Optimal Thresholds for Intrusion Detection Systems

A. Laszka[1], W. Abbas[2], S. Sastry[1], Y. Vorobeychik[2], X. Koutsoukos[2]

[1] University of California, Berkeley
[2] Vanderbilt University

Berkeley UNIVERSITY OF CALIFORNIA

VANDERBILT UNIVERSITY

Cyber Attacks Against Cyber Physical Systems

- Cyber physical systems are vulnerable to (cyber) **attacks.**
- Successful attacks might result in **severe damages**.

Maroochy water breach (2000)

Stealthy

Stuxnet worm (2010)

Cyber attack on German steel plant (2014)

Cyber attack on Turkish oil pipeline (2008)

Cyber Attacks Against Cyber Physical Systems

More recently,

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

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

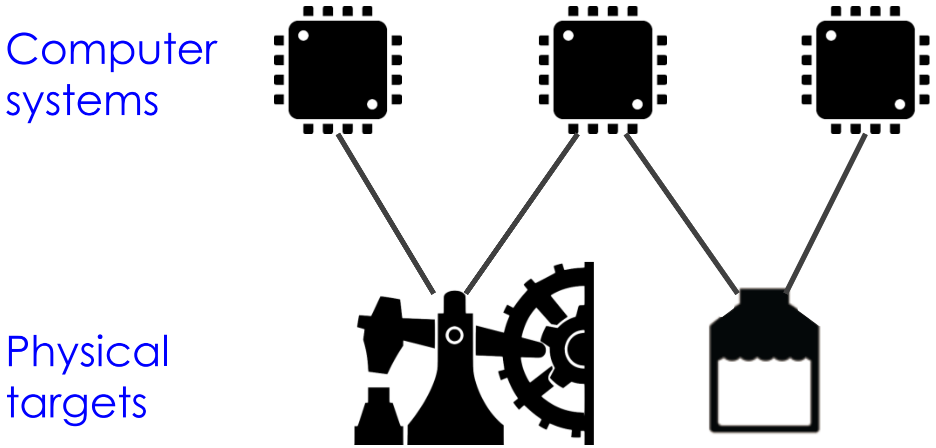
- Practical IDS are **imperfect**

might not detect attacks that do not result in “sufficient suspicious” activity

might raise false alarms for unusual but non-malicious activities.

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.



Water distribution networks

Objective

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

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

Water distribution networks

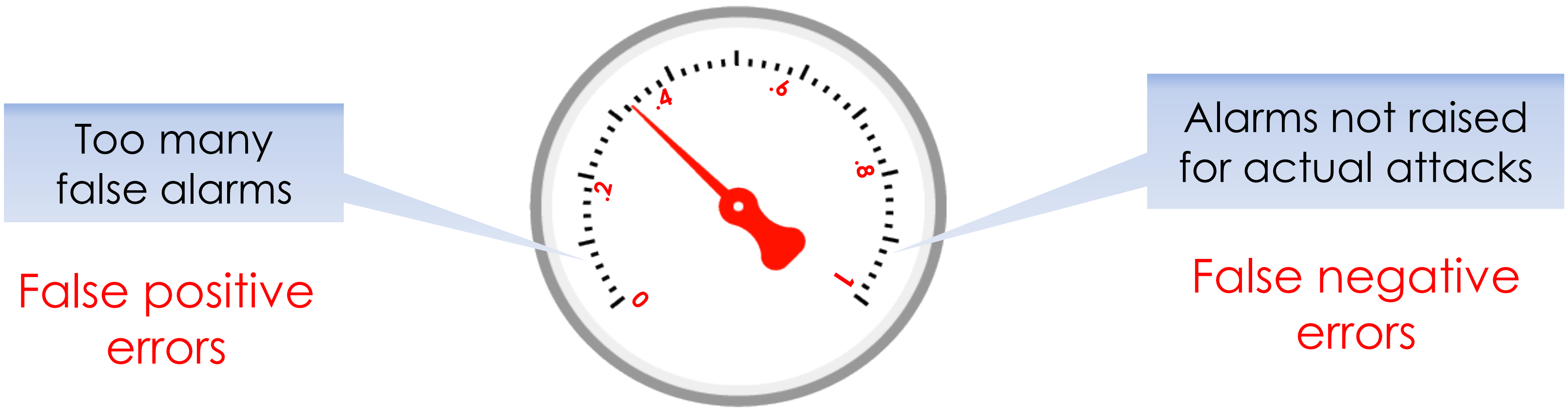
Smart grids.

Outline

- Introduction and Motivation
- Model of attacker and defender
- Attacker Defender game
- Best response attack
- Optimal Intrusion detection thresholds
- Numerical Evaluation
- Future Directions

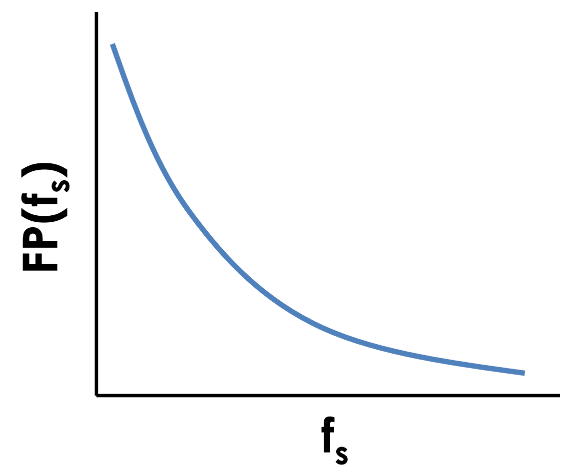
System Model

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



Detection threshold

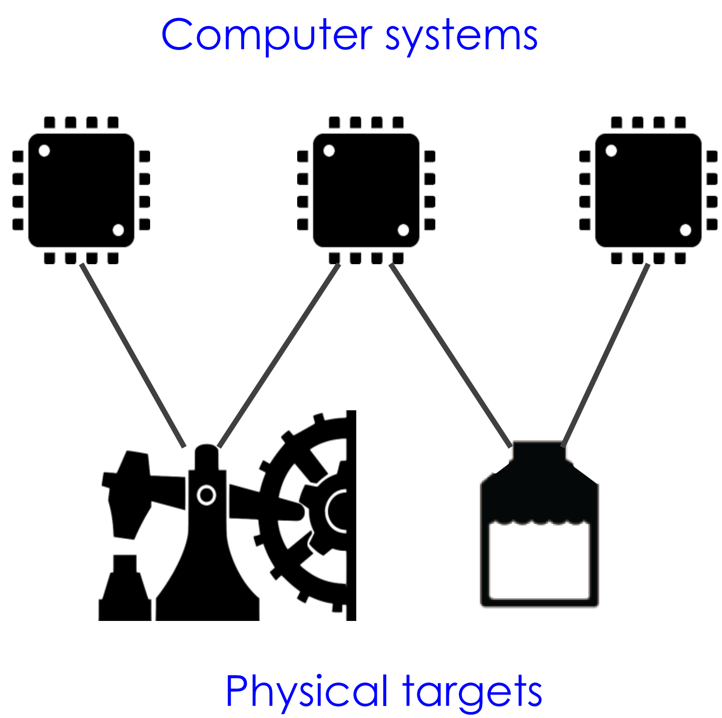
- False negative probability: $\mathbf{f_s}$
- Fale positive rate: $\mathbf{FP(f_s)}$



System Model

- An attacker could attack a subset of systems, $\boldsymbol{A} \subseteq \boldsymbol{S}$
- 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** $\mathcal{D}(A)$.

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:

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

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

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

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 $S, f, \mathcal{D}$
Initialize: $A \leftarrow \emptyset$, $P^* \leftarrow 0$
while **do**$A \neq S$
  $s \leftarrow \mathrm{argmax}_{i \in S \setminus A} \mathcal{P}(f, A \cup \{i\})$
  if $\mathcal{P}(f, A \cup \{s\}) > P^*$ **then**
    $A \leftarrow A \cup \{s\}$
    $P^* = \mathcal{P}(f, A)$
  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.

Alternate Heuristic for Best Response Attack

Algorithm 2 Alternate Linear-Time Attack

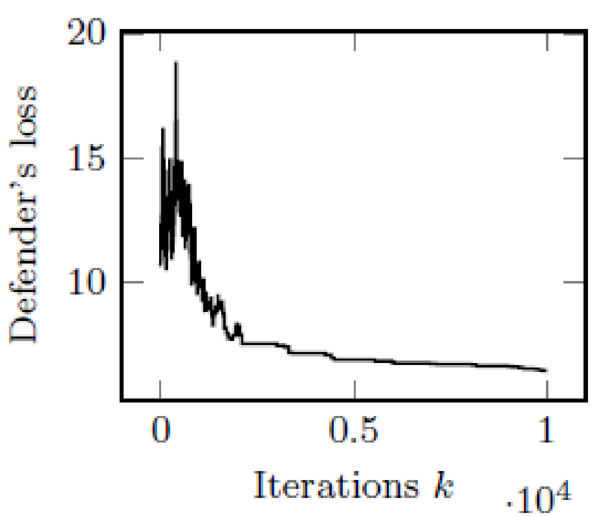
Input $S, \boldsymbol{f}, \mathcal{D}$
Initialize: $X \leftarrow \emptyset$, $Y \leftarrow S$,
Arrange elements of S in an arbitrary order
for $i = 1$ to $|S|$ **do**
    $x_i \leftarrow \mathcal{P}(\boldsymbol{f}, X \cup \{i\}) - \mathcal{P}(\boldsymbol{f}, X)$
    $y_i \leftarrow \mathcal{P}(\boldsymbol{f}, Y \cup \{i\}) - \mathcal{P}(\boldsymbol{f}, Y)$
    if $x_i \geq y_i$ **then**
        $X \leftarrow X \cup \{i\}$
    else
        $Y \leftarrow Y \setminus \{i\}$
    end if
end for
$A \leftarrow X$ (or equivalently Y since $X = Y$)
return A

- Runs in linear time.
- Gives a (1/3)-approximate solution if $\mathcal{P}(\boldsymbol{f}, A)$ is submodular.

Buchbinder et al. SIAM J Computing, 2015

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 $\boldsymbol{f}$

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:

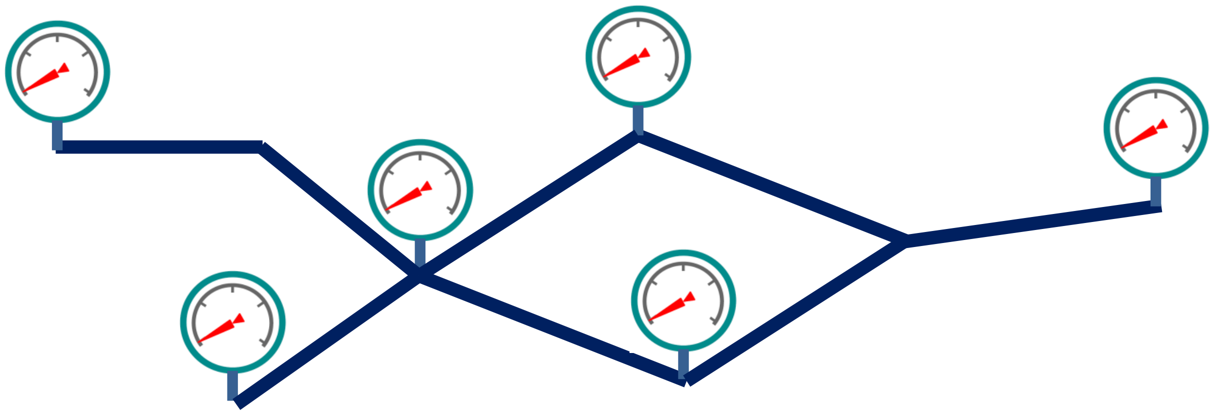
- For each system s, f_s is individually optimized.
- For each s, f_s is chosen to minimize

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

Numerical Illustration – Water Distribution Network

Leakages in Water Distribution Networks:

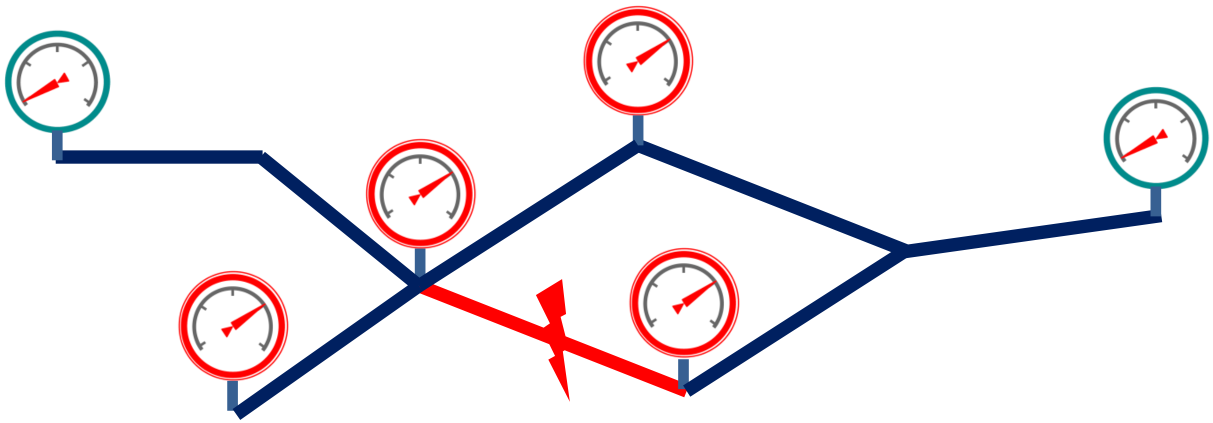
- **Leakages** in water distribution networks can cause significant losses and third-party damage
- **Pressure sensors** can detect "nearby" pipe bursts



Numerical Illustration – Water Distribution Network

Leakages in Water Distribution Networks:

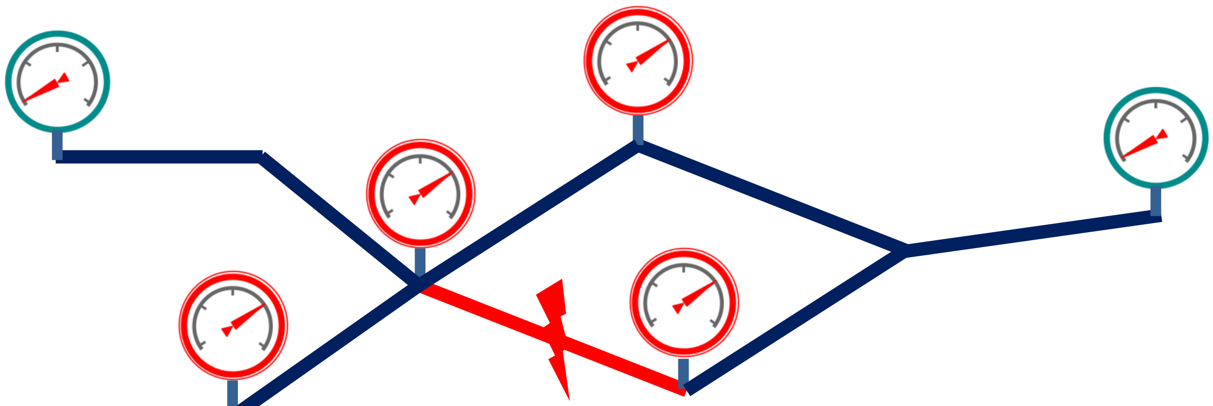
- **Leakages** in water distribution networks can cause significant losses and third-party damage
- **Pressure sensors** can detect "nearby" pipe bursts



Numerical Illustration – Water Distribution Network

Leakages in Water Distribution Networks:

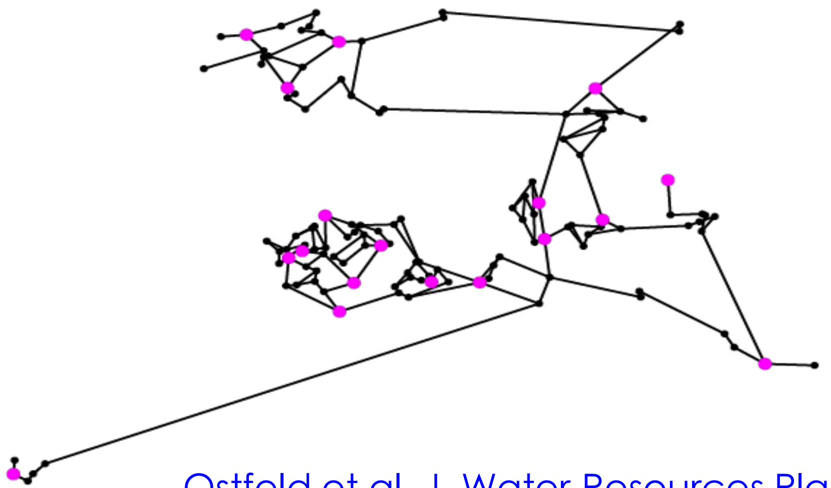
- **Leakages** in water distribution networks can cause significant losses and third-party damage
- **Pressure sensors** can detect "nearby" pipe bursts



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

Numerical Illustration – Water Distribution 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.

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

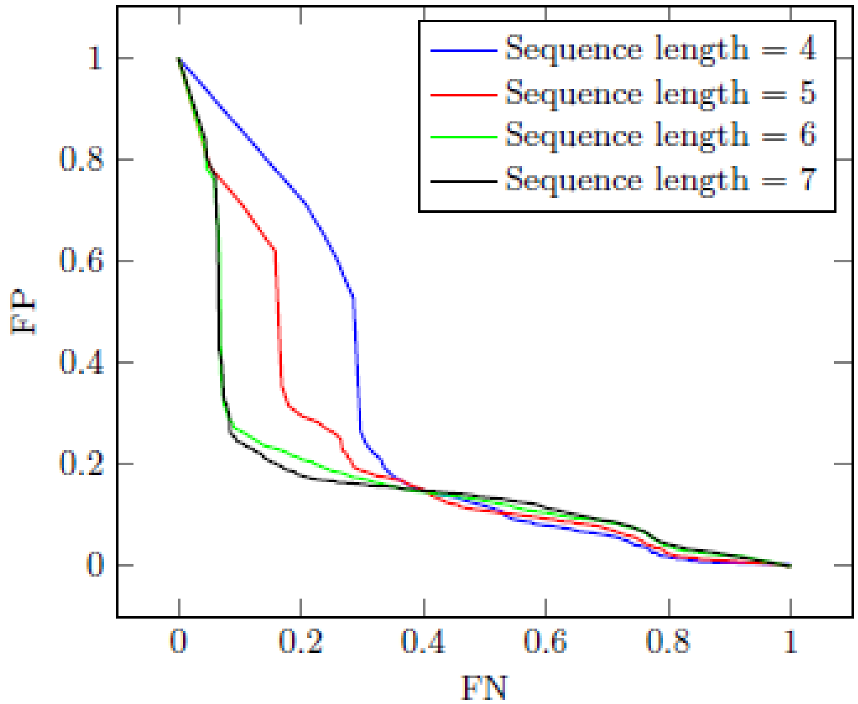


Figure: Attainable false-positive and false-negative error rates (i.e., fractions of misreported normal and attack traces, respectively) of the IDS for various sequence lengths.

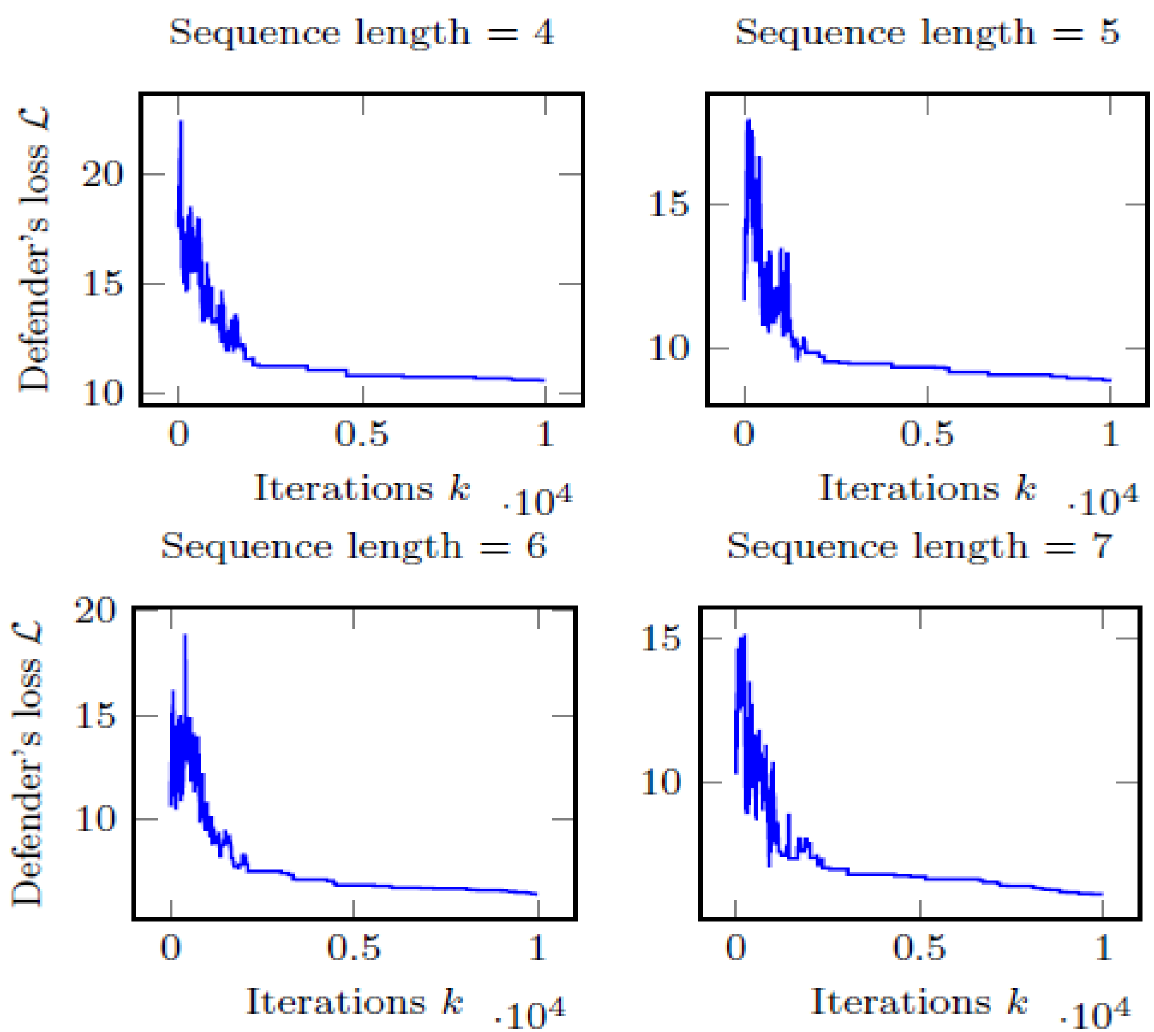
Greedy Attack vs. Best Response Attack

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

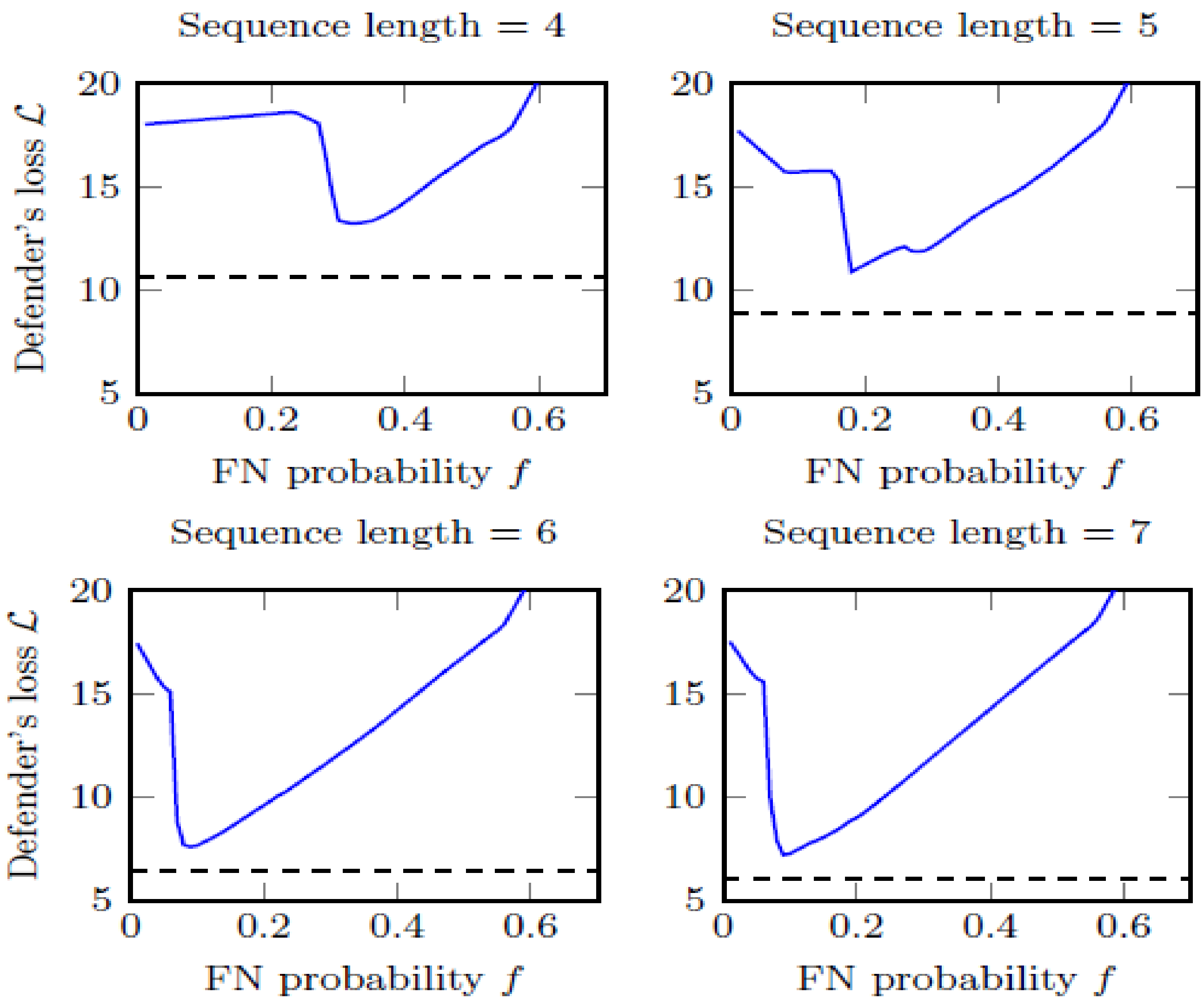
n	Fraction of instance where greedy and best-response payoffs are equal	Worst case ratio between greedy and best-response payoffs
2	100 %	100 %
3	99.9 %	97.99 %
4	99.5 %	93.41 %
5	98.2 %	86.03 %
6	98.1 %	85.62 %
7	96.1 %	75.27 %
8	94.9 %	82.72 %
9	95.2 %	82.7 %
10	95.7 %	77.32 %

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

Convergence of Algorithm for Detection Thresholds

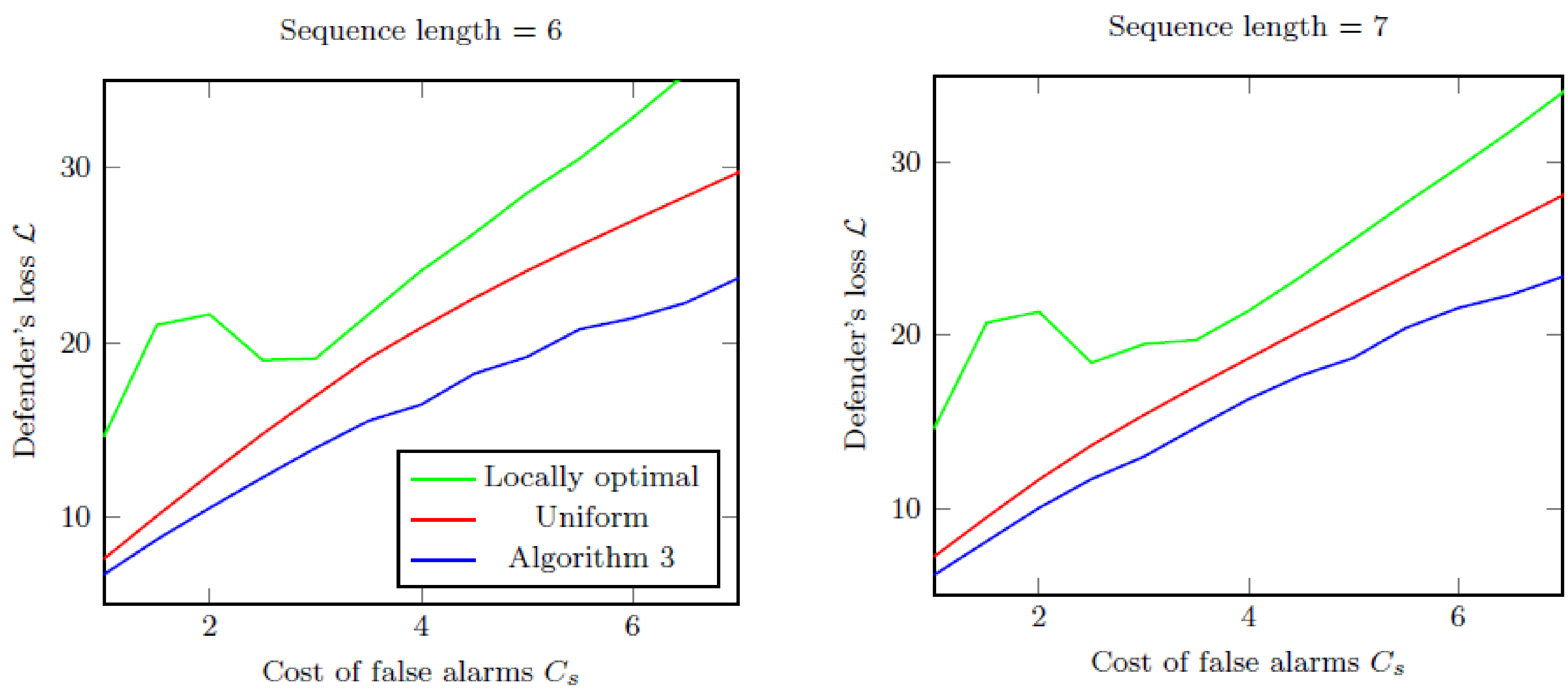


Numerical Results - Comparisons



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

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.

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
  - more realistic IDS models, and
  - more generalized damage functions to accommodate a wide variety of applications.
  - Moreover, simultaneous scheduling and configuration of IDS could further improve the overall detection performance.

Thank You