1

Acknowledgments


Henrik Sandberg, KTH


André Teixeira, UU


Ehsan Nekouei, City U HK


Song Fang, NYU

- H. Sandberg, A. Teixeira, K. H. Johansson, "On security indices for state estimators in power networks," Workshop on Secure Control Systems, 2010.
- A. Teixeira, I. Shames, H. Sandberg, K. H. Johansson, "A secure control framework for resource-limited adversaries," Automatica, 2015.
- A. M. H. Teixeira, K. C. Sou, H. Sandberg, K. H. Johansson, "Secure control systems: A quantitative risk management approach," IEEE Control Systems Mag, 2015.
- E. Nekouei, H. Sandberg, M. Skoglund, K. H. Johansson, "Privacy-aware minimum error probability estimation: an entropy constrained approach," ArXiv, 1808.02271 and MTNS, 2018
- E. Nekouei, T. Tanaka, M. Skoglund, K.H. Johansson, "Information-theoretic approaches to privacy in estimation and control," Annual Reviews in Control, 2019
- S. Fang, K.H. Johansson, M. Skoglund, H. Sandberg, H. Ishii, "Two-way coding in control systems under injection attacks: from attack detection to attack correction," ACM/IEEE International Conference on Cyber-Physical Systems, 2019

Knut och Alice Wallenbergs Stiftelse · Vetenskapsrådet · VINNOVA · SEVENTH FRAMEWORK PROGRAMME · HORIZON 2020

2

2

Outline

- What is control systems security?
- Motivation: Why should we care?
- Q1: How to conduct a false-data injection attack?
- Q2: How to defend against such attacks?
- Q3: How to do remote estimation under privacy constraints?
- Outlook and conclusions

3

3

Information Security

Confidentiality: information is not disclosed to unauthorized individuals

Integrity: information cannot be modified in an unauthorized manner

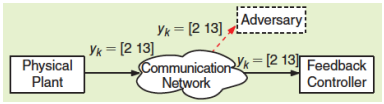
Availability: information must be available when it is needed

[Bishop, 2002]

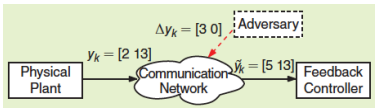
4

4

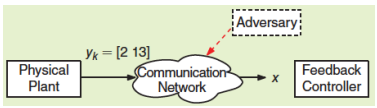
Control Systems Security



Confidentiality: information is not disclosed to unauthorized individuals



Integrity: information cannot be modified in an unauthorized manner

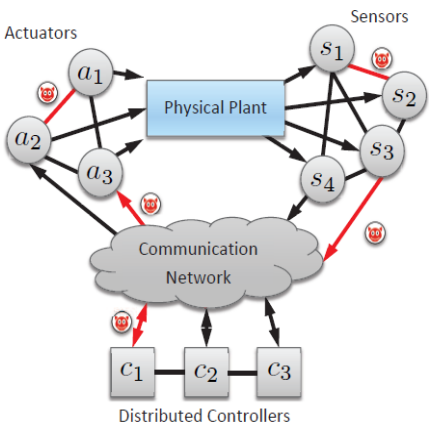


Availability: information must be available when it is needed

5

5

Vulnerabilities in Networked Control Systems

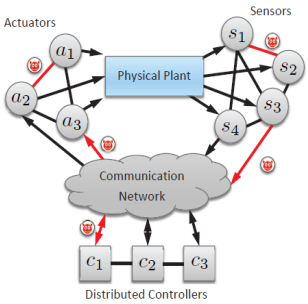


6

6

Why is not IT Security Enough?

- **Interfaces between cyber** and **physical systems** cannot be protected by conventional IT security
  - Cf. a physical variable vs its sampled sensor measurement
- **Physics can help** to detect if data have been manipulated
- **Malicious attackers** have an **intent** and can act strategically, as opposed to system faults or disturbances
- **Risk analysis** for safety-critical control systems needs to include the interactions between **cyber, physical,** and **human systems**
- Modern security of infrastructure systems requires **defense in depth**: multiple layers of security are needed for protection

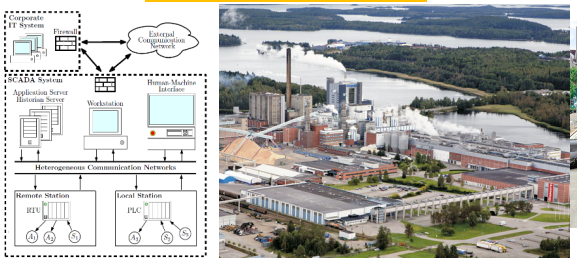


7

7

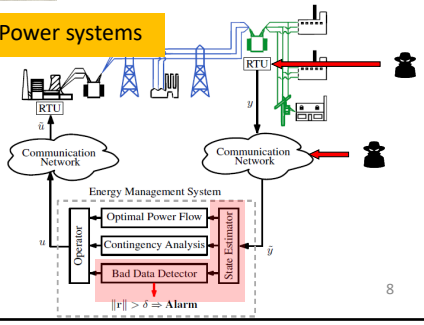
Motivating Applications

Industrial control systems

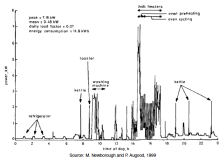


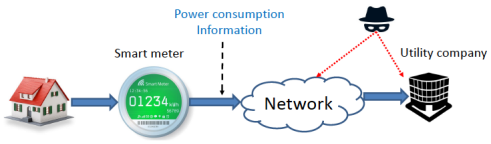
Intelligent transport

Power systems



Building automation

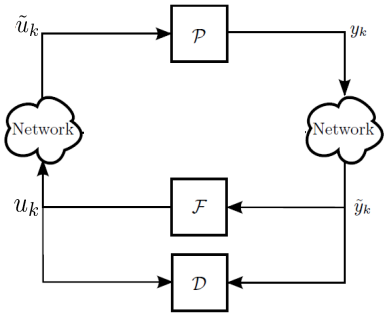




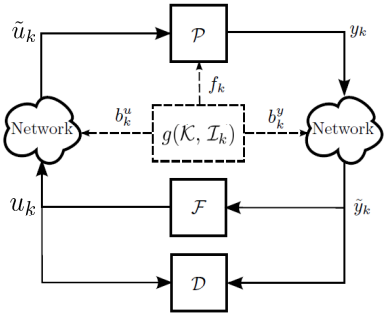
8

8

Networked Control System under Attack



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



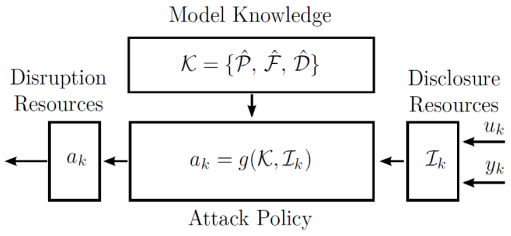
- Physical attacks f_k
- Data injection attacks

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

[Teixeira *et al.*, Automatica, 2015]

9

Adversary Model

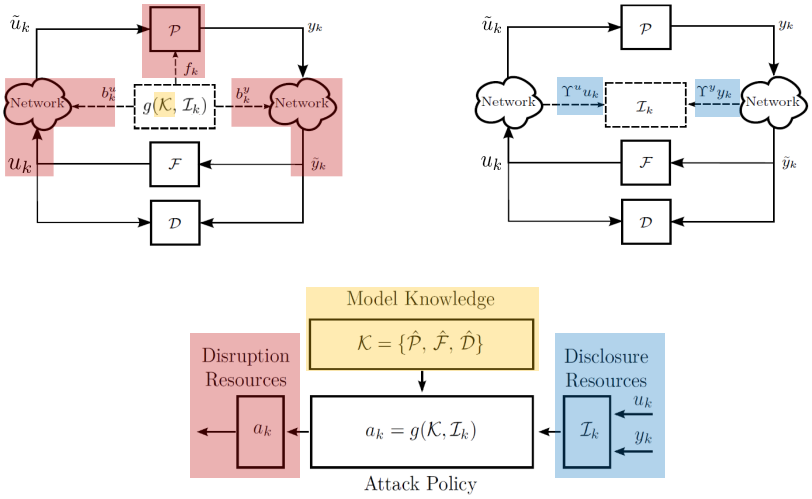


- **Attack policy:** Goal of the attack?
- **Model knowledge:** Adversary has some info about plant and controller?
- **Disclosure resources:** Which channels can the adversary listen to?
- **Disruption resources:** Which channels can the adversary manipulate?

[Teixeira *et al.*, Automatica, 2015]

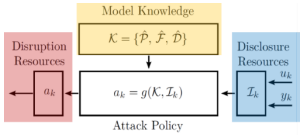
10

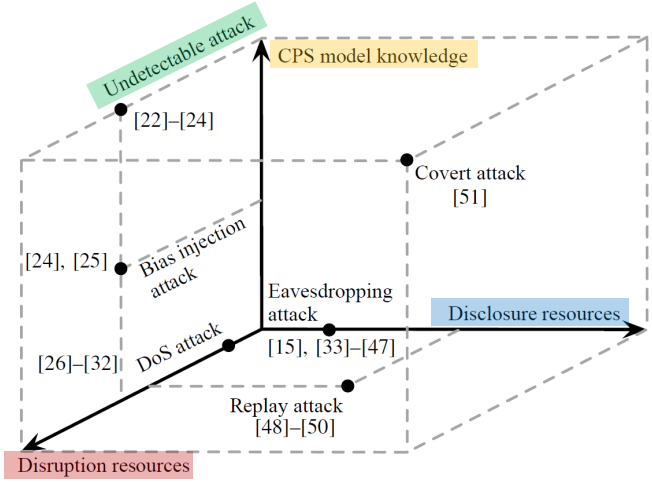
Networked Control System with Adversary Model



11

Attack Space





[Teixeira *et al.*, Automatica, 2015; Chong *et al.*, ECC, 2019]

12

Outline

- What is control systems security?
- Motivation: Why should we care?
- Q1: How to conduct a false-data injection attack?
- Q2: How to defend against such attacks?
- Q3: How to do remote estimation under privacy constraints?
- Outlook and conclusions

13

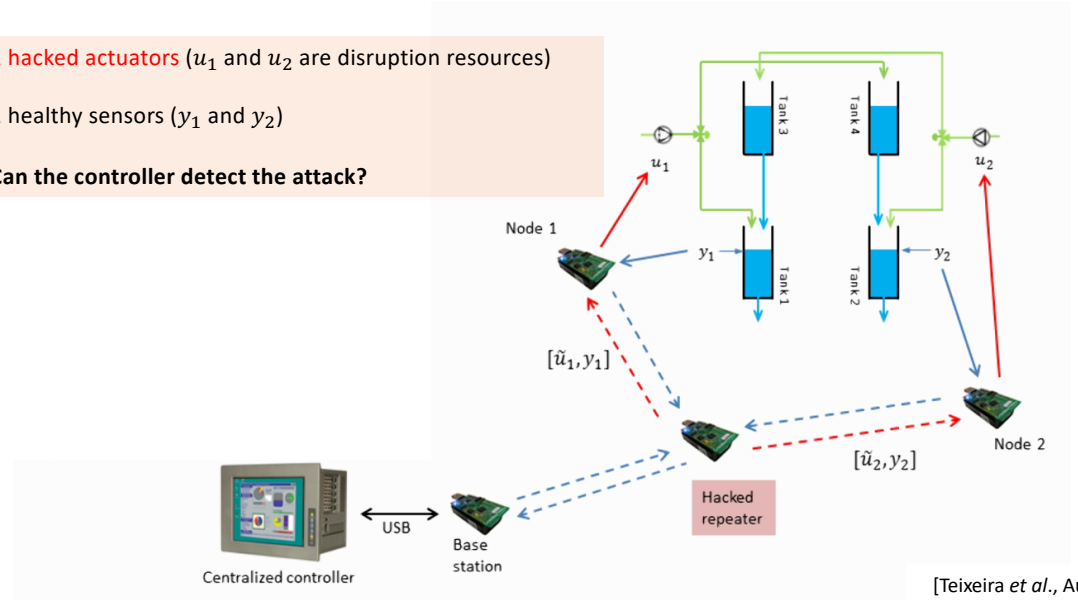
13

False-data Injection Attack on Quadruple Tank System

2 hacked actuators (u_1 and u_2 are disruption resources)

2 healthy sensors (y_1 and y_2)

Can the controller detect the attack?

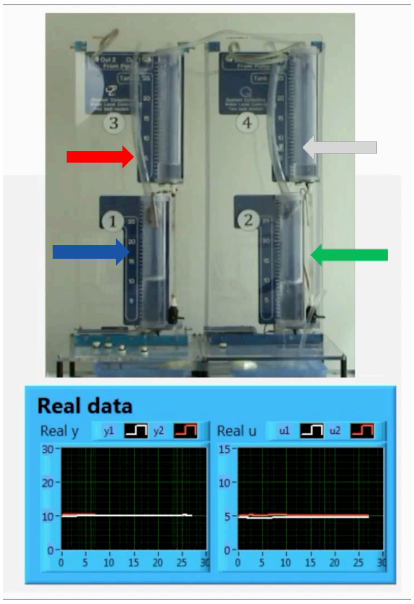


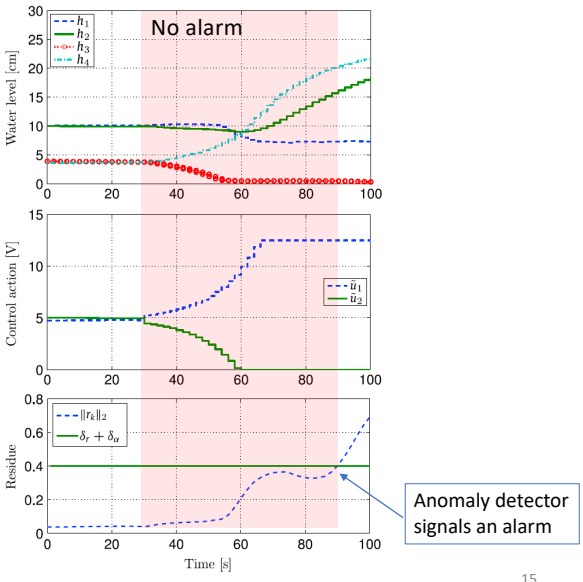
[Teixeira *et al.*, Automatica, 2015]

14

14

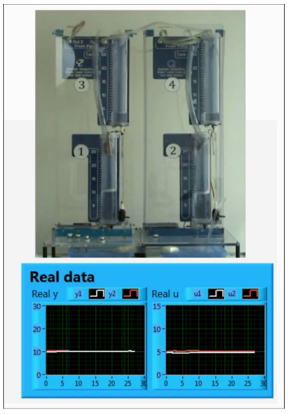
False-data Injection Attack on Quadruple Tank System





[Teixeira *et al.*, 2015]

Attack Analysis



Transfer function matrix from attack to sensor signals

$$G_a(z) = C(zI - A)^{-1}B = \begin{pmatrix} \frac{0.0289}{z-0.8076} & \frac{(1.277z+1.182)\cdot 10^{-3}}{z^2-1.784z+0.7928} \\ \frac{(1.356z+1.24)\cdot 10^{-3}}{z^2-1.754z+0.7643} & \frac{0.02954}{z-0.8347} \end{pmatrix}$$

Poles = $\{0.8076, 0.8347, 0.9464, 0.9498\}$

Zeros = $\{0.8675, 1.0362\}$ ⇒ Non-minimum phase system

Applied attack signal (for small ϵ)

$$a(k) = 1.0362^k \begin{pmatrix} 0.2281\epsilon \\ -0.2281\epsilon \end{pmatrix}, \quad x_0 = \begin{pmatrix} 0 & 0 & -0.6521\epsilon & 0.6876\epsilon \end{pmatrix}^T$$

satisfies **zero dynamics** and is thus **masked by** system transient:

$$0 = y(k) = CA^k x_0 + (g_a * a)(k), \quad k \geq 0$$

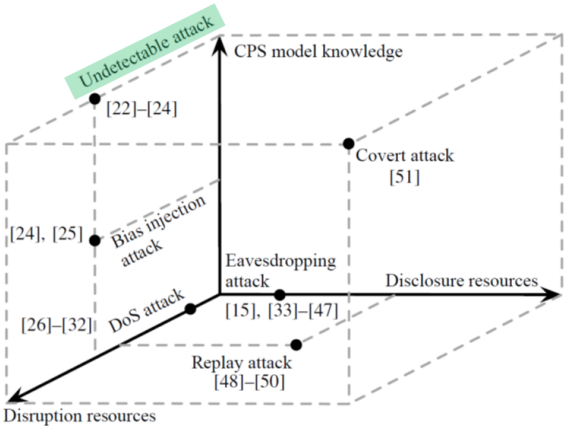
Undetectable Attack on Quadruple Tank System

2 hacked actuators (u_1 and u_2 are disruption resources)

2 healthy sensors (y_1 and y_2)

Can the controller detect the attack?

Not against an adversary with plant model knowledge



[Teixeira *et al.*, Automatica, 2015]

17

17

Undetectable Attacks: General Linear Systems

Consider the linear system $y = G_a a$ (the closed-loop control system):

$$x(k+1) = Ax(x) + Ba(k)$$
$$y(k) = Cx(k) + Da(k)$$

Operator: State $x(k) \in \mathbb{R}^n$ and malicious attack $a(k) \in \mathbb{R}^m$ *unknown*.
Measurement $y(k) \in \mathbb{R}^p$ and model A, B, C, D *known*

Attacker: Model A, B, C, D *known* and can manipulate attack signal $a(k)$ arbitrarily

Definition: Attack signal a is *undetectable* if there exists an initial state $x(0)$ such that $y(k) = 0, k \geq 0$

[Pasqualetti et al., 2013; Sandberg and Teixeira, 2016]

Remark: Less strict undetectable attacks ($y(k) \approx 0$) have been proposed in both deterministic [Teixeira et al., 2015] and stochastic [Bai et al., 2017] settings

18

18

Undetectable Zero-Dynamics Attack

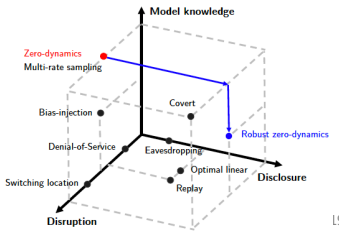
Theorem: Attack signal $a(k) = z_0^k a_0$, $0 \neq a_0 \in \mathbb{C}^m$, $z_0 \in \mathbb{C}$, is *undetectable* if and only if there exists $x_0 \in \mathbb{C}^n$ such that

$$P(z_0)\begin{bmatrix} x_0 \\ a_0 \end{bmatrix} = 0$$

with the Rosenbrock system matrix $P(z) = \begin{bmatrix} A - zI & B \\ C & D \end{bmatrix}$

[Pasqualetti et al., 2013; Sandberg and Teixeira, 2016]

Perfect model knowledge is not needed for attacker:
Extensions to robust zero-dynamics attack, where model knowledge is replaced by observing plant inputs and outputs [Park et al., 2019]



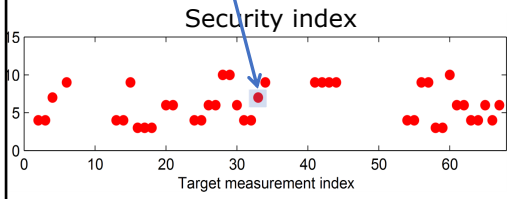
19

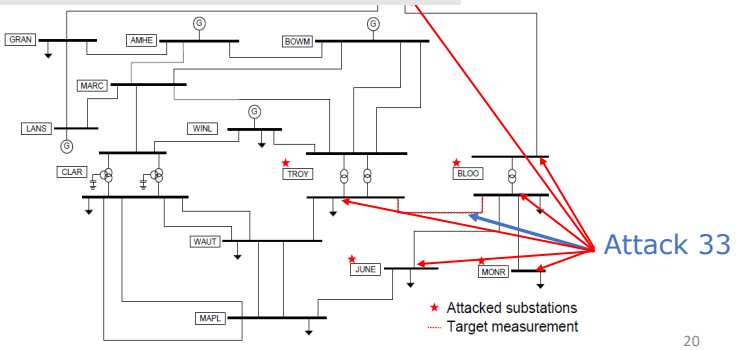
19

False-data Injection Attack on a Power Network

- Attack applied to power system state estimator by manipulating sensor measurements
- The number of measurements that needs to be manipulated for an undetectable attack on one targeted state measurement is denoted its **security index**
  - Validated on industrial SCADA testbed
  - Cyber attack based on physical information

At least 7 measurements have to be involved in an undetectable attack against measurement 33





[Teixeira et al., 2011; Hendrickx et al., 2014]

20

20

Outline

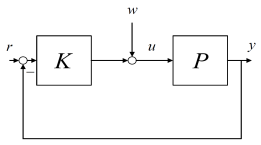
- What is control systems security?
- Motivation: Why should we care?
- Q1: How to conduct a false-data injection attack?
- Q2: How to defend against such attacks?
- Q3: How to do remote estimation under privacy constraints?
- Outlook and conclusions

21

21

Coding to Counteract False-data Injection Attacks

Data injection attack $w = w_0 e^{zt}$ is undetectable if z is a plant zero: $P(z) = 0$



Can we introduce coding to render the attack harmless?

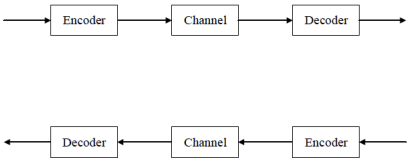
[Fang *et al.*, ICCPS, 2019]

22

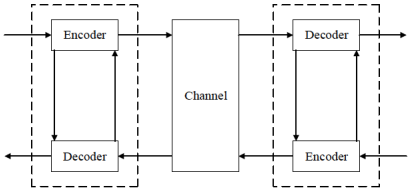
22

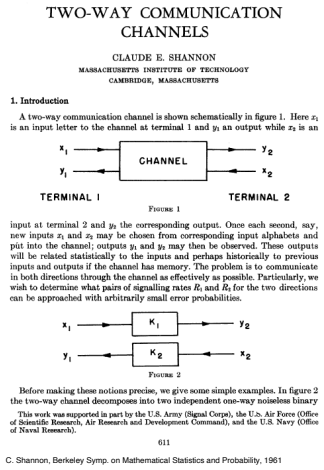
Shannon Two-Way Communication Channel

One-way coding for one-way channels



Two-way coding for two-way channels





TWO-WAY COMMUNICATION CHANNELS

CLAUDE E. SHANNON
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS

1. Introduction

A two-way communication channel is shown schematically in figure 1. Here x_1 is an input letter to the channel at terminal 1 and y_1 an output while x_2 is an

FIGURE 1

input at terminal 2 and y_2 the corresponding output. Once each second, say, new inputs x_1 and x_2 may be chosen from corresponding input alphabets and put into the channel; outputs y_1 and y_2 may then be observed. These outputs will be related statistically to the inputs and perhaps historically to previous inputs and outputs if the channel has memory. The problem is to communicate in both directions through the channel as effectively as possible. Particularly, we wish to determine what pairs of signalling rates R_1 and R_2 for the two directions can be approached with arbitrarily small error probabilities.

FIGURE 2

Before making these notions precise, we give some simple examples. In figure 2 the two-way channel decomposes into two independent one-way noiseless binary

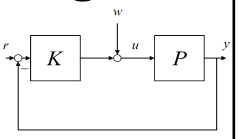
This work was supported in part by the U.S. Army (Signal Corps), the U.S. Air Force (Office of Scientific Research, Air Research and Development Command), and the U.S. Navy (Office of Naval Research).

C. Shannon, Berkeley Symp. on Mathematical Statistics and Probability, 1961

[Fang *et al.*, ICCPS, 2019]

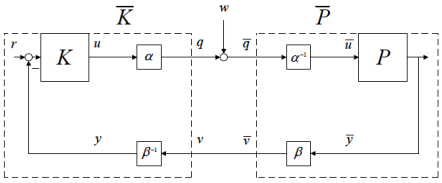
Data injection attack under **one-way coding**

Data injection attack $w = w_0 e^{zt}$, $\mathrm{Re}(z) > 0$, is undetectable if z is a plant zero: $P(z) = 0$



Data injection attack is undetectable also under **one-way coding**, because unstable zeros cannot be cancelled:

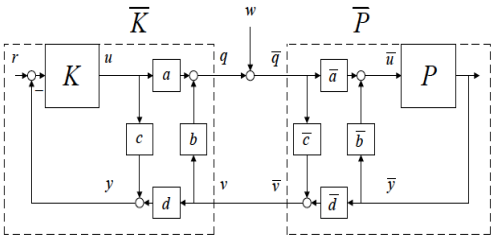
$$\overline{P}(s) = \beta P(s) \alpha^{-1}$$



[Fang *et al.*, ICCPS, 2019]

Data injection attack under **two-way coding**

Suppose the attacker knows the plant model $P(s)$ and performs a data injection attack $w = w_0 e^{zt}, \mathrm{Re}(z) > 0$



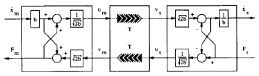
The data injection attack w is detectable under **two-way coding** because plant zeros can be moved:

$$\overline{P}(s) = \frac{P(s) - c}{ad - bc + bP(s)}$$

$$\begin{bmatrix} \bar{a} & \bar{d} \\ \bar{c} & \bar{d} \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1}$$

Remark

- If the attacker knows or learns $\bar{P}(s)$, a new undetectable data injection attack w can be performed
- The modified plant $\bar{P}(s)$ can be designed to have only stable zeros, so the attack $w = w_0 e^{zt} \to 0$ as $t \to \infty$
- Two-way coding resembles the scattering transformation used in teleoperated robots [Hokayem & Spong, 2006]



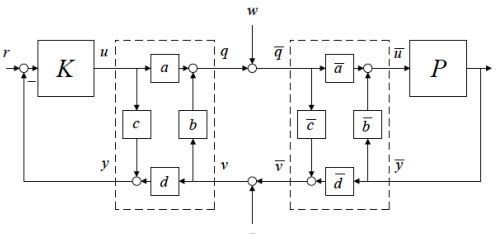
[Fang *et al.*, 2019]

25

25

Attack decoupling with **two-way coding**

Definition An attack is **decoupled** if the transfer function from attack to plant input/output can be made zero, without nullifying the transfer function from reference to plant input/output



Theorem Suppose $P(s)$ is stabilized by static output feedback controller $K(s) = K$.

- If $c = -1/K$, then any attack $w = w(t)$ is decoupled.
- If $b = (ad - bc)K$, then any attack $z = z(t)$ is decoupled.

Remark

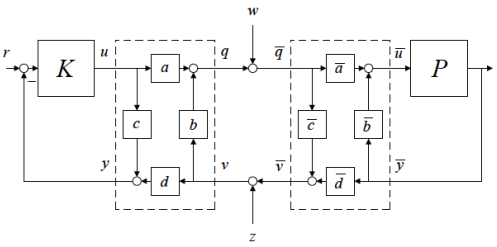
- Decoupling for w should be interpreted in the sense of "almost decoupling," cf., [Weiland and Willems, 1989] notion of "almost disturbance decoupling"

[Fang *et al.*, ICCPS, 2019]

26

26

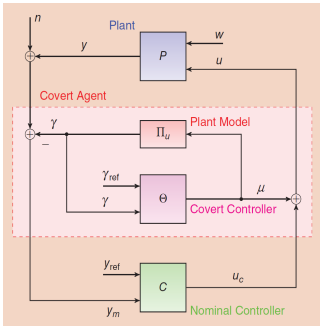
Impossibility Theorem

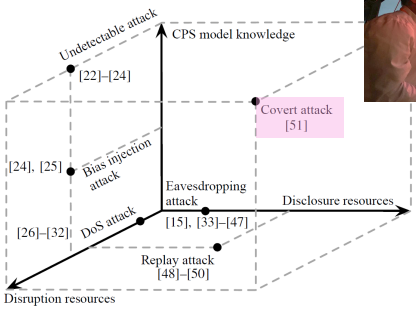


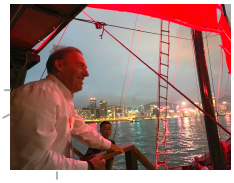
Theorem *w* and *z* cannot be decoupled simultaneously.

Remark

- Corresponds to the covert attack proposed in [Smith, IEEE CSM, 2015]:







[Fang *et al.*, ICCPS, 2019]

27

27

Outline

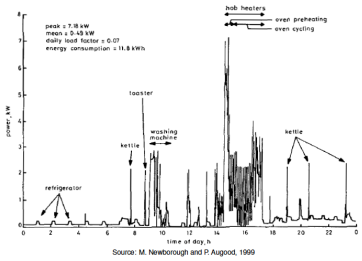
- What is control systems security?
- Motivation: Why should we care?
- Q1: How to conduct a false-data injection attack?
- Q2: How to defend against such attacks?
- Q3: How to do remote estimation under privacy constraints?
- Outlook and conclusions

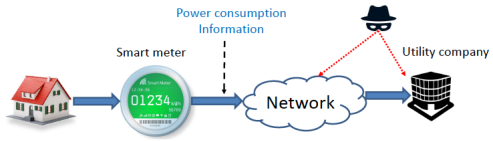
28

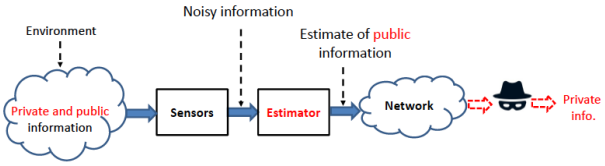
28

Remote Estimation under Privacy Constraints

Can we preserve consumers privacy despite cloud-based online temparature monitoring and other apps?



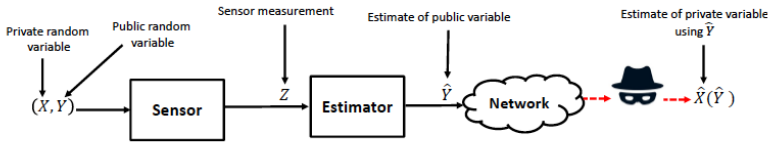




Nekouei et al., 2018 29

29

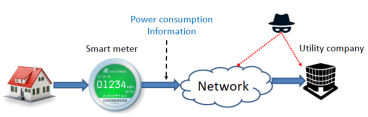
Remote Estimation under Privacy Constraints



- Y: Public random variable, e.g., HVAC data.
- X: Private random variable, e.g., occupancy information.
- Z: sensor measurement, noisy information regarding X and Y.
- $\hat{Y}$: estimate of Y based on Z.

Objectives:

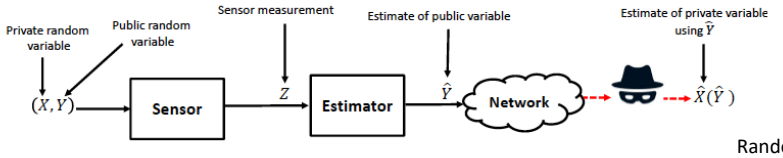
- Estimate Y reliably.
- Ensure private information cannot be inferred from $\hat{Y}$.



Nekouei et al., 2018 30

30

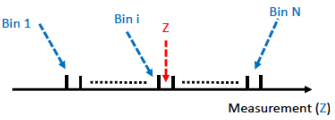
Privacy Metric



Randomized estimator

X (private variable) takes values in $\{x_1, \ldots, x_n\}$.

Y (public variable) takes values in $\{y_1, \ldots, y_m\}$.



$$\hat{Y}(z) = \begin{cases} y_1 & \text{w.p. } P_{1l}, \text{ if } z \in B_l \\ \vdots & \vdots \\ y_m & \text{w.p. } P_{ml}, \text{ if } z \in B_l \end{cases}$$

Privacy metric is chosen as conditional (discrete) entropy

$$\mathsf{H}\left[X \middle| \hat{Y}(Z)\right] = -\sum_{y \in \mathcal{Y}} \Pr\left(\hat{Y}(Z) = y\right) \sum_{x \in \mathcal{X}} \Pr\left(X = x | \hat{Y}(Z) = y\right) \log \Pr\left(X = x | \hat{Y}(Z) = y\right)$$

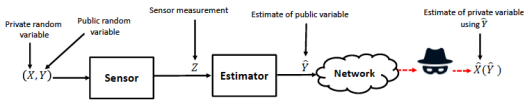
Captures the ambiguity of X after observing $\hat{Y}(Z)$.

Nekouei et al., 2018 31

31

Privacy-aware Estimator Design Problem

$$\begin{aligned} \underset{\{P_{il}\}_{i,l}}{\text{minimize}} \quad & \mathsf{E}\left[L\left(Y, \hat{Y}(Z)\right)\right] \\ & P_{il} \geq 0, \forall i, l \\ & \sum_i P_{il} = 1, \quad \forall l \\ & \mathsf{H}\left[X \middle| \hat{Y}(Z)\right] \geq \mathsf{H}_0 \end{aligned}$$



Randomized estimator

$$\hat{Y}(z) = \begin{cases} y_1 & \text{w.p. } P_{1l}, \text{ if } z \in B_l \\ \vdots & \vdots \\ y_m & \text{w.p. } P_{ml}, \text{ if } z \in B_l \end{cases}$$

- Loss function $L\left(Y, \hat{Y}(Z)\right) = \begin{cases} 1 & Y \neq \hat{Y}(Z) \\ 0 & Y = \hat{Y}(Z) \end{cases}$
- Randomization probabilities $\{P_{ji}\}$ are the optimization variables
- Privacy constraint $\mathsf{H}\left[X \middle| \hat{Y}(Z)\right] \geq \mathsf{H}_0$

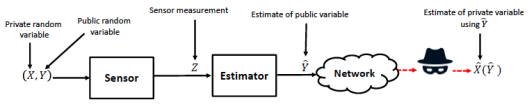
- Classical randomized estimator is recovered if privacy constraints are removed

Nekouei et al., 2018 32

32

Privacy-aware Estimator Design is Convex

$$\underset{\{P_{il}\}_{i,l}}{\text{minimize}} \quad \mathsf{E}\left[L\left(Y,\hat{Y}_{\cdot}(Z)\right)\right]$$
$$P_{il} \geq 0, \forall i,l$$
$$\sum_i P_{il} = 1, \quad \forall l$$
$$\mathsf{H}\left[X\left|\hat{Y}_P(Z)\right.\right] \geq \mathsf{H}_0$$



Randomized estimator

$$\hat{Y}_{\cdot}(z) = \begin{cases} y_1 & \text{w.p. } P_{1l}, \text{ if } z \in B_l \\ \vdots & \vdots \quad\quad \vdots \\ y_m & \text{w.p. } P_{ml}, \text{ if } z \in B_l \end{cases}$$

Theorem The privacy-aware estimator design problem is convex

"Proof:"

- Cost function $L\left(Y,\hat{Y}_{\cdot}(Z)\right) = \begin{cases} 1 & Y \neq \hat{Y}_{\cdot}(Z) \\ 0 & Y = \hat{Y}_{\cdot}(Z) \end{cases}$ is linear in $\{P_{ji}\}$
- Privacy constraint is convex (because Kullback-Leibler divergence is convex):

$$\mathsf{H}\left[X\left|\hat{Y}_{\cdot}(Z)\right.\right] = \mathsf{H}[X] - \sum_j \Pr(X = x_j)\, \mathsf{D}\left[p_{\hat{Y}_{\cdot}}(y|X = x_j)\left\|p_{\hat{Y}_{\cdot}}(y)\right.\right]$$

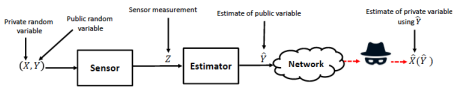
Nekouei et al., 2018 33

33

When Can Perfect Privacy be Achieved?

Perfect privacy corresponds to

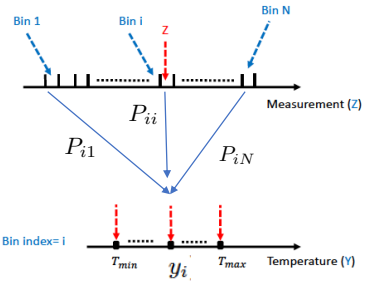
$$\Pr\left(\hat{Y}(Z) = y_i, X = x_j\right) = \Pr\left(\hat{Y}(Z) = y_i\right)\Pr(X = x_j)\,\forall i,j$$



Theorem Perfect privacy can be achieved if and only if $P_i \in \text{Null}(\Phi)$

$$P_i = \begin{bmatrix} P_{i1} \\ \vdots \\ P_{iN} \end{bmatrix}$$ is the probabilities associated with choosing y_i at bin 1,…,N

$$\Phi = \begin{bmatrix} \phi_{11} & \cdots & \phi_{1N} \\ \vdots & & \vdots \\ \phi_{n1} & \cdots & \phi_{nN} \end{bmatrix} \qquad \phi_{jl} = \Pr(Z \in B_l | X = x_j) - \Pr(Z \in B_l)$$



Nekouei et al., 2018 34

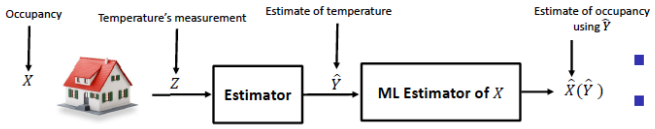
34

Example: Privacy-aware Estimator

X: occupancy (private)

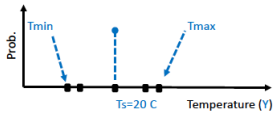
- $X = 1 \Longrightarrow$ Tenant is at home.
- $X = 0 \Longrightarrow$ Tenant is absent.

$\Pr(X = 1) = \Pr(X = 0) = 0.5.$



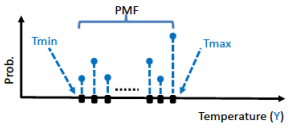
- Estimate temperature reliably.
- Ensure the occupancy remains private.

Temperature is equal to the set temperature.



Tenant is present (X=1)

Temperature is modeled as a random variable.



Tenant is absent (X=0)

- Z: sensor measurement

Sensor measurement ----> $Z = Y + N$ <---- Measurement noise

Nekouei et al., 2018 35

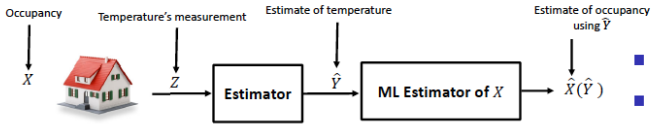
35

Example: Privacy-aware Estimator

X: occupancy (private)

- $X = 1 \Longrightarrow$ Tenant is at home.
- $X = 0 \Longrightarrow$ Tenant is absent.

$\Pr(X = 1) = \Pr(X = 0) = 0.5.$



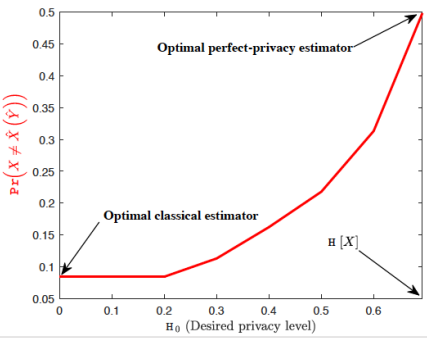
- Estimate temperature reliably.
- Ensure the occupancy remains private.

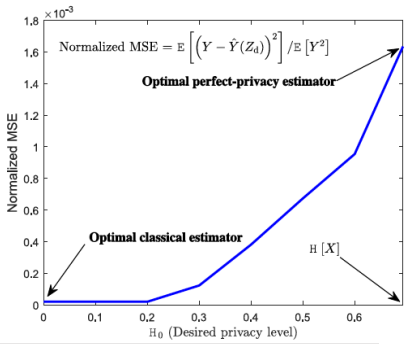
$$\underset{\{P_{ji}\}}{\text{minimize}} \quad \mathbb{E}\left[L\left(Y, \hat{Y}(Z)\right)\right]$$

$$P_{ji} \geq 0, \forall i$$

$$\sum_j P_{ji} = 1, \quad \forall i$$

$$\mathrm{H}\left[X \middle| \hat{Y}(Z)\right] \geq \mathrm{H}_0$$





Tradeoff between privacy and accuracy: guaranteed estimation loss in X while not too much MSE excess in $\hat{Y}$

Nekouei et al., 2018

36

Outline

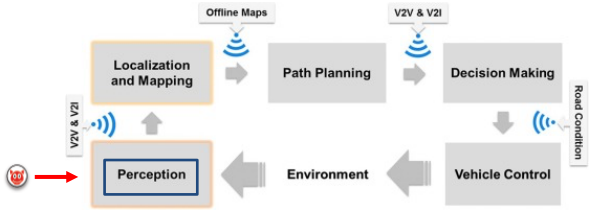
- What is control systems security?
- Motivation: Why should we care?
- Q1: How to conduct a false-data injection attack?
- Q2: How to defend against such attacks?
- Q3: How to do remote estimation under privacy constraints?
- Outlook and conclusions

37

37

Outlook: Adversarial learning

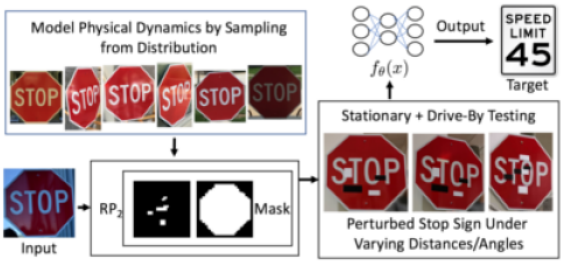
- Perception models for some automated vehicles are based on deep neural networks
- Trained from real or simulated data



[Brummelen et al., 2018]

- Vulnerable to attacks during training or operation, e.g., small perturbations to expected images

[Eykholt et al., CVPR, 2018] demonstrated attacks on deep learning visual classification algorithms, which then classfied stop signs with small markers as other road signs



How design safe learning schemes for automated vehicles?

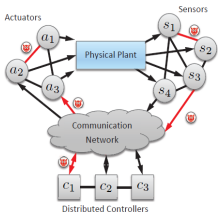
38

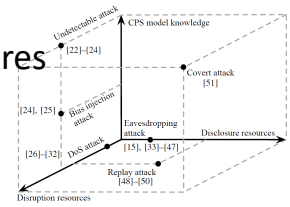
38

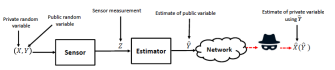
Conclusions

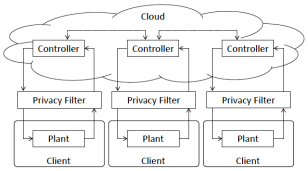
- Growing importance of **cyber-secure control systems,** motivated by evolving safety-critical infrastructure
- Design tools needed to complement other security measures
- Essential to model **adversary objectives** and **resources**
- False-data injection attacks and protection
- Remote estimation under privacy constraints
- Many **open problems**, including adversarial learning, cloud-based control, and operations under attack

people.kth.se/~kallej









39