FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Optimal Thresholds for Intrusion Detection Systems

Xenofon Koutsoukos

Waseem Abbas, Yevgeniy Vorobeychik, Amin Ghafouri

Aron Laszka, Shankar Sastry

Berkeley UNIVERSITY OF CALIFORNIA

Massachusetts Institute of Technology

NSF

MICHIGAN

VANDERBILT UNIVERSITY

(Recent) Cyber Attacks Against Cyber Physical Systems

Hackers caused power cut in western Ukraine

BBC NEWS
12 January 2016

Google Tool Aided N.Y. Dam Hacker

THE WALL STREET JOURNAL.
28 March, 2016

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Intrusion Detection Systems

Monitor a system for malicious activity

* When a malicious activity is detected, the **IDS raises an alarm** which can be investigated by operators.

For example

* By detection suspicious system call sequences
* By monitoring system files for modifications

Challenges

* Practical IDS are **imperfect**
  * might not detect attacks that do not result in "sufficient suspicious" activity
  * might raise false alarms for unusual but non-malicious activities.

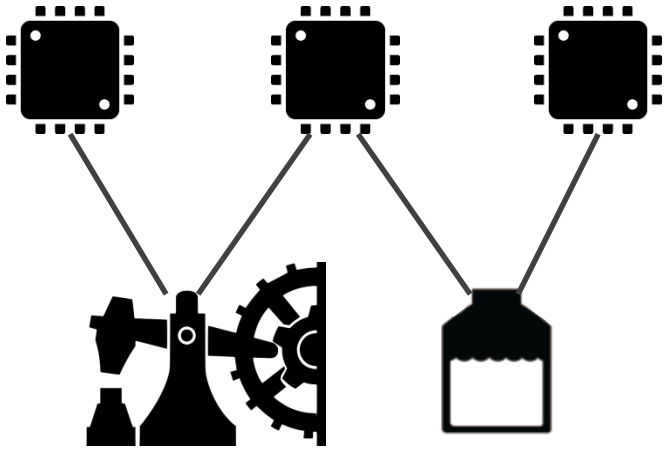
FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Configuration of IDS

* Finding an **optimal detection threshold** can prove to be a challenging problem even for a single IDS.
* Much more challenging when IDS are deployed on **multiple computer systems** that are interdependent with respect to the damage that could be caused by compromising hem.

Computer systems

Physical targets



Water distribution networks

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Objective

We study the problem of finding detection thresholds for multiple IDS in the face of strategic attacks.

Aron Laszka, Waseem Abbas, S. Shankar Sastry, Yevgeniy Vorobeychik, and Xenofon Koutsoukos. 2016. Optimal thresholds for intrusion detection systems. In *Proceedings of the Symposium and Bootcamp on the Science of Security* (HotSos '16). ACM, New York, NY, USA, 72-81.

Outline

- Introduction and Motivation
- Model of attacker and defender
- Attacker Defender game
- Best response attack
- Optimal Intrusion detection thresholds
- Numerical Evaluation
- Optimizing Thresholds for Time-Dependent Damage
- Future Directions

System Model

- Investigation of an alarm on system s cost, C_s.
- IDS are imperfect

Too many false alarms

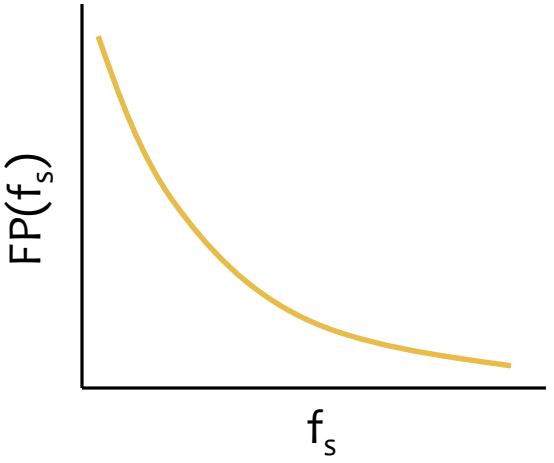
False positive errors

Detection threshold

Alarms not raised for actual attacks

False negative errors

- False negative probability: $\mathbf{f_s}$
- False positive rate: $\mathbf{FP(f_s)}$



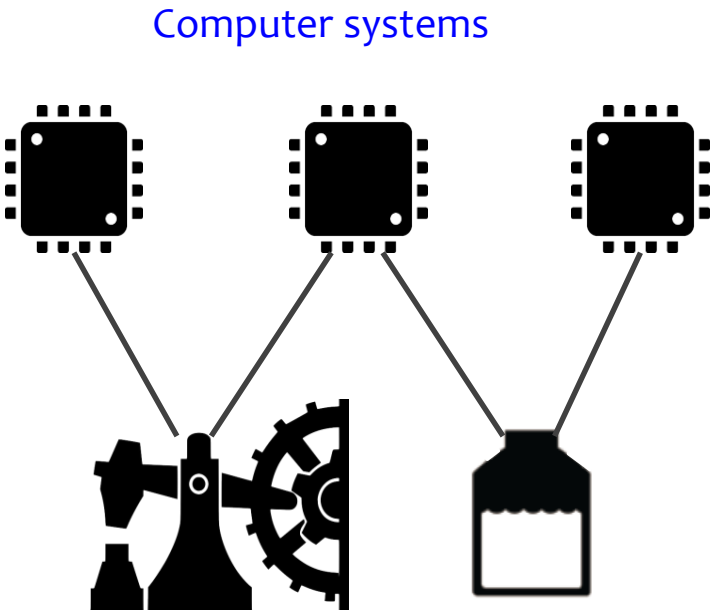
FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

System Model

* An attacker could attack a subset of systems, $A \subseteq S$
* The defender will detect the attack if the IDS of at least one targeted system raises an alarm.
* Probability that attack against systems in *A* is not detected is

$$\Pr\left[A \text{ is not detected}\right] = \prod_{s \in A} f_s$$

* An undetected attack will enable the attacker to cause **damage**



FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Attacker-Defender Game

Strategic Choices:

Defender:
Select false-negative probability f_s for each system.

Attacker:
Select a subset A of systems to attack.

Defender's Loss:

$$\mathcal{L}(\boldsymbol{f}, A) = \mathcal{D}(A) \prod_{s \in A} f_s + \sum_{s \in S} C_s \cdot FP_s(f_s),$$

Attacker's payoff:

$$\mathcal{P}(\boldsymbol{f}, A) = \mathcal{D}(A) \prod_{s \in A} f_s$$

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Attacker-Defender Game

* Attacker knows defender's algorithm, implementation etc.
* The defender cannot respond to the attacker's strategy, and must choose her strategy anticipating that the attacker will play a best response.

Best Response Attack:

$$\arg\max_{A \subseteq S} \mathcal{P}(\boldsymbol{f}, A)$$

Defender's Optimal Strategy

$$\arg\min_{\substack{\mathbf{0} \leq \boldsymbol{f} \leq \mathbf{1} \\ A \in \texttt{Best\_Response}(\boldsymbol{f})}} \mathcal{L}(\boldsymbol{f}, A)$$

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Best Response Attack

Theorem:

Given an instance of the model and configuration for the IDS, determining whether there exists an attack that causes at least a certain amount of damage is an NP-hard problem.

* Using reduction from a well-known NP-hard problem, the **Maximum Independent Set Problem**.
* In other words, it is computationally challenging even to determine how resilient a given configuration is.

Greedy Heuristic for Best Response Attack

Algorithm 1 Greedy Attack

```
Input S, f, D
Initialize: A ← ∅, P* ← 0
while doA ≠ S
    s ← argmax_{i∈S\A} P(f, A ∪ {i})
    if P(f, A ∪ {s}) > P* then
        A ← A ∪ {s}
        P* = P(f, A)
    else
        return A
    end if
end while
return A
```

Basic idea:

In each iteration, choose an element from S\A that maximally increases the attacker's payoff.

Proposition:

For any k>0, there is an instant of best response attack such that

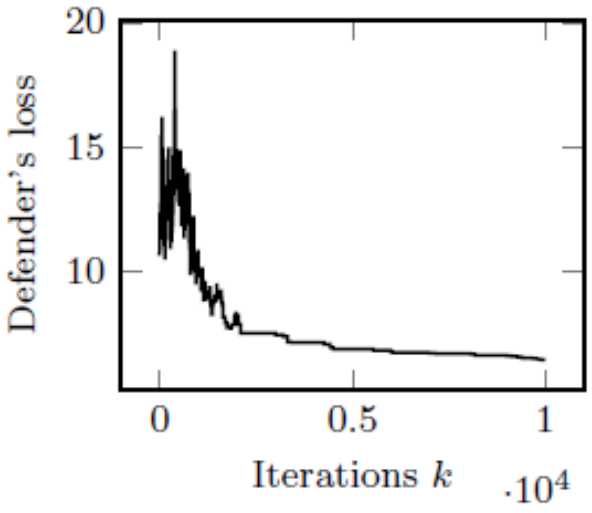
$$\frac{\mathcal{P}(\boldsymbol{f}, A^G)}{\mathcal{P}(\boldsymbol{f}, A^*)} < k$$

Where A^G is output of greedy heuristic and A* is the best response attack.

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Heuristics for Intrusion Detection Thresholds

- **Simulated Annealing** based polynomial time meta-heuristic.
- Iterative improvements until convergence.



Input S, $\mathcal{D}$, $\boldsymbol{C}$, $k_{\max}$
Initialize: $\boldsymbol{f}$, $k \leftarrow 1$, T_0, β
$A \leftarrow \texttt{Best\_Response\_Attack}(\boldsymbol{f})$
$L \leftarrow \mathcal{L}(\boldsymbol{f}, A)$
while $k \leq k_{\max}$ **do**
  $\boldsymbol{f}' \leftarrow \texttt{Perturb}(\boldsymbol{f}, k)$
  $A' \leftarrow \texttt{Best\_Response\_Attack}(\boldsymbol{f}')$
  $L' \leftarrow \mathcal{L}(\boldsymbol{f}', A')$
  $c \leftarrow e^{(L'-L)/T}$
  if $(L' < L) \vee (\texttt{rand}(0,1) \leq c)$ **then**
    $\boldsymbol{f} \leftarrow \boldsymbol{f}'$, $L \leftarrow L'$
  end if
  $T \leftarrow T_0 \cdot e^{-\beta k}$
  $k \leftarrow k + 1$
end while
return $\boldsymbol{f}$

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Baseline Strategies for Comparison

* **Uniform Threshold Strategy:**

- All systems are assigned the same false negative probability, i.e., $f_s = f$, for all s in S.
- The value of f is chosen to minimize the defender's loss

* **Locally Optimum Strategy:**

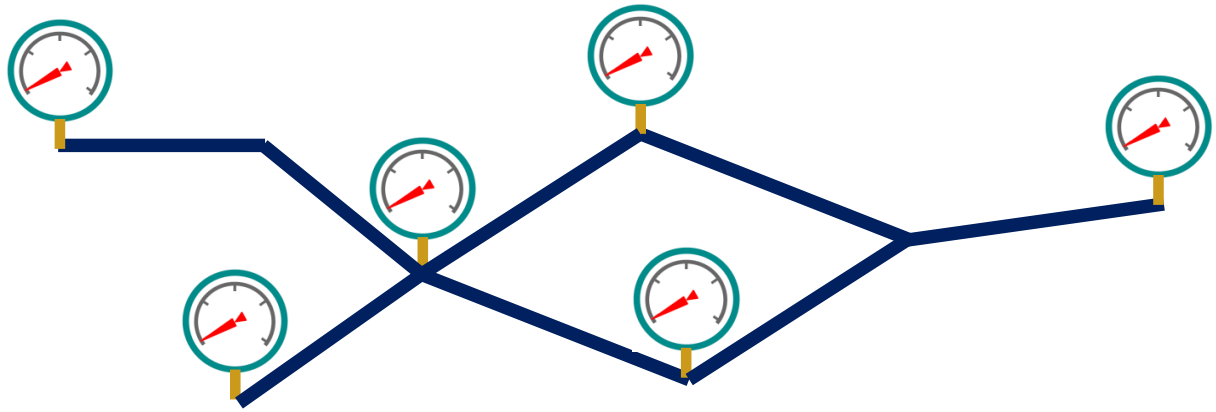
- For each system s, f_s is individually optimized.
- For each s, f_s is chosen to minimize

$$\mathcal{L}(f_s, \{s\}) = \mathcal{D}(\{s\}) f_s + C_s \,.\, FP(f_s)$$

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Numerical Illustration – Water Dist. Network

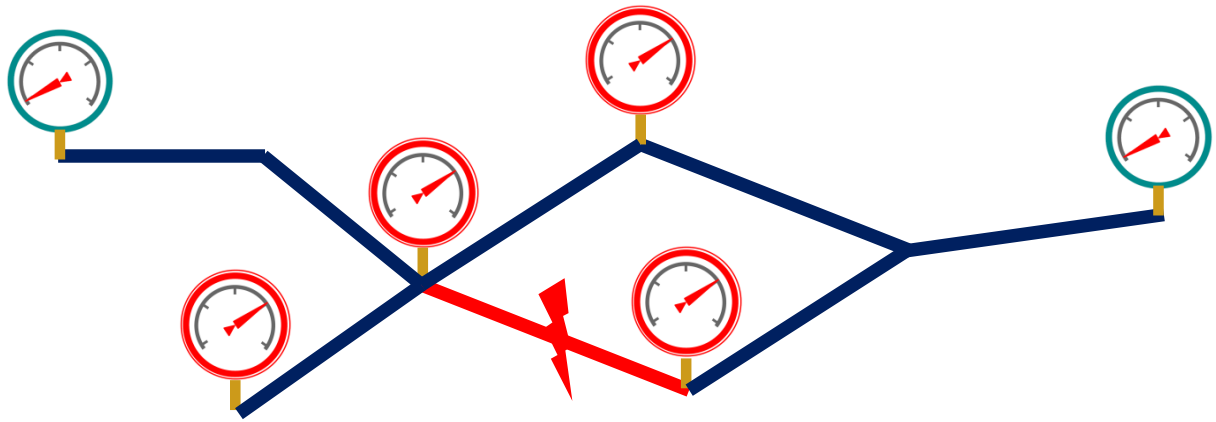
* **Leakages in Water Distribution Networks:**
  - **Leakages** in water distribution networks can cause significant losses and third-party damage
  - **Pressure sensors** can detect "nearby" pipe bursts



FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Numerical Illustration – Water Dist. Network

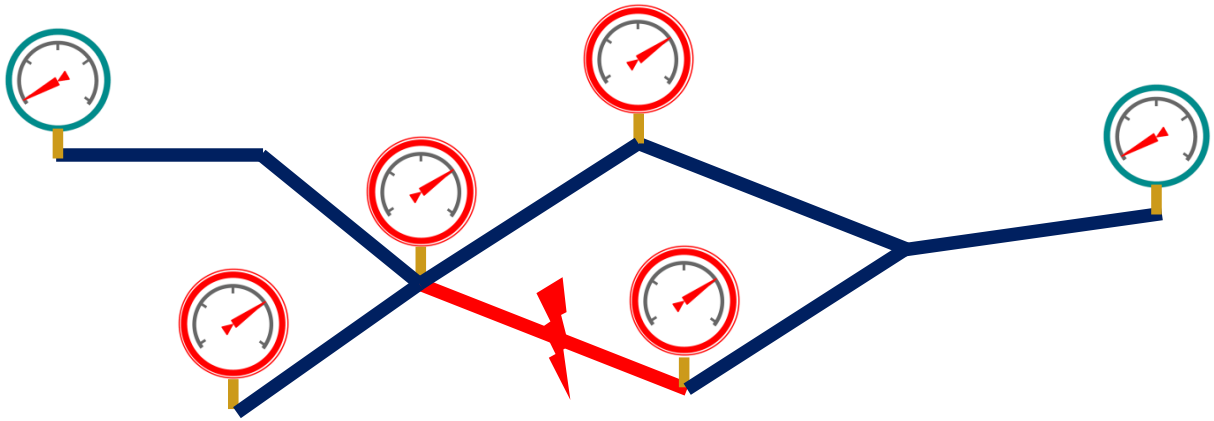
* **Leakages in Water Distribution Networks:**
  - **Leakages** in water distribution networks can cause significant losses and third-party damage
  - **Pressure sensors** can detect "nearby" pipe bursts



FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Numerical Illustration – Water Dist. Network

* **Leakages in Water Distribution Networks:**
  - **Leakages** in water distribution networks can cause significant losses and third-party damage
  - **Pressure sensors** can detect "nearby" pipe bursts

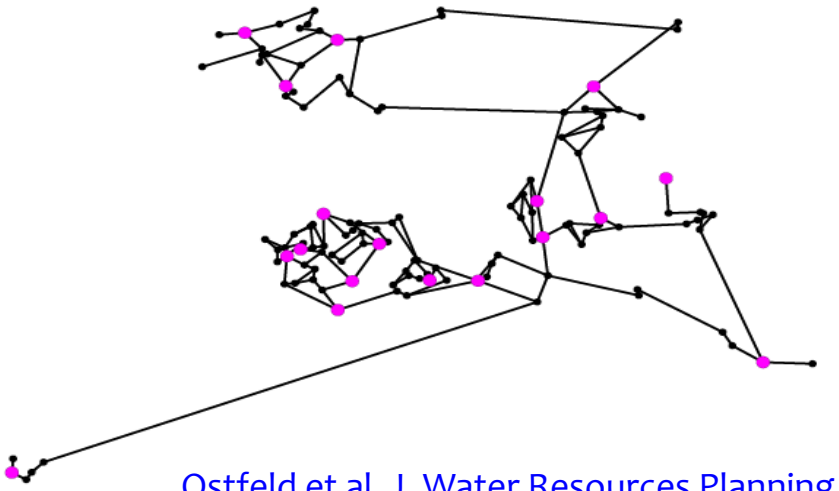


  - **Attacker** may tamper with sensors to cause damage
  - IDSs can be deployed on the sensors to detect **cyber-attacks**

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Numerical Illustration – Water Dist. Network

Water Network:



- **168 pipes** and **126 nodes**
- A sensor monitors pipes that are at most D = 3 distant from the sensing node.
- **18 sensors** are sufficient to monitor the whole network.

Ostfeld et al. J. Water Resources Planning and Management, 2008.

- **S** : set of sensors that need to be defended.
- **D(A)**: number of pipes monitored by the sensors in A.
- C_s: cost of investigating a false alarm on sensor s.

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

False Positive and False Negative Error Rates

* As an example, we use the ADFA-LD dataset to train an IDS that monitors system-call sequences

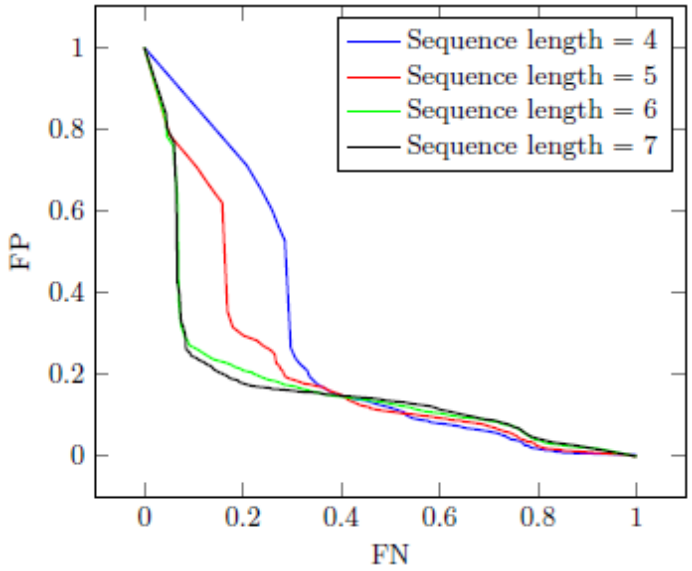


Figure: Attainable false-positive and false-negative error rates (i.e., fractions of misreported normal and attack traces, respectively) of the IDS for various sequence lengths.

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

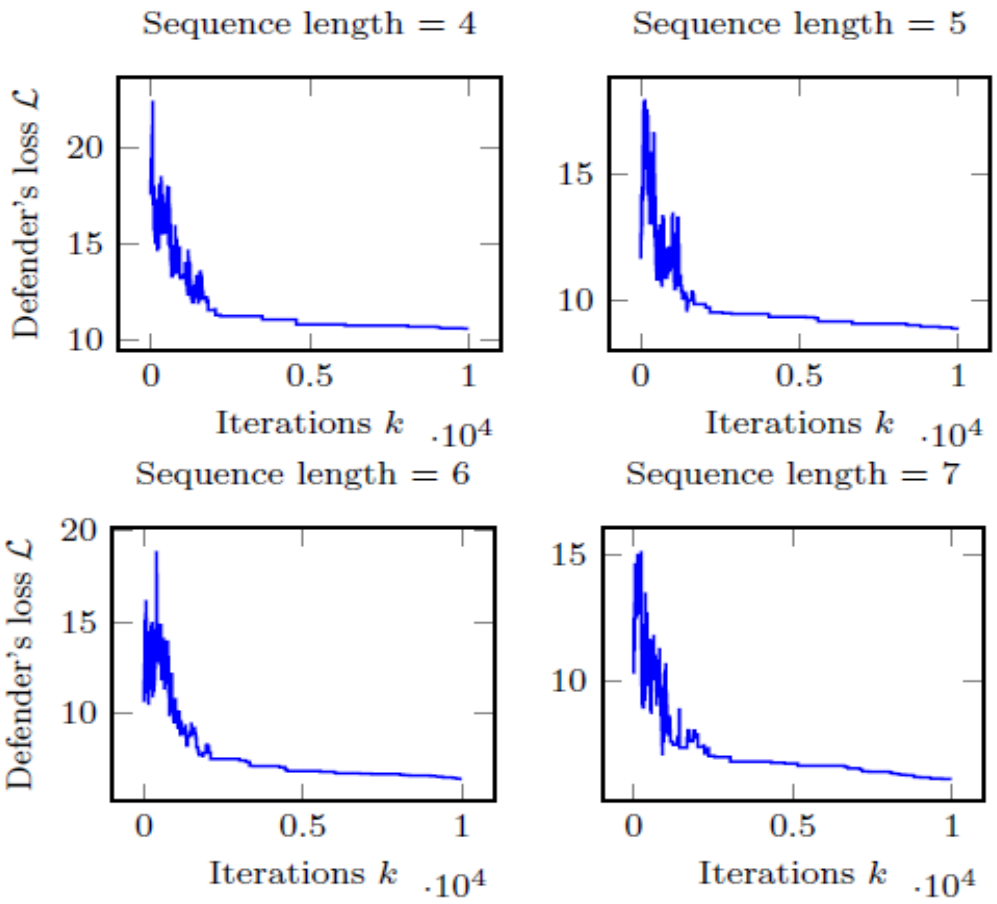
Greedy Attack vs. Best Response Attack

* Comparison Between Best-Response Attacks and the Output of Algorithm 1

n	Fraction of instance where greedy and best-response payoffs are equal	Worst case ratio between greedy and best-response payoffs
2	100 %	100 %
3	99.9 %	97.99 %
4	99.5 %	93.41 %
5	98.2 %	86.03 %
6	98.1 %	85.62 %
7	96.1 %	75.27 %
8	94.9 %	82.72 %
9	95.2 %	82.7 %
10	95.7 %	77.32 %

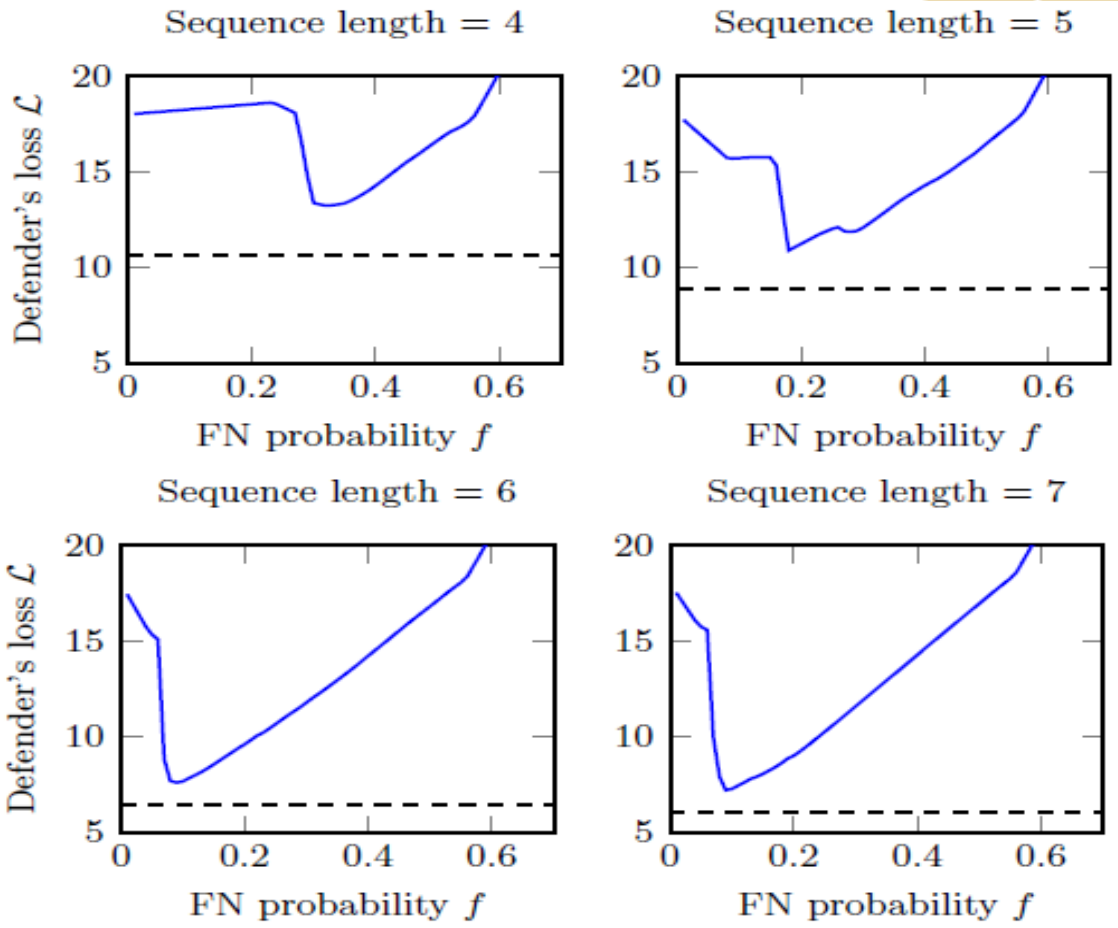
* Greedy heuristics provide a good way to approximate the best response attacks for practical purposes.

Convergence of Algorithm



FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

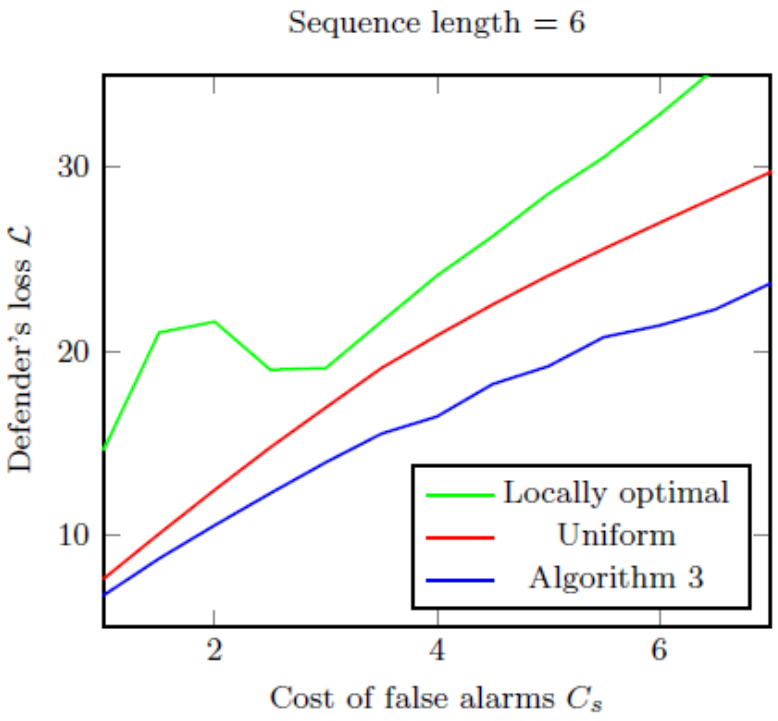
Numerical Results - Comparisons

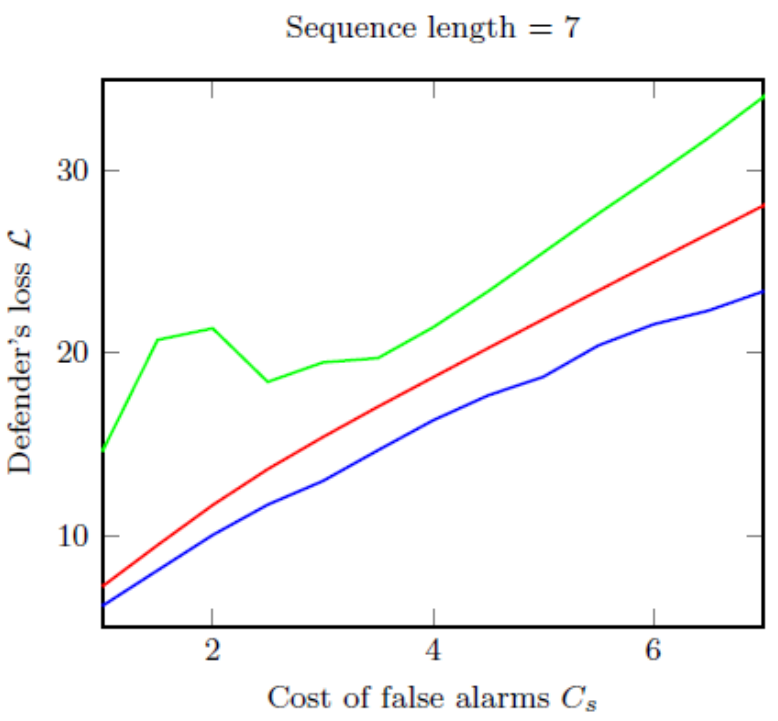


Comparison of proposed algorithm with uniform threshold and locally optimum threshold strategies.

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Numerical Results - Comparisons



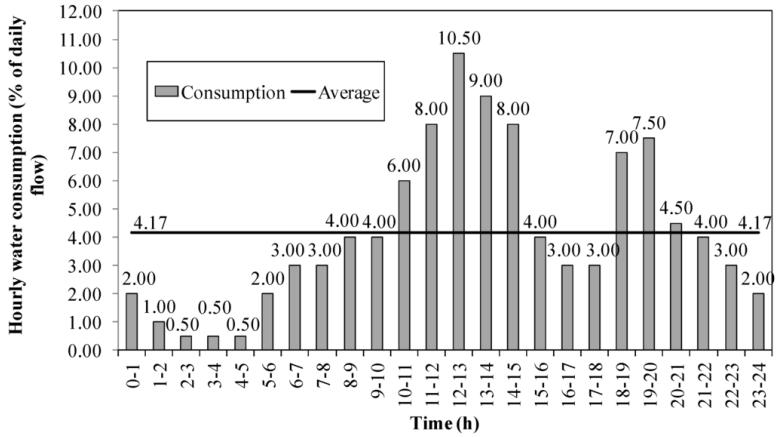


Defender's loss using three different strategies (uniform, locally optimal, and our Algorithm) as a function of the cost of false alarms.

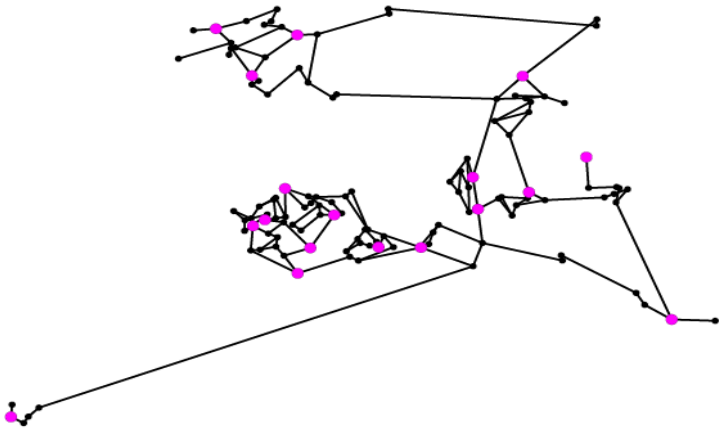
FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Time-Dependent Damage

* In CPS, the expected damage incurred from undetected attacks varies by time, and depends on the state of the physical systems



* There is the need to incorporate the dynamic behavior in computing optimal thresholds when facing strategic attackers



FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

System Model

* System's time horizon of interest, $T = \{k_0, ..., k_m\}$.
* Attack period, $\{k_a, ..., k_e\} \subseteq T$.
* Damage function $\mathcal{D} : T \to \mathbb{R}_+$ represents the expected damage $\mathcal{D}(k)$ incurred to a CPS from an undetected attack at time $k \in T$.
* Expected total damage represents the expected damage $\sum_{k_a}^{k_e} \mathcal{D}(k)$ from an undetected attack in a period $\{k_a, ..., k_e\} \subseteq T$.

Sequential Change Detection

* Sequential change detection: tradeoff between false alarm rate and detection delay

Too many false alarms

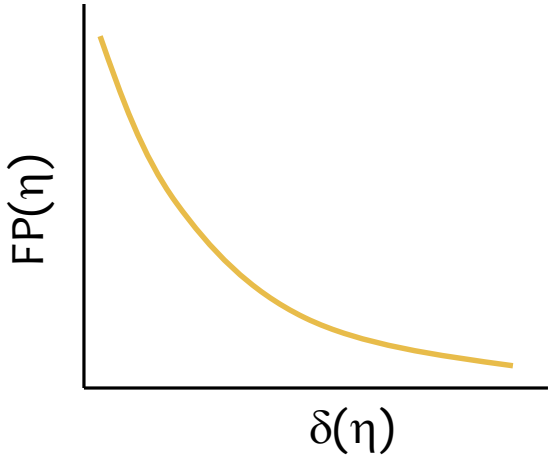
False positive errors

Too much detection delay

Detection delay

Detection threshold

* Threshold: η
* Detection delay: $\delta(\eta)$
* False positive rate: **FP(η)**



FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Attacker-Defender Game

Strategic Choices:

Defender:
Select threshold η for the detector.

Attacker:
Select a time k_a to start an attack.

Defender's Loss:

$$\mathcal{L}(\eta, k_a) = C \cdot FP(\eta) \cdot |T| + \sum_{k=k_a}^{k_a+\delta(\eta)} \mathcal{D}(k)$$

Attacker's payoff:

$$\mathcal{P}(\eta, k_a) = \sum_{k=k_a}^{k_a+\delta(\eta)} \mathcal{D}(k)$$

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Adaptive Threshold

* Reduce detector's sensitivity during less critical periods and increase sensitivity during more critical periods
* Significantly decreases the defender's loss
* Adaptive threshold $\Psi = (\gamma, \eta)$ represents a set of pairs $\Psi_j = (\gamma_j, \eta_j)$ where γ_j is a threshold change time, η_j is a corresponding threshold value, $j \in \{0, ..., N-1\}$.
* Expected detection delay $\Delta(\Psi, k_a)$ changes as threshold changes.

Adaptive Threshold Game

Strategic Choices:

Defender:
Select threshold schedule ψ for the detector.

Attacker:
Select a time k_a to start an attack.

Defender's Loss:

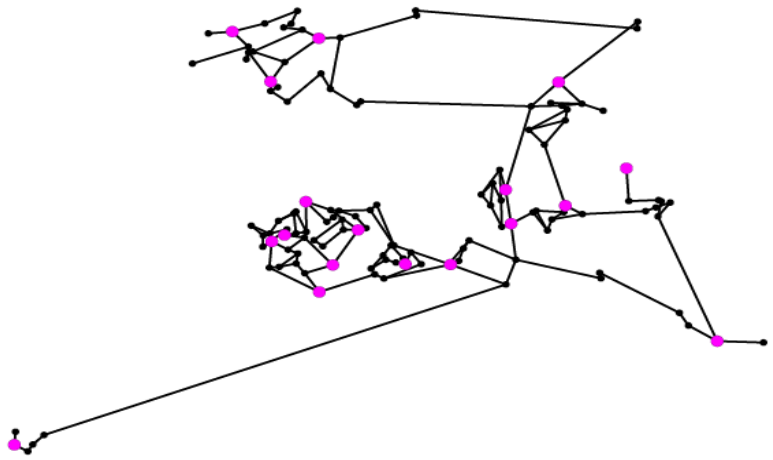
$$\mathcal{L}(\Psi, k_a) = (N-1)\, C_d + \sum_{j=0}^{N-1} \frac{|\Gamma_j|}{|T|} \cdot C \cdot FP(\eta_j) + \sum_{k_a}^{k_a+\delta(\Psi,k_a)} \mathcal{D}(k)$$

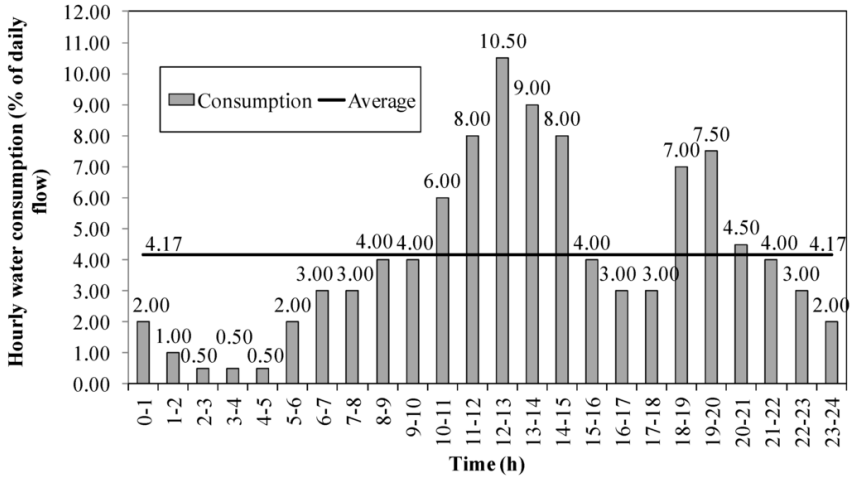
Attacker's payoff:

$$\mathcal{P}(\Psi, k_a) = \sum_{k_a}^{k_a+\delta(\Psi,k_a)} \mathcal{D}(k)$$

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Water Distribution System Example





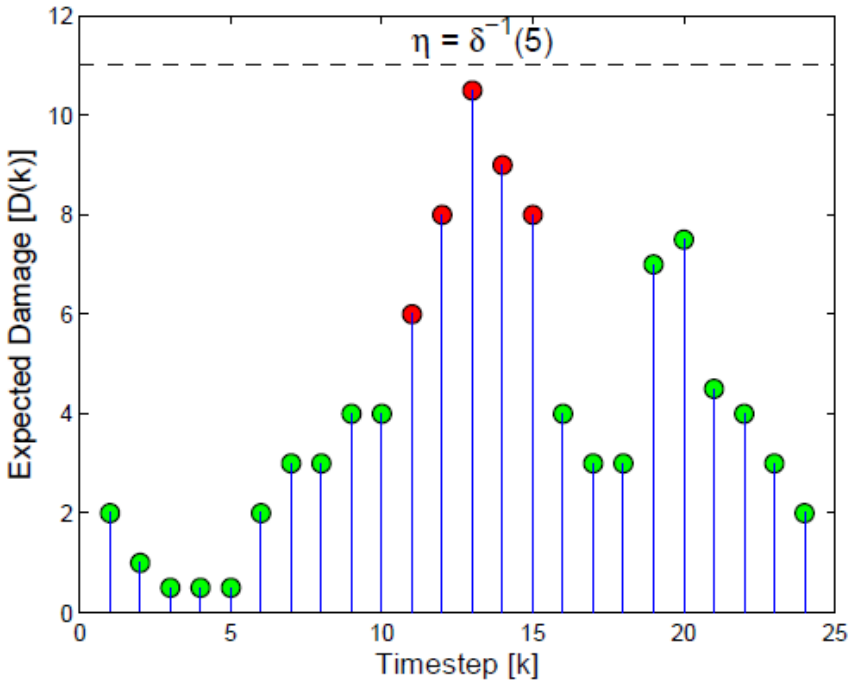
* Detectors installed on a pressure sensing devices
* Attacker alters sensor measurements
* Damage is a function of demand pattern

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

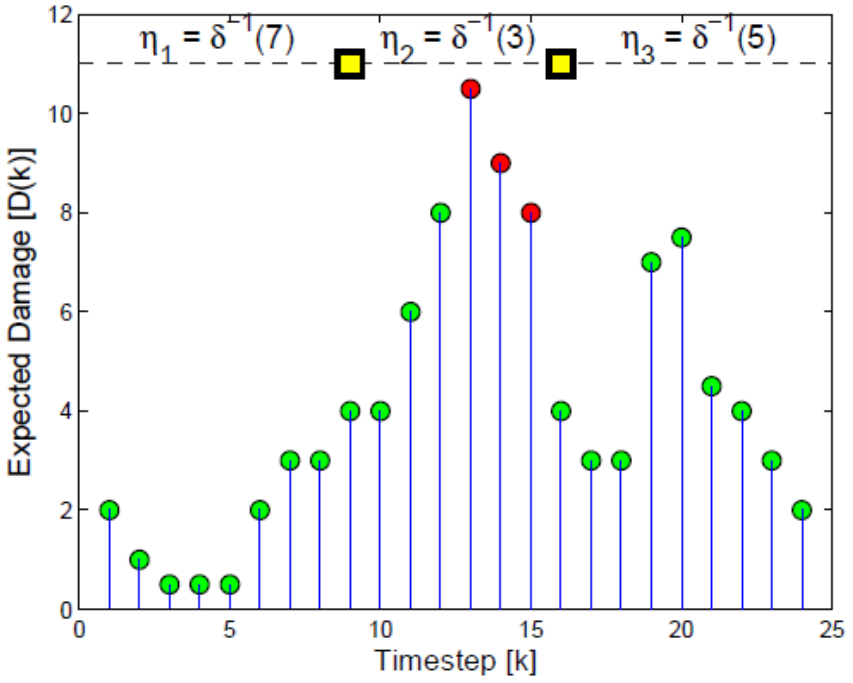
Numerical Results

* Optimal Threshold and Best-Response Attack

Fixed Threshold



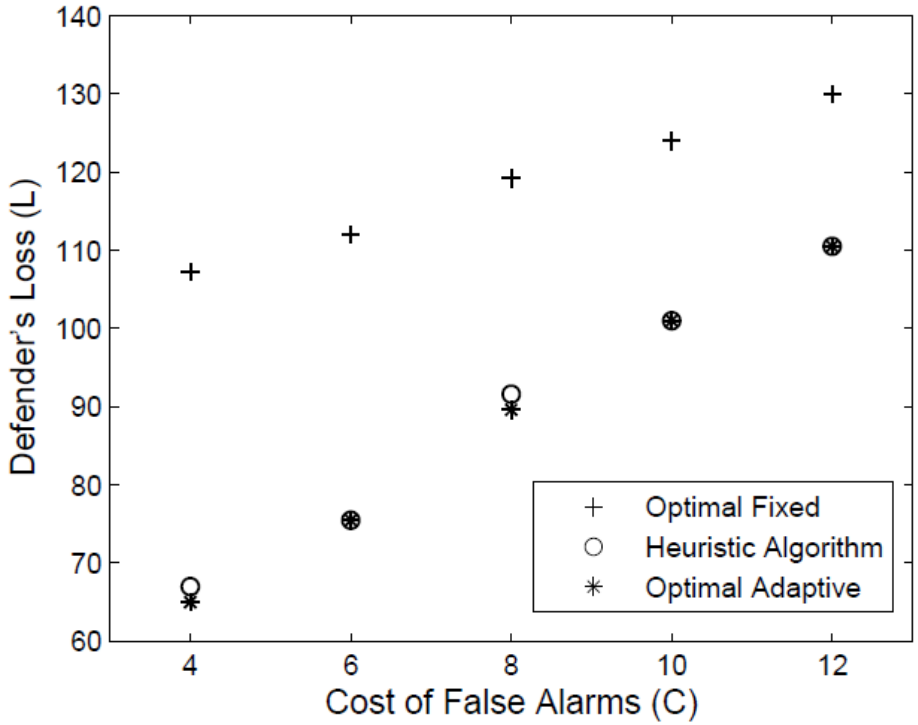
Adaptive Threshold



FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Numerical Results

* Adaptive threshold reduces the defender's loss by almost 40%.



FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS

Conclusions and Future Directions

* By taking into account the characteristics of the physical processes controlled by the computational elements, we can
  * Increase the probability of detecting cyber-attacks
  * Decrease losses due to cyber-attacks and false alarms
* In future, we would like to incorporate
  * Multiple systems: Different time-varying damage for each subsystem
  * Hypothesis testing: Tradeoff between false alarm rate, missed detection rate, and detection delay
  * Moving target defense techniques based on randomized thresholds

FORCES
FOUNDATIONS OF RESILIENT CYBER-PHYSICAL SYSTEMS