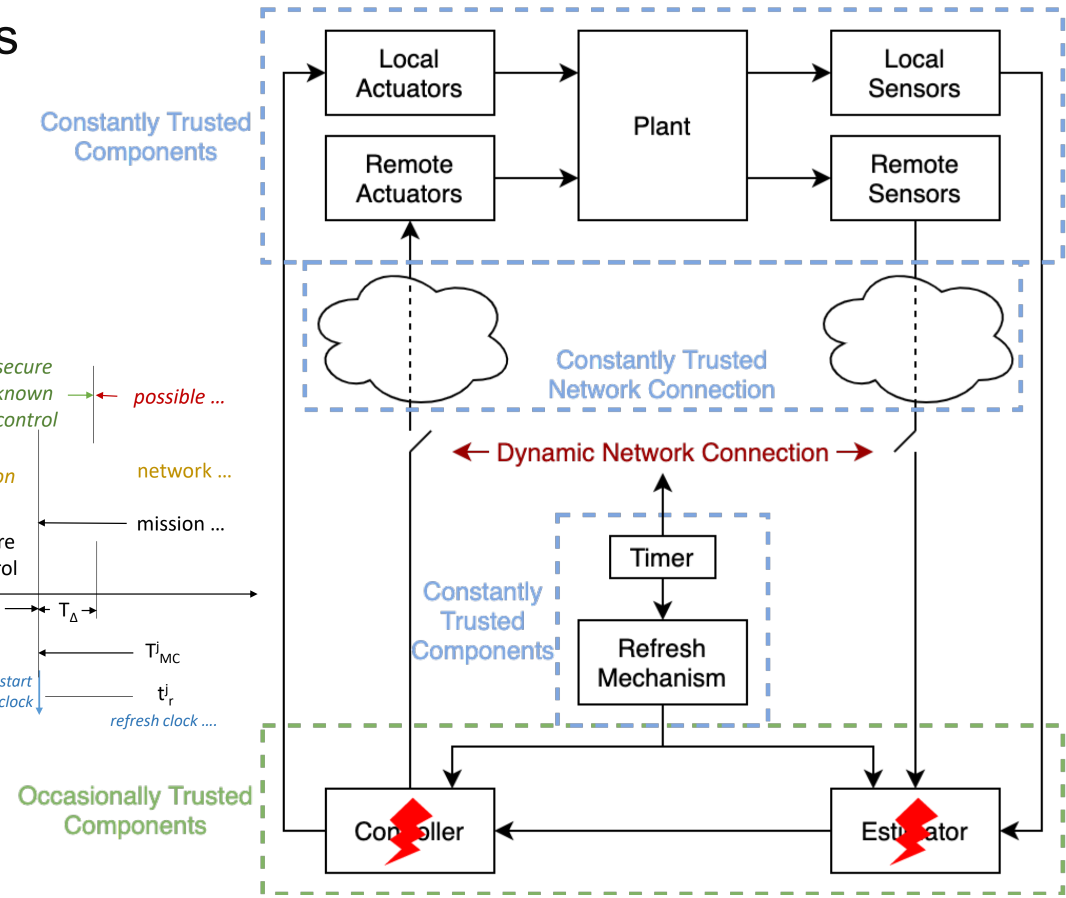


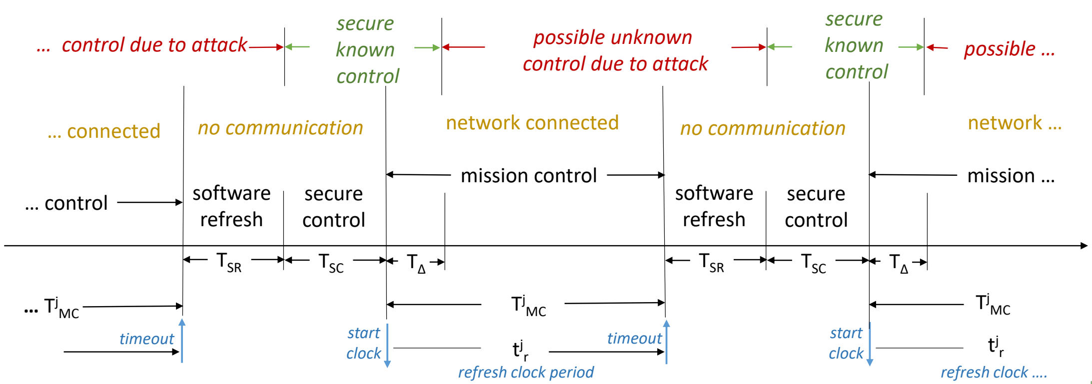


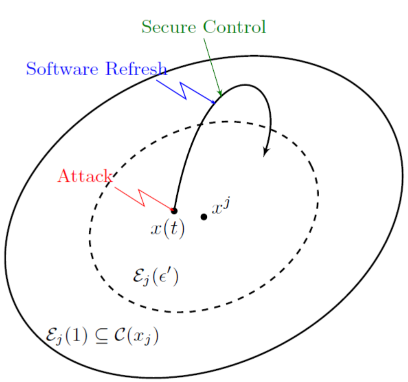
Software Rejuvenation: Environmental Constraints

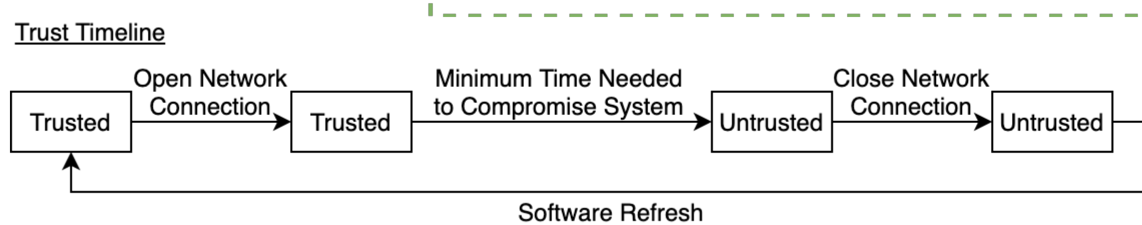
- Physical environmental constraints and persistent attacks may hinder reference tracking
- A secure recovery algorithm is needed to drive the system to a safer place

Root of trust: secure onboard hardware module





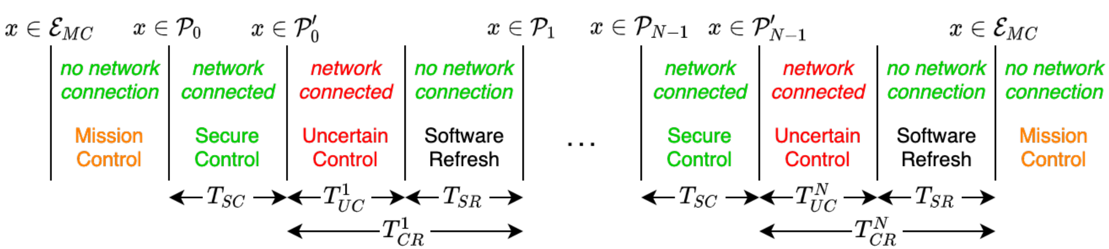


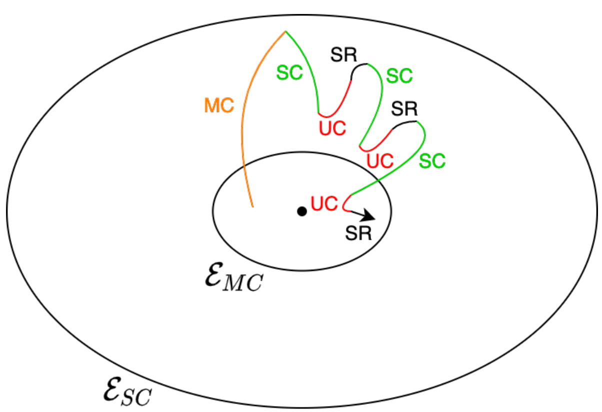


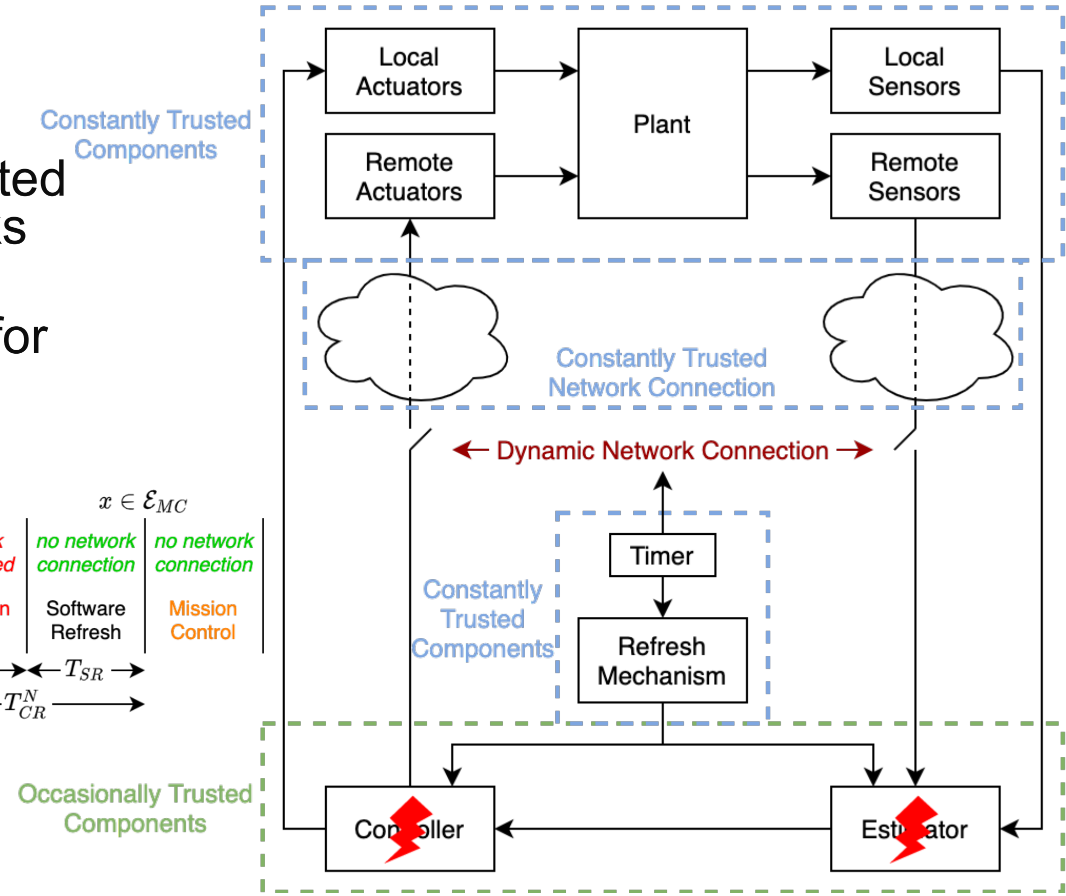
Complementary Software Rejuvenation

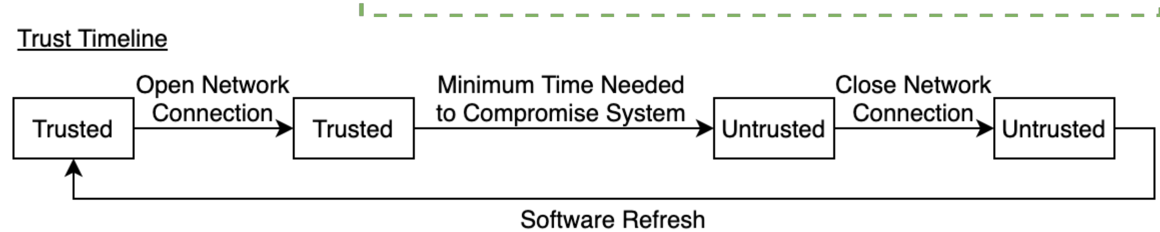
Root of trust: secure onboard hardware module

- The system is normally disconnected from the network to prevent attacks from occurring
- Remote information is necessary for reference tracking or recovering from dangerous disturbances





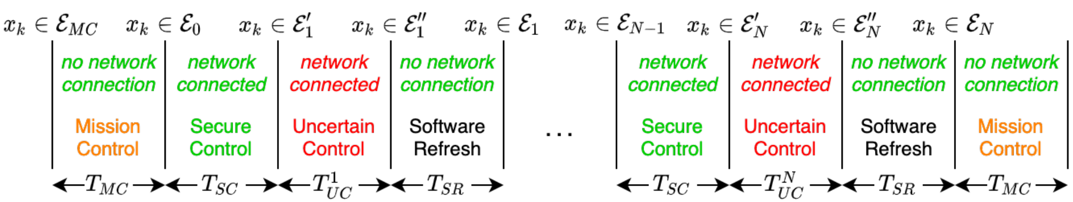


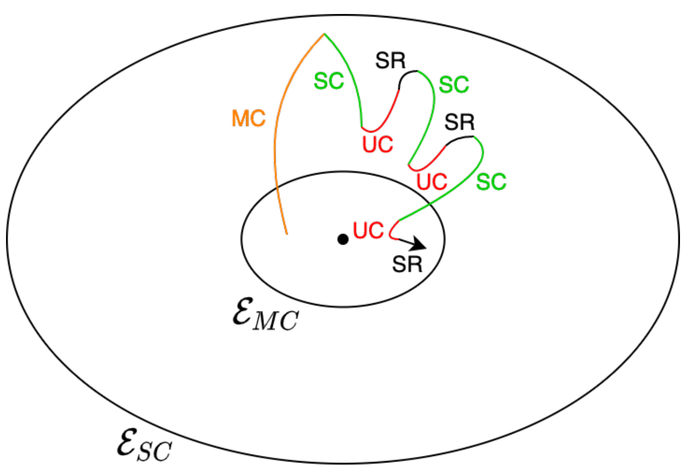


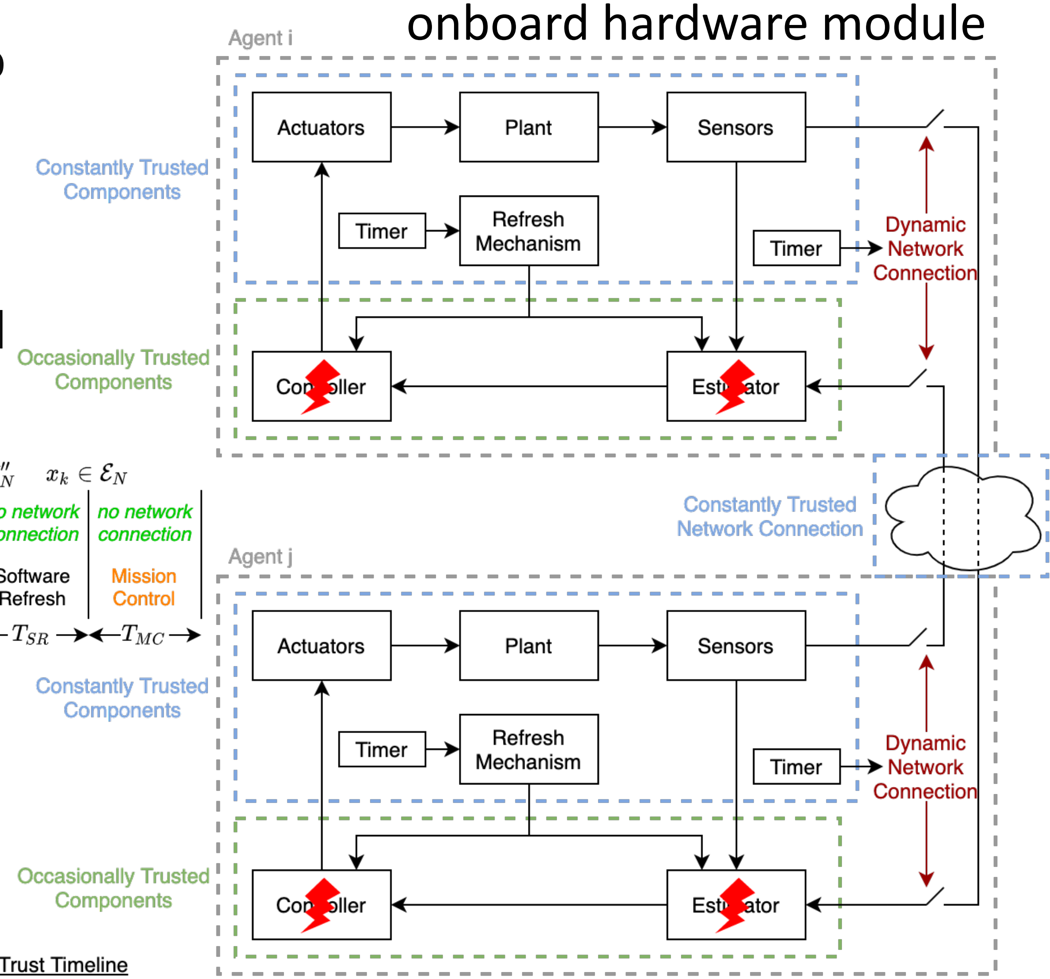
Decentralized Software Rejuvenation

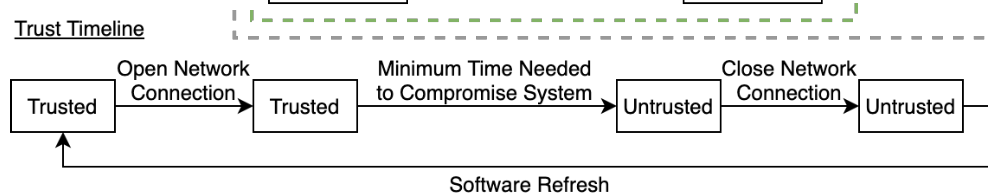
Root of trust: secure onboard hardware module

- Each agent is normally disconnected from the network to prevent attacks from occurring
- Decentralized systems require occasional communication between agents to ensure overall system safety



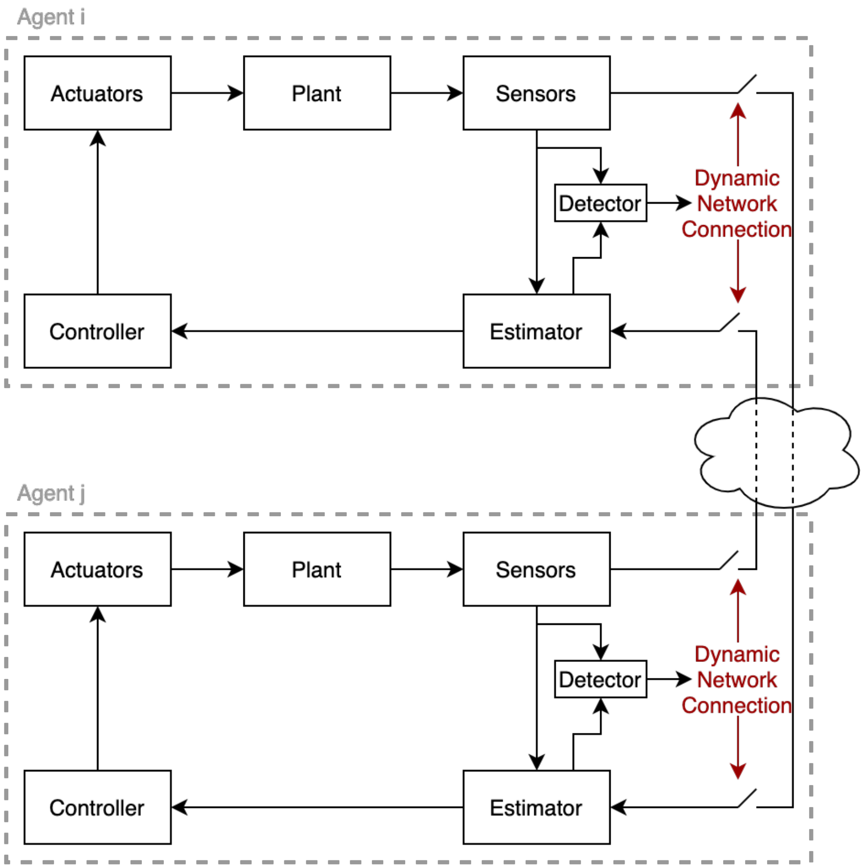






Decentralized Event-Triggered Control

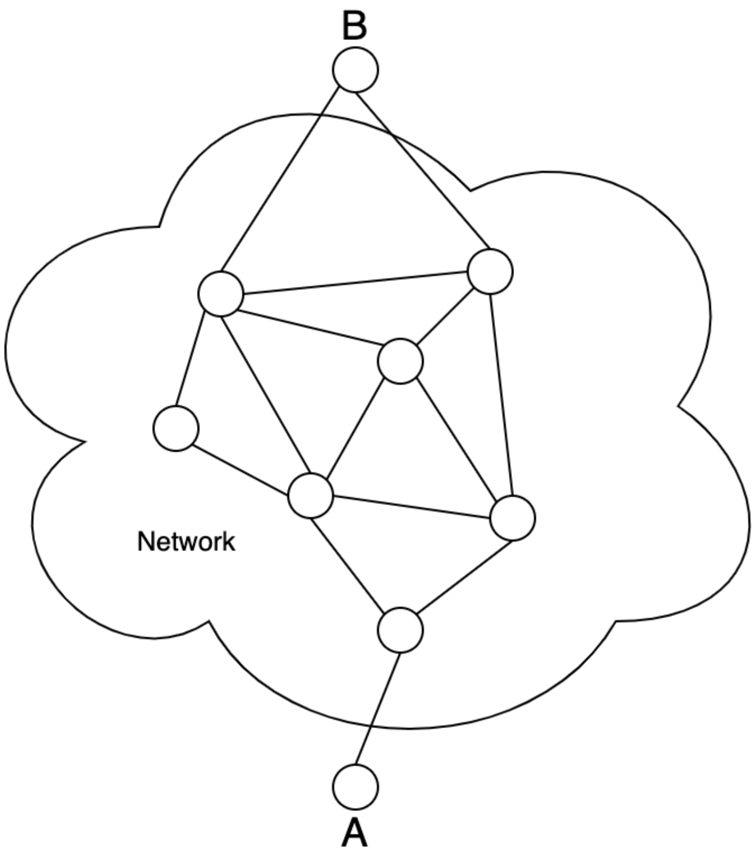
- Decentralized control systems require communication between agents to ensure overall safety and stability
- Communication results in
  - Connecting to the network and becoming vulnerable to malicious attacks
  - Increasing communication costs
- Intermittent network connections are therefore desirable

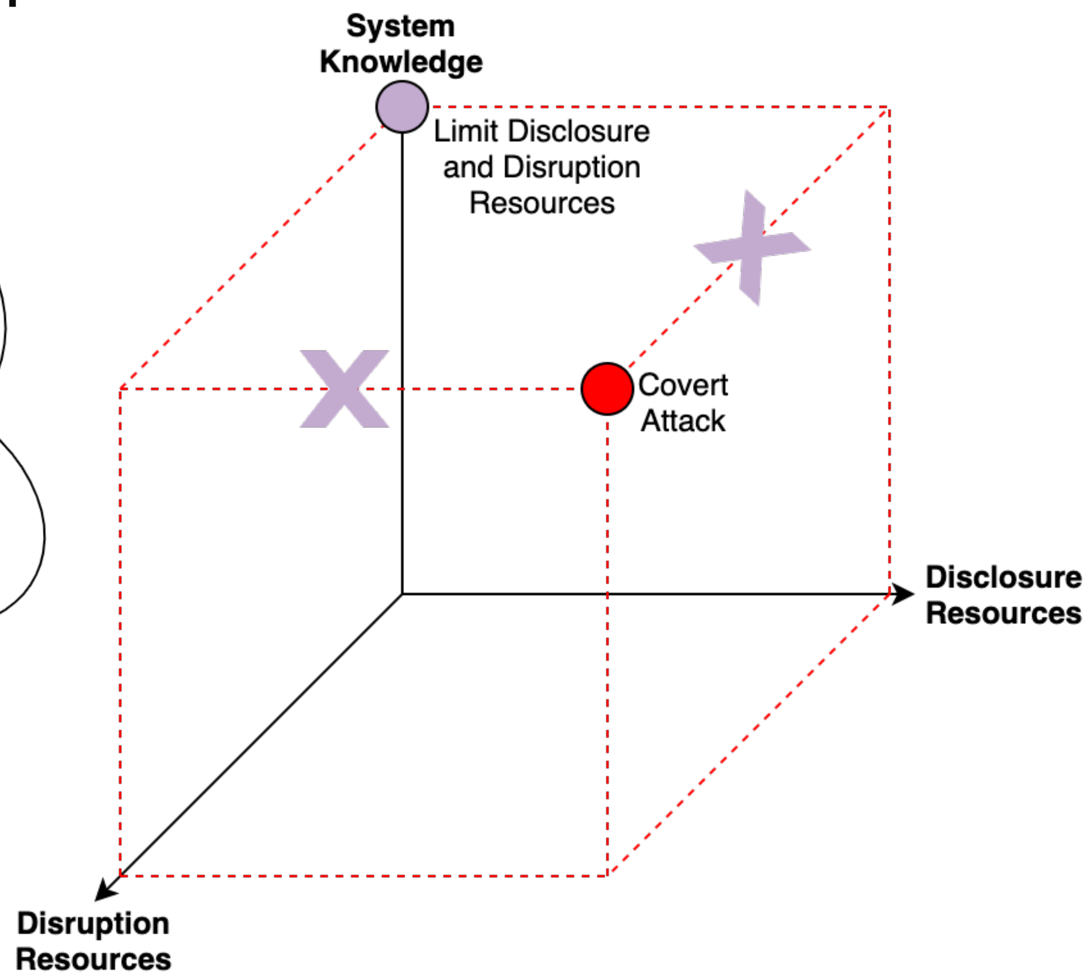


Goal: design a decentralized event-triggered network connection and communication protocol which ensures the stability of the overall system in attack-free scenarios

Resilient Overlay Networks

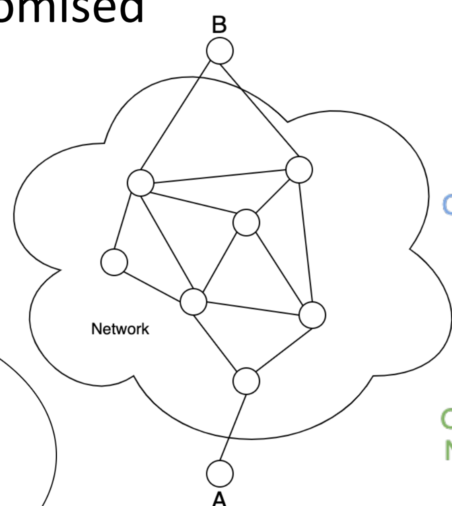
Goal: periodically limit the adversary’s disclosure and disruption resources

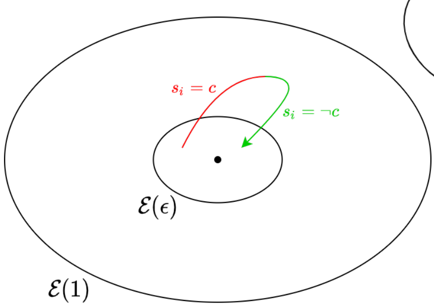


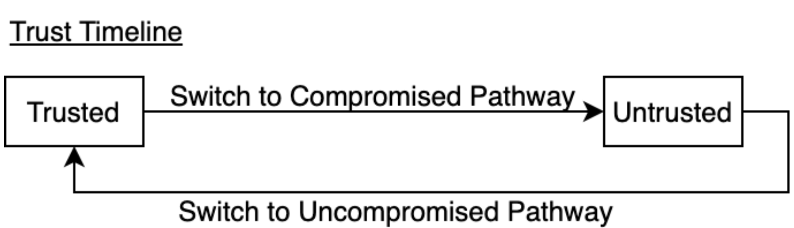


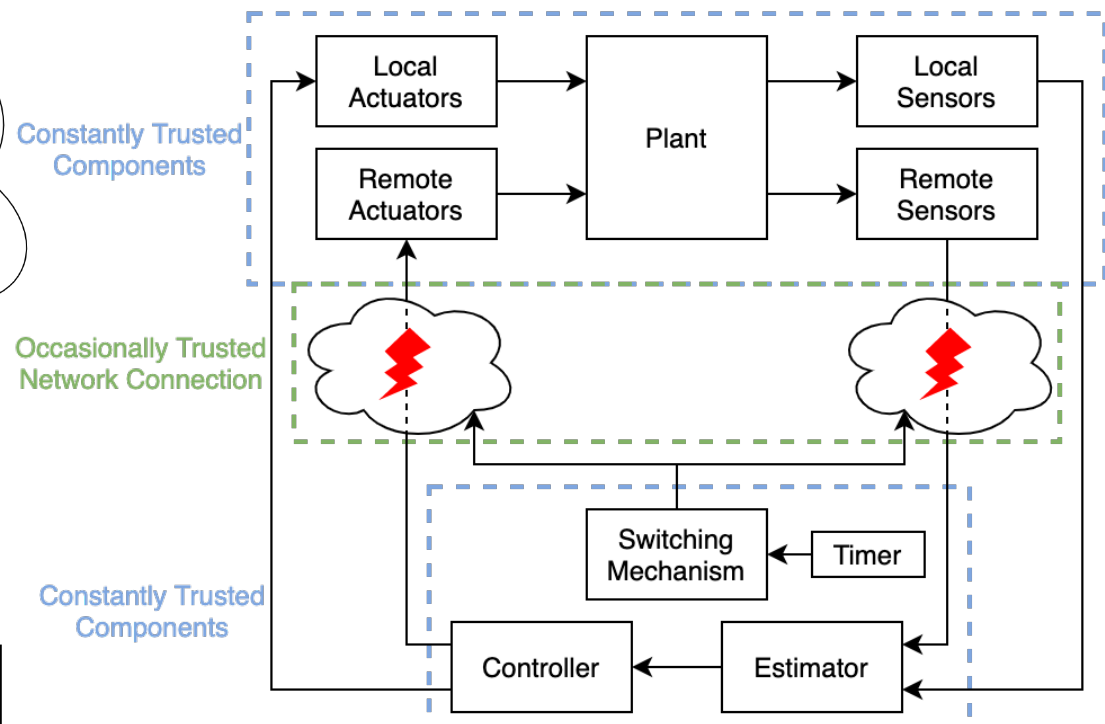
Resilient Overlay Networks

- The communication pathway over which data is sent is periodically switched to avoid continually sending data over a compromised pathway
- Is a prevention mechanism against man-in-the-middle and denial of service attacks
- Ensures safety when up to a certain percentage of pathways are compromised

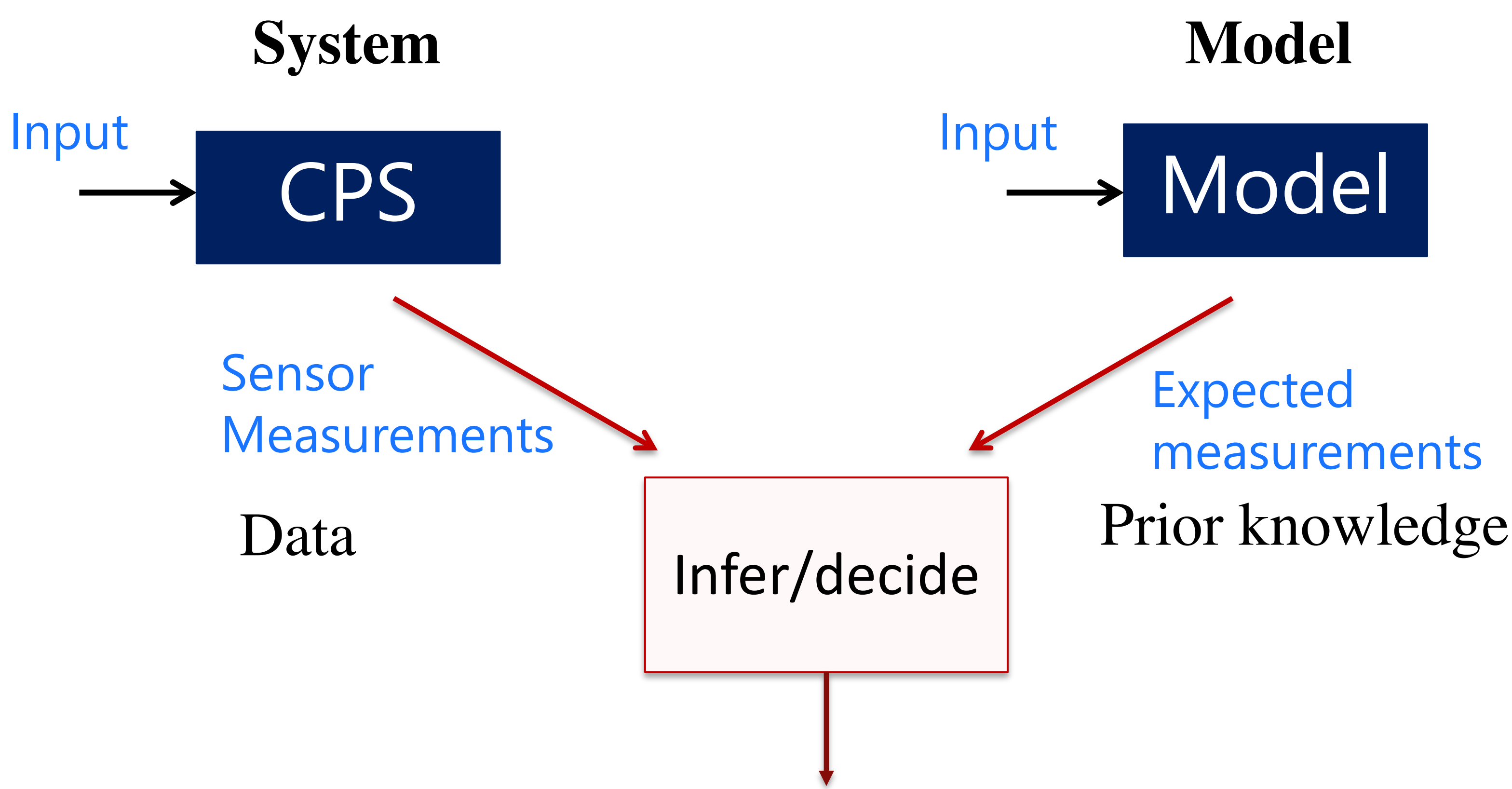




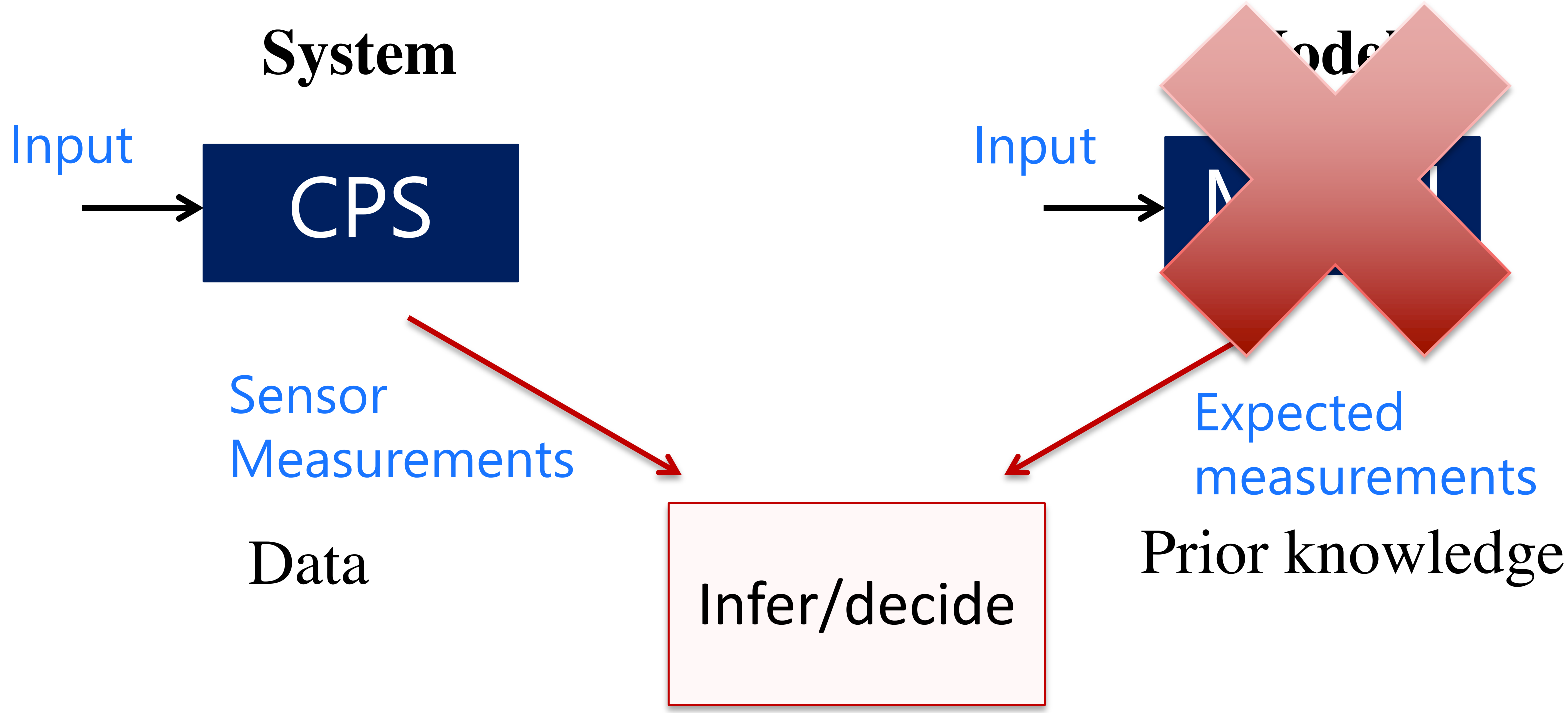




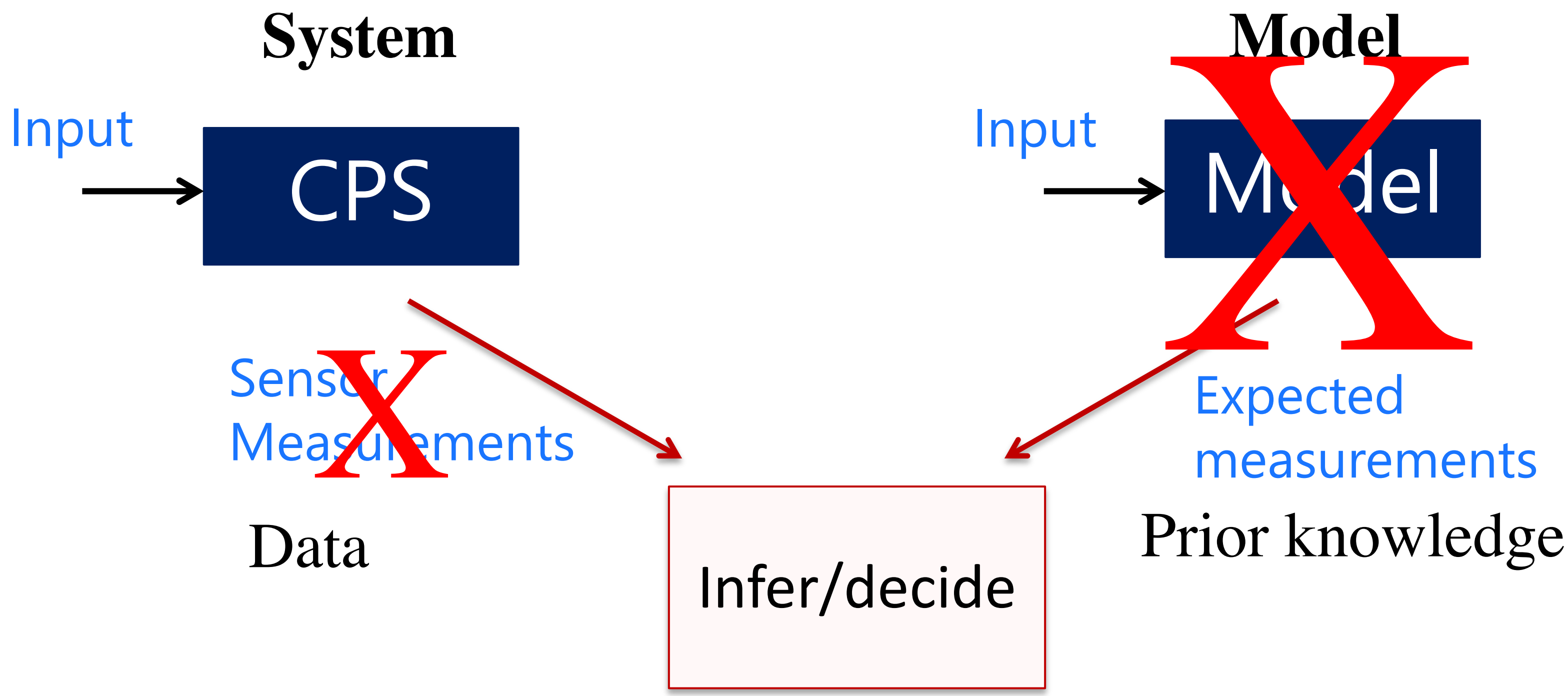
The issue with these sets of results



The issue with these sets of results

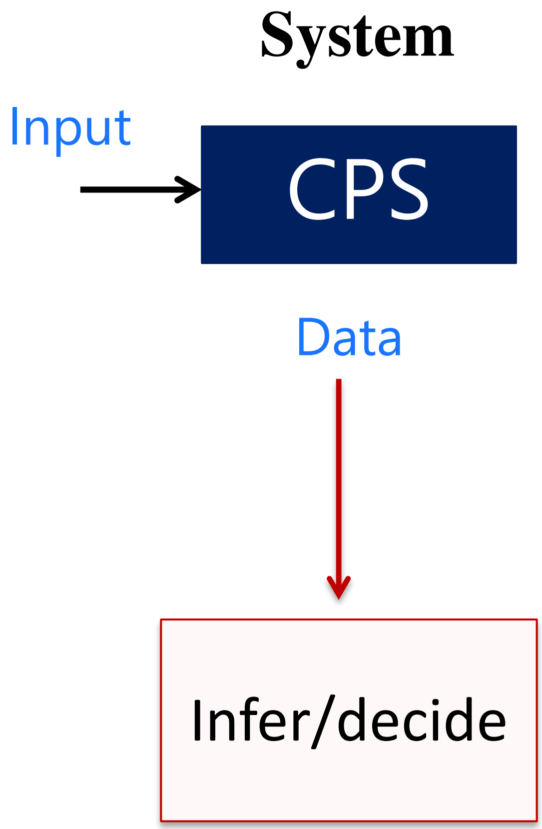


The issue with these sets of results



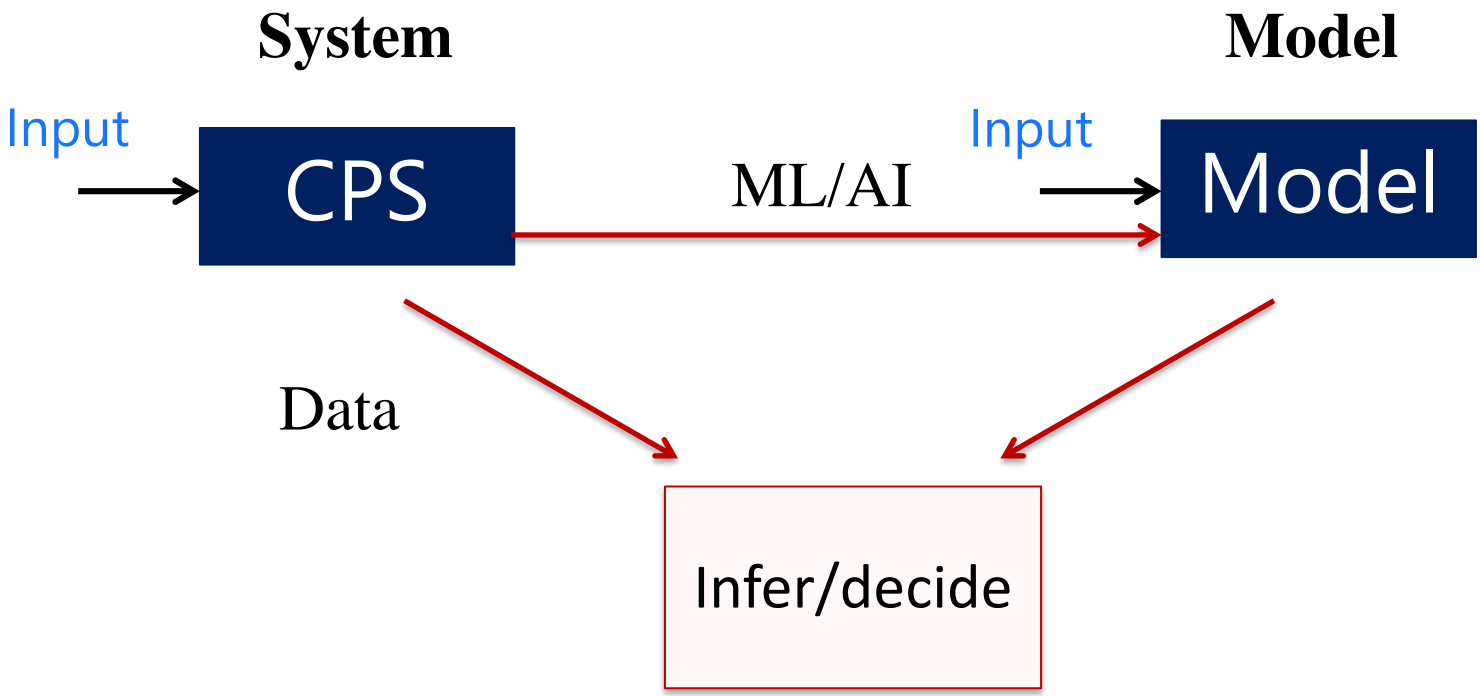
Complex perception problems
Lack of adequate first principle modeling

Black box paradigm (e.g. RL)



ML/AI-based perception/modeling

Grey Box (?): add understanding



The role of AI

- AI-ML is a tool and needs to be used as such
- Pros:
  - Modeling
  - Design
- Challenges
  - Analysis
  - Data need
  - Bias
  - Privacy
  - Security
- Interesting directions
  - Use data to further understanding of phenomena, modeling
  - Adaptivity
  - Analysis methods/certification
  - Accountability
  - Tradeoff between data complexity and performance
  - Human in the loop

Efforts at WashU

Washington University in St. Louis

Center for Trustworthy AI in CPS

FACULTY RESEARCH

Mission: The Center conducts research to advance trustworthy AI-driven CPS engineering. The Center will develop methods, tools, procedures, solutions, hardware, software, and integrated systems that result in AI-driven CPS that are secure, safe, reliable, and resilient.

Vision: The Center is known as a leading academic institution of global consequence in trustworthy AI-driven cyber-physical systems.

Impact: To achieve this vision, we will be at the vanguard of trustworthy AI in CPS research, generate innovations that can be leveraged by society, and engage in meaningful collaborations with industry, government, and academia on a regional, national, and global basis.

Multi university effort on Trustworthy AI in CPS

Washington University in St. Louis

University of Illinois Urbana-Champaign

University of Michigan

Berkeley University of California

Reflecting on 15 years of CPS

Thank you

References

Active Detection

- P. Griffioen, S. Weerakkody, B. Sinopoli, O. Ozel, and Y. Mo, “A tutorial on detecting security attacks on cyber-physical systems,” in *2019 18th European Control Conference (ECC)*. IEEE, 2019, pp. 979–984.
- S. Weerakkody, O. Ozel, P. Griffioen, and B. Sinopoli, “Active detection for exposing intelligent attacks in control systems,” in *2017 IEEE Conference on Control Technology and Applications (CCTA)*. IEEE, 2017, pp. 1306–1312.

Physical Watermarking

- S. Weerakkody, O. Ozel, and B. Sinopoli, “A bernoulli-gaussian physical watermark for detecting integrity attacks in control systems,” in *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2017, pp. 966–973.
- Y. Mo, S. Weerakkody, and B. Sinopoli, “Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs,” *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 93–109, 2015.
- S. Weerakkody, Y. Mo, and B. Sinopoli, “Detecting integrity attacks on control systems using robust physical watermarking,” in *53rd IEEE Conference on Decision and Control*. IEEE, 2014, pp. 3757–3764.
- Y. Mo, R. Chabukswar, and B. Sinopoli, “Detecting integrity attacks on scada systems,” *IEEE Transactions on Control Systems Technology*, vol. 22, no. 4, pp. 1396–1407, 2013.
- Y. Mo and B. Sinopoli, “Secure control against replay attacks,” in *2009 47th annual Allerton conference on communication, control, and computing (Allerton)*. IEEE, 2009, pp. 911–918.
- M. Hosseinzadeh, B. Sinopoli, and E. Garone, “Feasibility and Detection of Replay Attack in Networked Constrained Cyber-Physical Systems,” in *Proceedings of the 57th Annual Allerton Conference on Communication, Control, and Computing*, Allerton Park & Retreat Center, Monticello, IL, USA, Sep.24-27, 2019, pp. 712-717.
- R. Romagnoli, S. Weerakkody and B. Sinopoli, "A Model Inversion Based Watermark for Replay Attack Detection with Output Tracking," 2019 American Control Conference (ACC), Philadelphia, PA, USA, 2019, pp. 384-390, doi: 10.23919/ACC.2019.8814483.

References

Moving Target Defense

- P. Griffioen, S. Weerakkody, and B. Sinopoli, “A moving target defense for securing cyber-physical systems,” *IEEE Transactions on Automatic Control*, vol. 66, no. 5, pp. 2016–2031, 2021.
- P. Griffioen, S. Weerakkody, and B. Sinopoli, “An optimal design of a moving target defense for attack detection in control systems,” in *2019 American Control Conference (ACC)*. IEEE, 2019, pp. 4527–4534.
- S. Weerakkody and B. Sinopoli, “A moving target approach for identifying malicious sensors in control systems,” in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2016, pp. 1149–1156. [Online]. Available: http://arxiv.org/abs/1609.09043
- S. Weerakkody and B. Sinopoli, “Detecting integrity attacks on control systems using a moving target approach,” in *2015 54th IEEE Conference on Decision and Control (CDC)*. IEEE, 2015, pp. 5820–5826.

References

Software Rejuvenation

- R. Romagnoli, P. Griffioen, B. H. Krogh, and B. Sinopoli, "Software rejuvenation under persistent attacks in constrained environments," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 4088–4094, 21st IFAC World Congress. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405896320331190
- P. Griffioen, R. Romagnoli, B. H. Krogh, and B. Sinopoli, "Secure networked control for decentralized systems via software rejuvenation," in *2020 American Control Conference (ACC)*. IEEE, 2020, pp. 1266–1273.
- R. Romagnoli, B. H. Krogh, and B. Sinopoli, "Robust software rejuvenation for cps with state estimation and disturbances," in *2020 American Control Conference (ACC)*. IEEE, 2020, pp. 1241–1246.
- P. Griffioen, R. Romagnoli, B. H. Krogh, and B. Sinopoli, "Secure networked control via software rejuvenation," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 3878–3884.
- R. Romagnoli, B. H. Krogh, and B. Sinopoli, "Design of software rejuvenation for cps security using invariant sets," in *2019 American Control Conference (ACC)*. IEEE, 2019, pp. 3740–3745.
- R. Romagnoli, B. H. Krogh, and B. Sinopoli, "Safety and liveness of software rejuvenation for secure tracking control," in *2019 18th European Control Conference (ECC)*. IEEE, 2019, pp. 2215–2220.
- R. Romagnoli, B. H. Krogh, D. de Niz, and B. Sinopoli, "Software rejuvenation for secure tracking control," *arXivpreprint arXiv:1810.10468*, 2018.

References

Decentralized Event-Triggered Control

- P. Griffioen, R. Romagnoli, B. H. Krogh, and B. Sinopoli, "Decentralized event-triggered control in the presence of adversaries," in *2020 59th IEEE Conference on Decision and Control (CDC)*, 2020, pp. 3236–3242.

Resilient Overlay Networks

- P. Griffioen, R. Romagnoli, B. H. Krogh, and B. Sinopoli, "Resilient control in the presence of man-in-the-middle attacks," in *2021 American Control Conference (ACC)*. IEEE, 2021.

Secure Estimation/detection

- **N. Forti, G. Battistelli, L. Chisci, B. Sinopoli, "Joint Attack Detection and Secure State Estimation of Cyber-Physical Systems", *International Journal of Robust and Nonlinear Control, Special Issue: Privacy and Security of Cyber-Physical Systems,* Vol 30 (11), pp. 4303-4330, July 2020.**
- **N. Forti, G. Battistelli, L. Chisci, S. Li, B. Wang, and B. Sinopoli, "Distributed joint attack detection and secure state estimation, " *IEEE Transactions on Signal and Information Processing over Networks, Special Issue on Distributed Signal Processing for Security and Privacy in Networked Cyber-Physical Systems*, 2017.**
- **Y. Mo and B. Sinopoli, "Secure Estimation in the Presence of Integrity Attacks," *IEEE Transactions on Automatic Control*, vol. 60, no. 4, pp. 1145- 1151, 2015.**
- **Y. Mo, J.P. Hespanha, and B. Sinopoli, "Resilient Detection in the Presence of Integrity Attacks," *IEEE Transactions on Signal Processing*, vol. 62, no. 1, pp. 31-43, 2014.**
- **K.G. Vamvoudakis, J.P. Hespanha, B. Sinopoli, and Y. Mo, "Detection in Adversarial Environments," *IEEE Transactions on Automatic Control*, vol. 59, no. 12, pp. 3209-3223, 2014.**

Xie, Le, Yilin Mo, and Bruno Sinopoli. 2011. “*Integrity Data Attacks in Power Market Operations*.” IEEE Transactions on Smart Grid 2 (4): 659–66.

Liu, Yao, Peng Ning, and Michael K. Reiter. 2011. “*False Data Injection Attacks against State Estimation in Electric Power Grids*.” ACM Transactions on Information and System Security 14 (1): 1–33.

Pasqualetti, Fabio, Antonio Bicchi, and Francesco Bullo. 2012. “*Consensus Computation in Unreliable Networks: A System Theoretic Approach*.” IEEE Transactions on Automatic Control 57 (1): 90–104.

Pasqualetti, Fabio, Florian Dorfler, and Francesco Bullo. 2013. “*Attack Detection and Identification in Cyber-Physical Systems*.” IEEE Transactions on Automatic Control 58 (11): 2715–29.

Hendrickx, Julien M., Karl Henrik Johansson, Raphael M. Jungers, Henrik Sandberg, and Kin Cheong Sou. 2014. “*Efficient Computations of a Security Index for False Data Attacks in Power Networks*.” IEEE Transactions on Automatic Control 59 (12): 3194–3208.

Fawzi, Hamza, Paulo Tabuada, and Suhas Diggavi. 2014. “*Secure Estimation and Control for Cyber-Physical Systems Under Adversarial Attacks*.” IEEE Transactions on Automatic Control 59 (6): 1454–67.

Y. Mo, S. Weerakkody, and B. Sinopoli. 2015. “*Physical Authentication of Control Systems: Designing Watermarked Control Inputs to Detect Counterfeit Sensor Outputs*.” IEEE Control Systems 35 (1).

Teixeira, André, Iman Shames, Henrik Sandberg, and Karl Henrik Johansson. 2015. “*A Secure Control Framework for Resource-Limited Adversaries*.” Automatica 51 (C): 135–48.

Mo, Yilin, and Bruno Sinopoli. 2016. “*On the Performance Degradation of Cyber-Physical Systems Under Stealthy Integrity Attacks*.” IEEE Transactions on Automatic Control 61 (9): 2618–24.

Shoukry, Yasser, Pierluigi Nuzzo, Alberto Puggelli, Alberto L. Sangiovanni-Vincentelli, Sanjit A. Seshia, and Paulo Tabuada. 2017. “*Secure State Estimation for Cyber-Physical Systems Under Sensor Attacks: A Satisfiability Modulo Theory Approach*.” IEEE Transactions on Automatic Control 62 (10): 4917–32.

Satchidanandan, Bharadwaj, and P. R. Kumar. 2017. “*Dynamic Watermarking: Active Defense of Networked Cyber–Physical Systems*.” Proceedings of the IEEE 105 (2): 219–40.

Sundaram, Shreyas, and Bahman Gharesifard. 2019. “*Distributed Optimization Under Adversarial Nodes*.” IEEE Transactions on Automatic Control 64 (3): 1063–76.

S. Weerakkody, X. Liu, S. H. Son, and B. Sinopoli, "A Graph Theoretic Characterization of Perfect Attackability for the Secure Design of Distributed Control Systems," *IEEE Transactions on Control of Network Systems,* Vol 4, no. 1, pp. 1060-1070, 2017.

Extra Slides

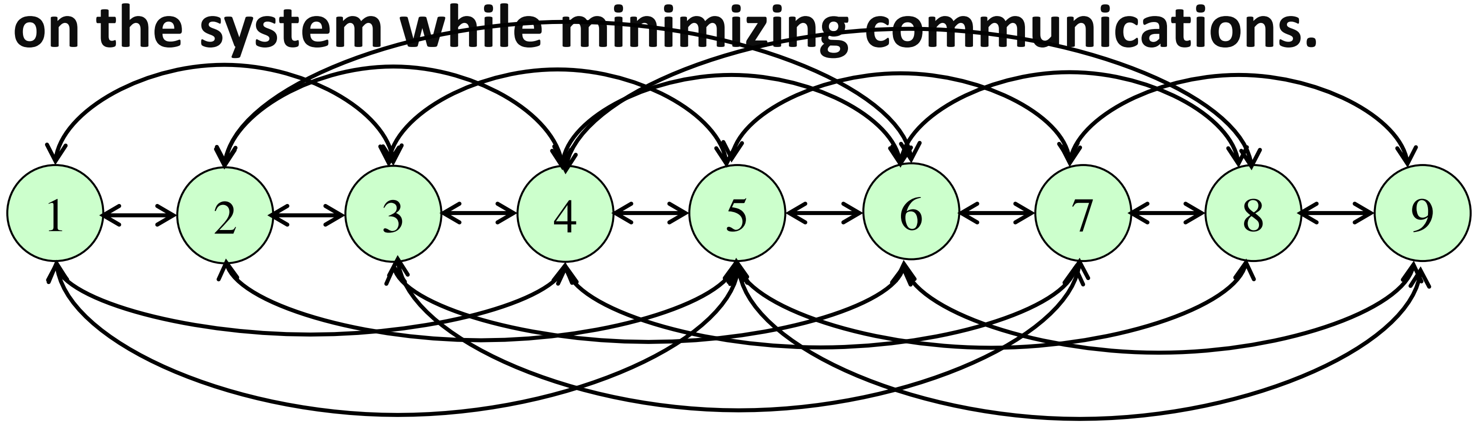
Secure Design of Distributed Control Systems

Design a sensing/communication topology to guarantee detection of misbehaving agents

S. Weerakkody, X. Liu, S. H. Son, and B. Sinopoli, "A Graph Theoretic Characterization of Perfect Attackability for the Secure Design of Distributed Control Systems," *IEEE Transactions on Control of Network Systems*, Vol 4, no. 1, pp. 1060-1070, 2017.

Example: Formation Control

- **9 vehicles want to keep the same speed and can only communicate with up to 4 vehicles ahead or behind them.**
- **An adversary attacks may up to 3 unknown vehicles or sensors .**
- **Design Problem 1: Which nodes should be observed by centralized detector?**
- **Design Problem 2: How can we remain robust to attacks on the system while minimizing communications.**



Attack characterization (Mo et al.)

- **Perfect Attack: The attacker could destabilize the system, without changing the residue. A system is perfectly attackable if there exists a feasible perfect attack.**

- **Nearly Perfect Attack: The attack could destabilize the system, with bounded change of the residue.**

Perfect Attack: Topological Characterization

- ***Definition:*** **A vertex separator between non-adjacent nodes *a* and *b* is a set of vertices whose removal, deletes all paths from *a* to *b***

- ***Theorem 1*****: Consider a graph *G* generated from agent *X,* sensor *Y,* and detector *d* interactions. Given *p* compromised agents, the system is generically perfectly attackable for some feasible attack configuration if and only if for some agent node *x*, the size of the minimum vertex separator from *x* to *d* is less than *p.***

Perfect Attack: Network Optimization

- *Theorem* 2: Given p compromised nodes, m observed nodes, and n agents, the minimum number of communications needed for a system not to be perfectly attackable is np-m.

- *Remark:* A feasible configuration for an unconstrained system exists if and only if $m \geq p$. The above theorem assumes there are no constraints on communication.

Perfect Attack: Graphical Realization

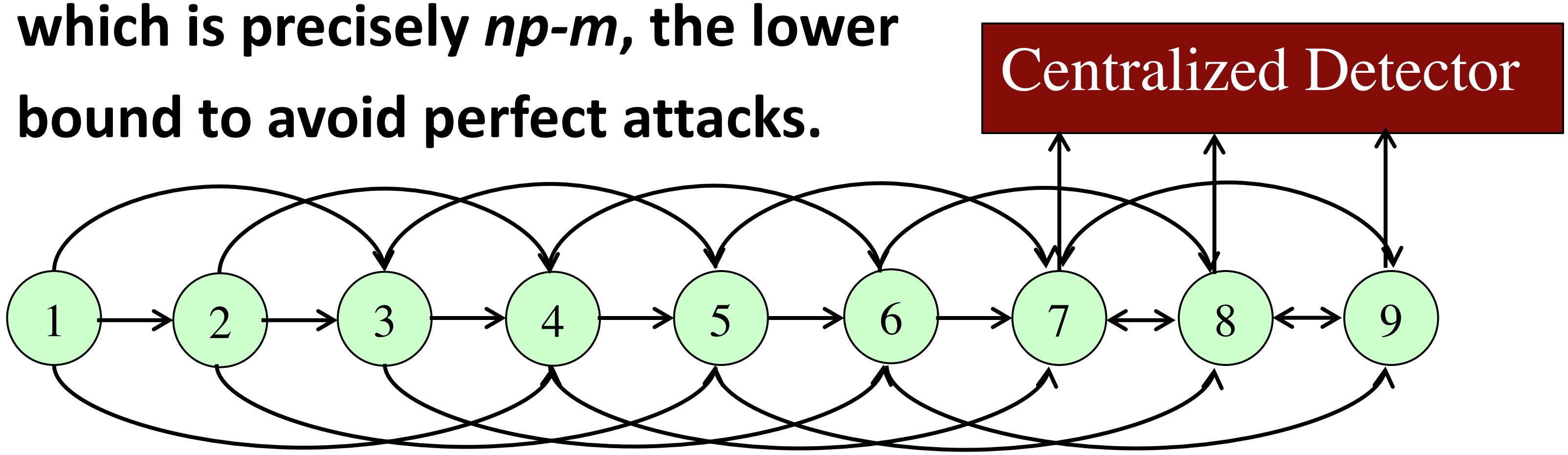
- ***Corollary 3*: Suppose there exist no cycles in graph *G* among unobserved nodes. Then the following conditions are necessary and sufficient for optimality.**

The out-degree (ignoring self loops) of each node is p.

Feasible Configuration

- An adversary may attack up to **3 unknown vehicles or sensors, $p = 3$.**
- Suppose the centralized detector **observes 3 vehicles** as shown, $m = 3$. The total number of vehicles $n = 9$.

- Each of the first 6 vehicles communicates with the 3 vehicles ahead of it. The last 3 vehicles are observed and communicate with 2 other vehicles. There are 24 edges which is precisely $np-m$, the lower bound to avoid perfect attacks.



Perfect Attack: Joint Sensor and Network Optimization

***Theorem 4*: Suppose in an unconstrained network we wish to minimize the number of sensors and communication**

$$\min_{G} C_1(\text{number of links}) + C_2 m$$

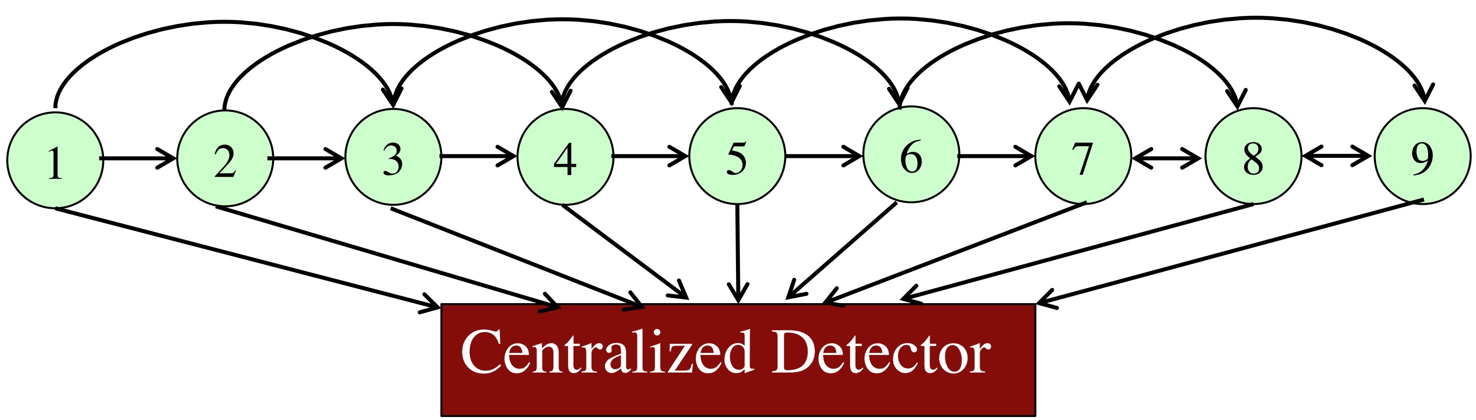
- **If sensing is more expensive than communicating, take *m=p.* (This is what we did before.)**
- **If communicating is more expensive, observe all nodes.**

Case: Communicating more Costly

- **An adversary may attack up to 3 unknown vehicles, $p = 3$.**

- **Suppose the centralized detector observes all the vehicles as shown, $m = 9$.**

- **Each of the 9 vehicles communicates with 2 other vehicles, thus we have 6 less communication links than before.**



Perfect Attack: Network Optimization with Constraints

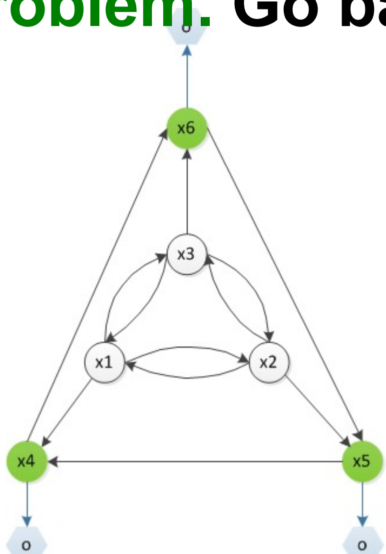
- ***Theorem 5*: Given *p* compromised nodes, *m* fixed observed nodes, and *n* agents, and a set of agents which are allowed to communicate, the minimum number of communications is *np-m.***

- ***Remark:* Even with constraints on the system we can obtain a minimal network as long as ensuring the system is not perfectly attackable is feasible**

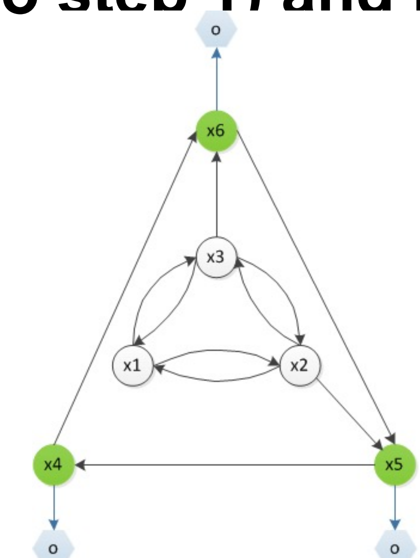
Obtaining a minimal network

1) **Consider node *x* with out-degree *p'* greater than *p*.**

2) **Remove edges to *p'-p* neighbors which are not necessary to ensure system is not perfectly attackable. Equivalent to solving a maximum flow problem. Go back to step 1) and repeat**

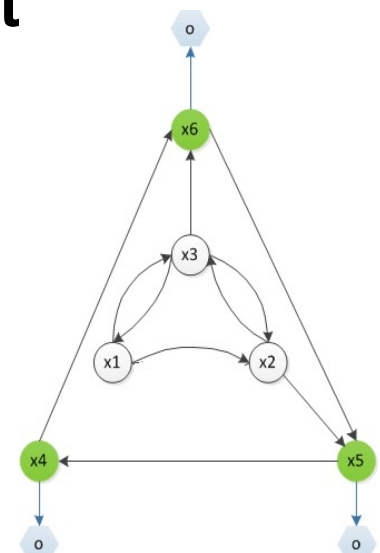


Original Network, $p = 2$



Node x1 has out-degree > 2

Remove (x1,x4)



Node x2 has out-degree > 2

Remove (x2,x1)

Perfect Attack: Joint Sensor and Network Optimization

***Theorem 6*: Suppose in a constrained network we wish to minimize the number of sensors and communication**

$$\min_{G \subseteq G^*} \; C_1(\text{number of links}) + C_2 m$$

- **If sensing is more expensive than communicating, take *m=p*,* the minimum number of observers needed to ensure system is not perfectly attackable**

- **If communicating is more expensive, observe all nodes.**