Attacks and Detection in ICS

CPS and IoT Security

Alessandro Brighente

Master Degree in Cybersecurity

Università degli Studi di Padova

SPRITZ Security & Privacy Research Group

Vulnerabilities of ICS

SPRITZ
SECURITY & PRIVACY
RESEARCH GROUP

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

- The major distinction of control systems with respect to other IT systems is the interaction with the physical world
- While measures for IT systems may be good for IT systems, they might not be sufficient for ICS
- It is important not only to protect information, but also to protect estimation and control algorithms and how they impact the physical world

Risk Assessment

- Risk management = shifting the odds in your favor by finding the alternative that minimizes the impact of certain events
- The best well-known metric is the average loss $R_\mu = \mathbb{E}[\dot{L}] \approx \sum_i L_i p_i$
- Multiply the loss implied by event i by the probability that event i occurs
- There are also other measures, for instance the variance of the losses
- $R_\chi = \mathbb{E}[L^2] - R_\mu$
- However, this is important over long periods of time so not very suitable for our case, average loss is more than enough

Attack Model

- We focus on attacks on sensor networks and their effect on process control
- In our case, p_i denotes the likelihood that an attacker will compromise sensor i, and L_i denotes the loss associated with that particular compromise
- We assume that the likelihood is the same for all sensors, and we focus on the estimate of the potential loss

Attack Model

- Let's consider a network of p sensors with measurement vector
  $$y(k) = \{y_1(k), y_2(k), \cdots, y_p(k)\}$$
- All sensors have a dynamic range that defines the domain of y for all k
- I.e., $\forall k, y_i(k) \in [y_i^{min}, y_i^{max}] = \mathcal{Y}_i$
- We assume each sensor has a unique identity protected by a cryptographic key
- Let $\tilde{y}(k) \in \mathbb{R}^p$ denote the received measurement by the controller at time k

Attack Model

- Based on these measurements, the control system defines control actions to maintain certain operational goals
- We assume that attacked signals lie within $\mathcal{Y}_i$
- If the malicious value is outside the legitimate range, it can be simply detected by fault-tolerant algorithms
- Let $\mathcal{K}_a = \{k_s, \ldots, k_e\}$ denote the attack duration
- Given attack signal $a_i(k)$, the general model for the observed signal

$$\tilde{y}_i(k) = \begin{cases} y_i(k) & \text{for } k \notin \mathcal{K}_a \\ a_i(k) & \text{for } k \in \mathcal{K}_a, a_i(k) \in \mathcal{Y}_i \end{cases}$$

Attack Model

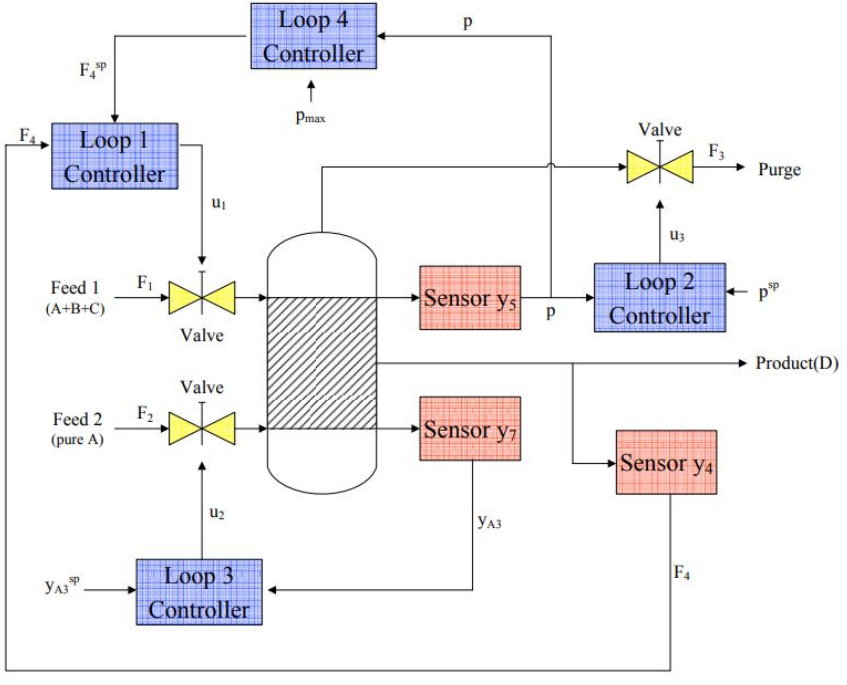
- This model can be used to represent two types of attacks
- **Integrity attacks**: in this case the attacker compromises the senor and the value it reports, therefore $a_i(k)$ can be some arbitrary non-zero value
- **DoS attack**: the controller will notice the lack of new measurements and will react accordingly. An intuitive response is the use of the last received valid signal

Experiments

- Controllers are not stupid…
- Let’s see what effect these attack might have on an industrial process and decide which attack has the most impact
- Tennessee-Eastman process control system (TE-PCS) and multi-loop PI control
- Chemical process introduced for education purposes and for benchmarking

Experiments

- Consists of an irreversible reaction which occurs in the vapour phase inside a reactor of a fixed volume V
- Non-condensable reactants A and C react in the presence of an inert B to form a non.volatile liquid D
- Different streams with flow indicated as $F_i\ (kmol\ h^{-1})$



Control Objectives

- Regulate the rate F_4 of production of product D at a set point F_4^{sp}
- Main the operating pressure of the reactor P below the shutdown limit
- Minimize C, the operating cost, that is a linear function of the purge loss of A and C relative to the production rate of D $r_D = k_0 y_{A3}^{v_1} y_{C3}^{v_2} P^{v_3}$
- Where y_{A3} and y_{C3} denote the respective fractions of A and C in the purge, the vs are given constants

Input Variables

- Four input variables (command signals) available to achieve the control objectives
- u_1, u_2 and u_3 are triggers of the actuators that can change the position of the respective valves
- u_4 Is the set point for the proportional controller for the liquid inventory
- The input variables are used by the controller as follows
- The production rate $y_4 = F_4$ is controlled using Feed 1 u_1 by loop-1 controller
- Pressure $y_5 = P$ is controlled using the purge rate u_3 by loop-2 control

Input Variables

- Partial pressure of product A in the purge $y_7 = y_{A3}$ is controlled using Feed 2 u_3 by loop-3 controller
- When u_3 saturates, the loop-4 controller uses u_1 to control the pressure P
- In steady state, the production rate F_4 is $100\ kmol\ h^{-1}$, the pressure is P is $2700\ KPa$ and the fraction of A in the purge is $47\ mol\%$
- We assume an attacker wanting to obtain an unsafe pressure in the chemical reactor, i.e., > 3000 kPa and has access to a single sensor at the time

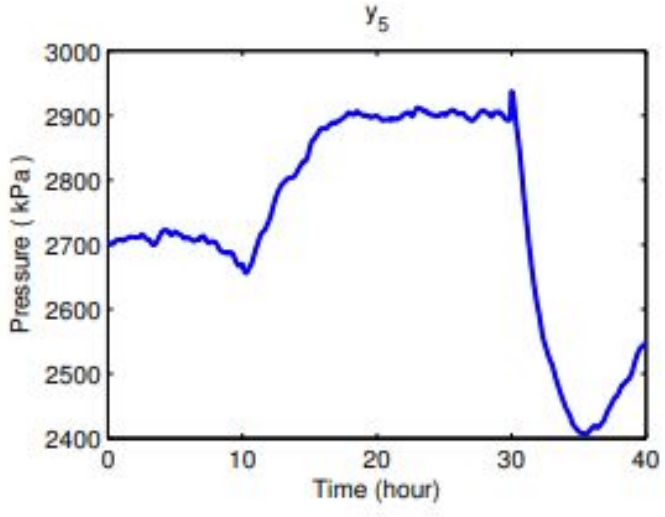
Attacking TE-PCS

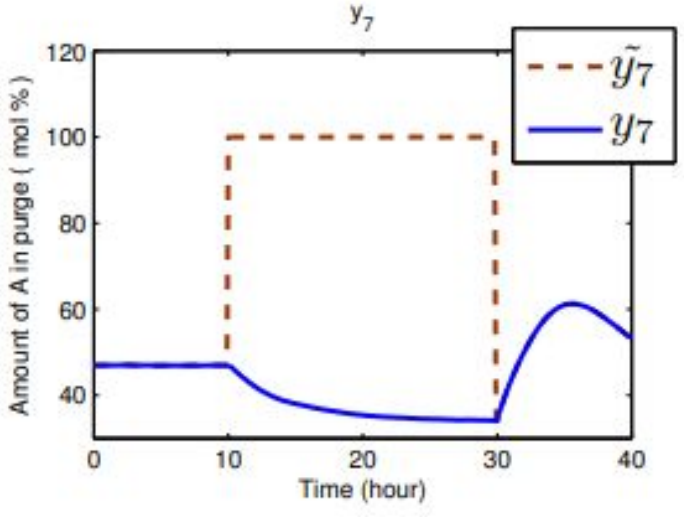
- We assume an attacker wanting to obtain an unsafe pressure in the chemical reactor, i.e., > $3000\ \mathrm{kPa}$ and has access to a single sensor at the time
- We also assume that $L_i > L_j$ when an attack i can drive the system to an unsafe state and an attack j cannot
- Spoiler: the most effective attacks are max/min attacks, i.e., when $a_i(k) = y_i^{min}$ or $a_i(k) = y_j^{max}$, although not all of them can drive the system to un unsafe state

Attacking TE-PCS

- By attacking a sensor we expect the controller to respond with incorrect control signals: by pushing y_7^{max} from t=0, to 30 the controller believes there is a large amount of component A in the tank
- However, the controller can bring back the system to the steady state when the attack finishes
- Furthermore, the pressure never reaches the critical 3000 kPa value

Attacking TE-PCS

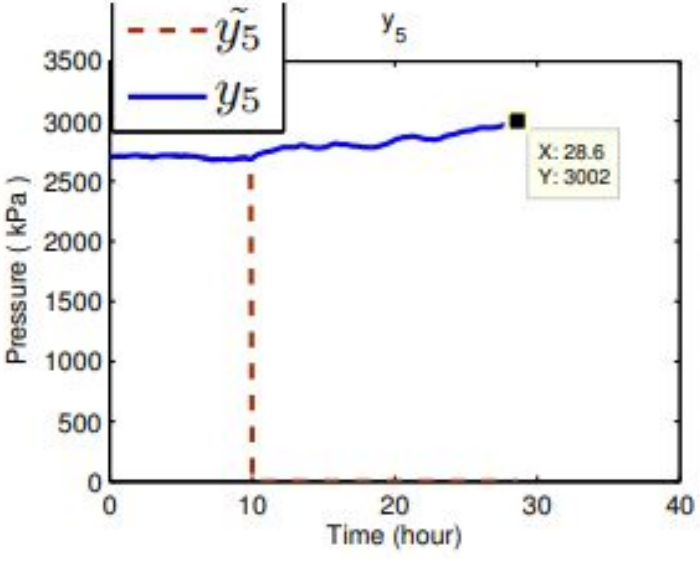


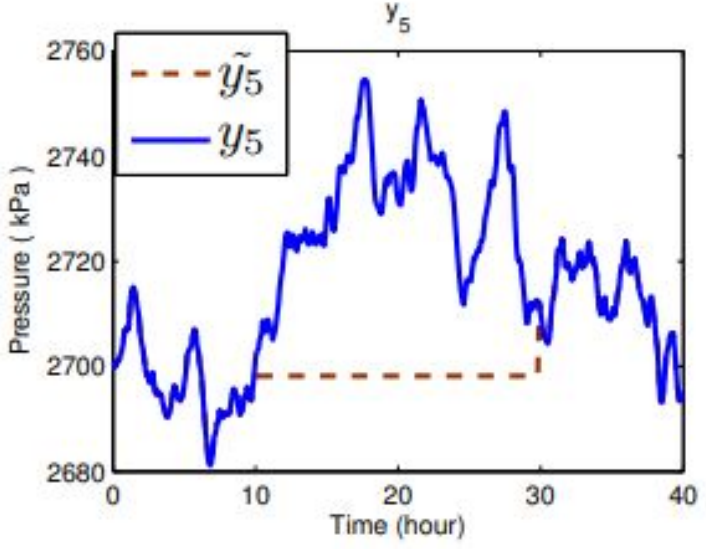


Attacking TE-PCS

- By launching attack y_5^{min} the attacker turns down the purge valve to increase the pressure and prevent the liquid products from accumulating
- The pressure of the tank keeps increasing past 3000 kPa and the system operates in an unsafe state
- However, it takes about 20 hours to cause an explosion
- This delay (slow-dynamics) gives human system operators enough time to observe the unusual phenomenon and take proper action

Attacking TE-PCS





Attack Detection

SPRITZ
SECURITY & PRIVACY
RESEARCH GROUP

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

- Detecting attacks to control systems can be formulated as an anomaly-based intrusion detection
- Instead of modelling traffic or software behavior, we model the physical system
- If we know how the output sequence of the physical system should react to the control input sequence, then any attack to the sensor data can be potentially detected by comparing the expected output with the received signal
- **Needs**: model of the system and anomaly detection algorithm

System Model

- Developing a model for the system under analysis is not trivial
- First principles model: derived from the laws of physics
- Empirical input and output data: learn the model
- To facilitate this task, most industrial control vendors provide *identification packages*, i.e., tools to develop models from physical data

System Model

- The most common models are linear systems, which model dynamics that are linear in state x and control u $x(k+1) = Ax(k) + Bu(k)$
- Assuming that the system is monitored by a sensor network of p senrs, we obtain the measurement sequence from the observation equation
- $\hat{y}(k) = Cx(k)$
- These values represents the estimated measurements collected by the sensor, and C is the output matrix

Detection Methods

- **Sequential detection theory:** methodology to detect anomalies in real-time without considering a fixed measurement time, but relying on an online-chosen detection time
- Problem formulations of this kind are called *optimal stopping problems*
- Two such problems formulation are sequential detection (sequential hypothesis testing) and quickest detection (change detection)
- Reference to this type of problems:

  T. Kailath and H. V. Poor. Detection of stochastic processes.

  IEEE Transactions on Information Theory, 44(6):2230–2258, October 1998

Optimal Stopping Problems

- Let us consider a given time series sequence $z(1), z(2), \ldots, z(N)$
- Goal: determine the minimum number N of samples to observe before making a decision on one of two hypothesis: H_0, i.e., normal behavior or H_1, i.e., attack
- Sequential detection strategy: we assume that the time series originated from either the normal or attack hypothesis and we want to decide on the hypothesis in the minimum amount of time
- Change detection: the sequence originated from the normal hypothesis and changed to the other. We want to detect this change ASAP

Sequential Detection

SPRITZ
SECURITY & PRIVACY
RESEARCH GROUP

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

- *False alarm probability*: detect an attack when there is none
- *Missed detection probability*: probability of not detecting an attack
- Goal: Given fixed false alarm and detection probability, minimize the number of observations needed to make a decision
- Solution given by the Sequential Probability Ratio Test (SPRT)
- Also referred to as Threshold Random Walk (TRW) in security papers and used to detect portscans, worms, and botnets

Sequential Detection: SPRT

- Assume that observations $z(k)$ under hypothesis H_j are generated with a probability distribution p_j
- The SPRT algorithm can be described by the following equations

$$S(k+1) = \log \frac{p_1(z(k))}{p_0(z(k))} + S(k)$$

Cum.Sum. of LLRs

$$N = \inf_n \{n : S(n) \notin [L, U]\},$$

- With S(0) = 0
- Decision rule $d_N = \begin{cases} H_1 & \text{if } S(N) \geq U \\ H_0 & \text{if } S(N) \leq L, \end{cases}$
- Where $L \approx \ln \frac{b}{1-a}$ and $U \approx \ln \frac{1-b}{a}$, a = desired P of FA and b= P of MD

Change Detection

- Detect a possible change in the generation problem at an unknown change point k_s
- CUSUM and Shiryaev-Roberts statistics are the two most commonly used algorithms for change detection
- Given a fixed false alarm rate, CUSUM aims at minimizing the time N > k_s for which the test stops and decides that a change has occurred
- Very similar to SPRT

Change Detection

- Let S(0) = 0
- CUSUM statistics is updated according to

$$S(k+1) = \left(\log \frac{p_1(z(k))}{p_0(z(k))} + S(k) \right)^+$$

- Where $(a)^+ = a$ if $a \geq 0$, 0 otherwise
- The stopping time is $N = \inf_n \{n : S(n) \geq \tau\}$
- Where the threshold tau is selected based on the false alarm constraint
- Notice: CUSUM is SPRT with L = 0, and $U = \tau$, and whenever the statistic reaches the lower bound L it restarts

Detecting Compromised Sensors

SPRITZ
SECURITY & PRIVACY
RESEARCH GROUP

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

- In general, we do not know the probability distribution for an attack
- The attacker can indeed select arbitrary and non-stationary sequence, hence assuming a fixed probability limits our detection capabilities
- We hence use ideas from nonparametric statistics
- We do not assume a parametric distribution for p_j , j = {0, 1}
- We place mild constraints on the observation sequence

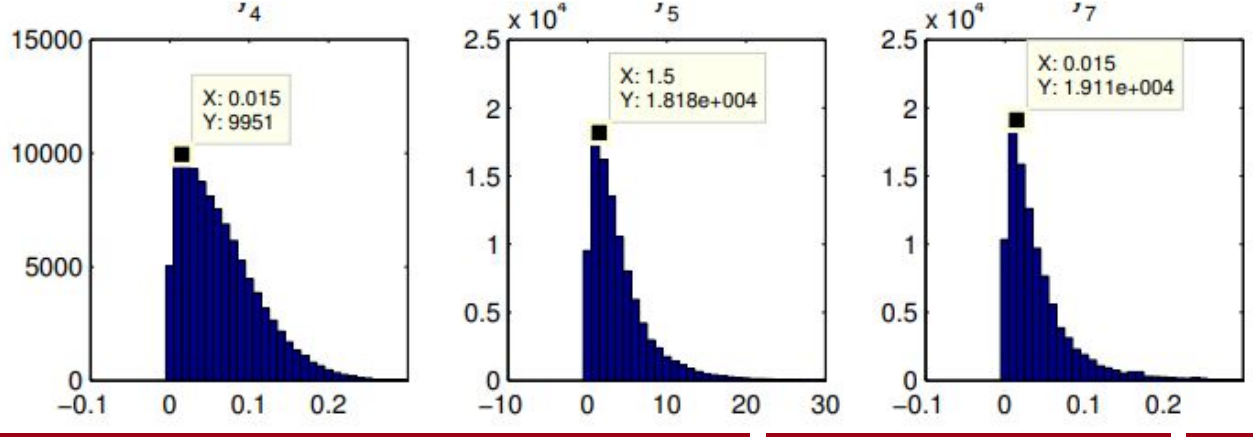
Detecting Compromised Sensors

- One of the simplest constraints is to assume the expected value of the random process $Z_i(k)$ that generates the sequence $z_i(k)$ under H_0 is less than 0 (i.e., $\mathbb{E}_0[Z_i] < 0$), and the expected value of $Z_i(k)$ under hypothesis H_1 is greater than zero (i.e., $\mathbb{E}_1[Z_i] > 0$)
- To achieve this condition, we define $z_i(k) := \|\tilde{y}_i(k) - \hat{y}_i(k)\| - b_i$ where the last terms is a small positive constant such that

$$\mathbb{E}_0[\|\tilde{y}_i(k) - \hat{y}_i(k)\| - b_i] < 0$$

Selecting b

- To select the value b for each sensor, we need to estimate the expected value of the distance $|\hat{y}_i(k) - y_i(k)|$ between the linear model estimate and the sensor measurement (i.e., without attack), respectively
- Run experiments to estimate probabilities



Detecting Compromised Sensors

- The nonparametric CUSUM statistic for sensor i is then

$$S_i(k) = (S_i(k-1) + z_i(k))^+, S_i(0) = 0$$

- The corresponding detection rule

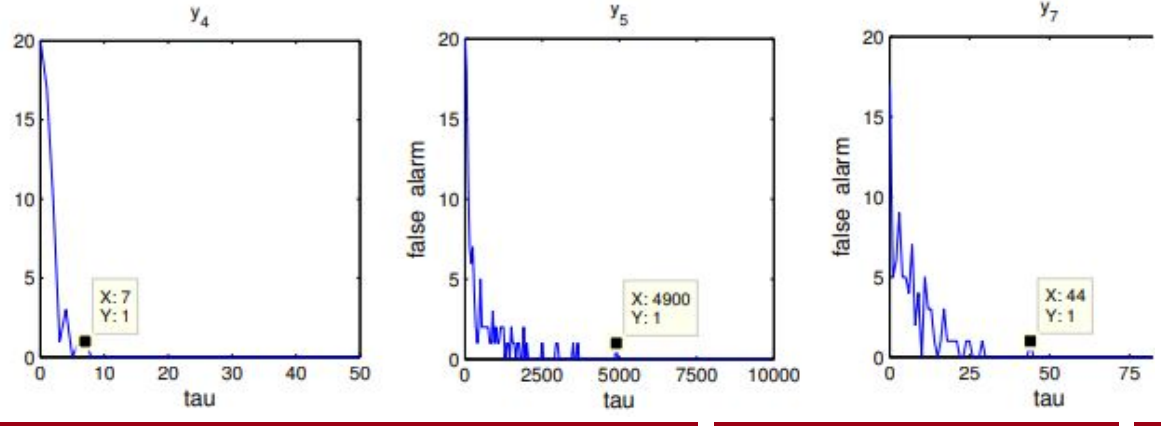
$$d_{N,i} \equiv d_\tau(S_i(k)) = \begin{cases} H_1 & \text{if } S_i(k) > \tau_i \\ H_0 & \text{otherwise.} \end{cases}$$

where tau is the threshold selected based on the false alarm rate for sensor i

- P. of FA exponentially decreases with increasing tau
- Time to detect attack inversely proportional to b

Computing tau

- Run the system many times without attacks and count how many times the system detects ana attack
- In general we would like to select tau as high as possible to avoid false alarms
- This however increases the detection time



Stealthy Attacks

SPRITZ
SECURITY & PRIVACY
RESEARCH GROUP

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

- Sometimes attackers are able to adapt to the system and develop solutions to evade detection schemes
- We now consider an attacker aware of the existence and working principle of our detection system
- We assume the attacker knows: the exact linear model that we use (i.e., matrices A, B, and C), the parameters tau and b, and the control command signals
- **Goal:** raise the pressure in the tank without being detected (i.e., keeping the controlled statistics below threshold tau)

Surge Attacks

- The attacker tries to maximize the damage as soon as possible
- However, when the statistics reaches the threshold it stays at the threshold level for the remaining time of the attack
- Means $S_i(k) = \tau$
- To do this, the attacker needs to solve

$$S_i(k) + \sqrt{(\hat{y}_i(k) - \tilde{y}_i(k))^2} - b_i = \tau_i$$

Surge Attacks

- The resulting attack for y_5 and y_4 is

$$\tilde{y}_i(k) = \begin{cases} y_i^{min} & \text{if } S_i(k+1) \leq \tau_i \\ \hat{y}_i(k) - |\tau_i + b_i - S_i(k)| & \text{if } S_i(k+1) > \tau_i \end{cases}$$

- For y_7

$$\tilde{y}_7(k) = \begin{cases} y_7^{max} & \text{if } S_{y_7}(k) \leq \tau_7 \\ \hat{y}_7 + |\tau_7 + b_7 - S_{y_7}(k)| & \text{if } S_{y_7}(k) > \tau_7 \end{cases}$$

Bias Attacks

- In a bias attack the attacker adds a small constant c_i at each time step

$$\tilde{y}_{i,k} = \hat{y}_{i,k} - c_i \in \mathcal{Y}_i$$

- In this case, the nonparametric CUSUM statistic can be written as

$$S_i(n) = \sum_{k=0}^{n-1} |\hat{y}_i(k) - \tilde{y}_i(k)| - nb_i$$

- Assuming the attack starts at time 0, and that wants to be undetected for n steps, the attacker needs to solve

$$\sum_{k=0}^{n-1} c_i = \tau_i + nb_i \quad \rightarrow \quad c_i = \tau_i/n + b$$

Bias Attacks

- This attack creates a bias of $\tau_i/n + b_i$ for each attacked signal
- If an attacker wants to maximize the damage, the attacker needs to select the smallest n
- Since $\tilde{y}_i \in \mathcal{Y}_i$, it is reduced to an impulse attack
- If the attacker wants to attack for a long time, then n will be very large
- If n is large, the bias is smaller

Geometric Attacks

- The attacker wants to drift the value very slowly at the beginning and maximize the damage at the end
- Combine slow initial drift of the bias attack with a surge attack at the end to cause maximum damage
- Given $\alpha \in (0, 1)$, the attack is $\tilde{y}_i(k) = \hat{y}_i(k) - \beta_i \alpha_i^{n-k}$
- We need to find alpha and beta such that $S_i(n) = \tau_i$

Geometric Attacks

- Assume the attack starts at time k = 0 and that the attacker wants to be undetected for n steps
- Need to solve the following equation $\sum_{k=0}^{n-1} \beta_i \alpha_i^{n-k} - nb_i = \tau_i$
- The addition is a geometric progression

$$\sum_{k=0}^{n-1} \beta_i \alpha_i^{n-k} = \beta_i \alpha_i^n \sum_{k=0}^{n-1} (\alpha_i^{-1})^k = \beta_i \frac{1-\alpha_i^n}{\alpha_i^{-1}-1}$$

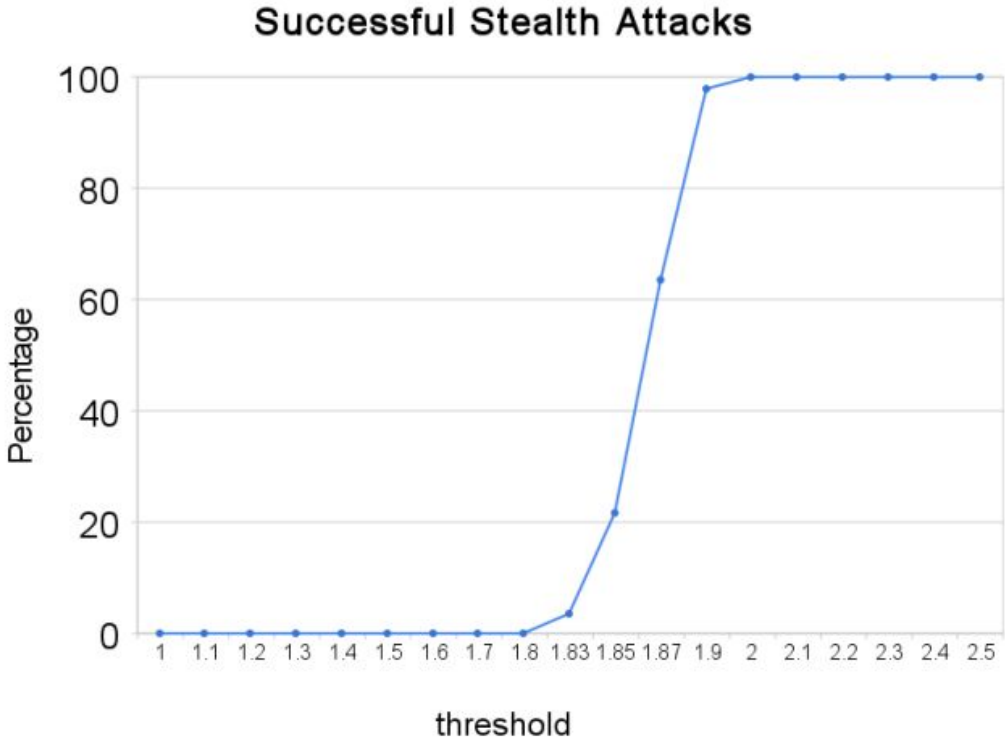
- By fixing alpha, the attacker can select proper beta to satisfy the above equation

Detectability

- The value tau impacts on how good we are in detecting attacks
- Thus, a relevant information is the percentage of successful stealthy attacks for different values of tau
- Paramterization of threshold as $\tau_i^{test} = p\tau_i$

Detectability

- Success of geometric stealthy attack in bringing pressure above 3000kPa
- Attack undetected



Safety

- Although attack may go undetected, they usually do not impact on the safety
- The most successful attack is the geometric one