An Integrated Cyber-Physical Risk Assessment Framework for Worst-Case Attacks in Industrial Control Systems*

Preprint, compiled July 21, 2023

Navid Aftabi[1†], **Dan Li, Ph.D.**[1‡], **and Thomas Sharkey, Ph.D.**[1§]

[1]*Department of Industrial Engineering, Clemson University, Clemson, SC 29634, USA*

Abstract

Industrial Control Systems (ICSs) are widely used in critical infrastructures that face various cyberattacks causing physical damage. With the increasing integration of the ICSs and information technology (IT), ensuring the security of ICSs is of paramount importance. In an ICS, cyberattacks exploit vulnerabilities to compromise sensors and controllers, aiming to cause physical damage. Maliciously accessing different components poses varying risks, highlighting the importance of identifying high-risk cyberattacks. This aids in designing effective detection schemes and mitigation strategies. This paper proposes an optimization-based cyber-risk assessment framework that integrates cyber and physical systems of ICSs. The framework models cyberattacks with varying expertise and knowledge by 1) maximizing physical impact in terms of failure time of the physical system, 2) quickly accessing the sensors and controllers in the cyber system while exploiting limited vulnerabilities, 3) avoiding detection in the physical system, and 4) complying with the cyber and physical restrictions. These objectives enable us to jointly model the interactions between the cyber and physical systems and study the critical cyberattacks that cause the highest impact on the physical system under certain resource constraints. Our framework serves as a tool to understand the vulnerabilities of an ICS with a holistic consideration of cyber and physical systems and their interactions and assess the risk of existing detection schemes by generating the worst-case attack strategies. We illustrate and verify the effectiveness of our proposed method in a numerical and a case study. The results show that a worst-case strategic attacker causes almost 19% further acceleration in the failure time of the physical system while remaining undetected compared to a random attacker.

***K**eywords* Cyber-Physical Systems · Cybersecurity · Industrial Control Systems · Worst-Case Attack Generation · Optimization · Reliability Engineering · Risk Assessment

1 Introduction

In the past few decades, the cybersecurity of ICSs has been drawing increasing attention and concerns. ICSs are widely used for large-scale automation and online management of critical infrastructures, including power systems, water supply systems, transportation systems, and oil and gas pipelines [1, 2]. ICSs were initially designed to work under local area networks without exposure to the internet. Nowadays, they are becoming more intelligent with the increasing penetration of Internet of Things devices, edge and cloud computing, and 5G networks. At the same time, they are becoming more exposed and vulnerable to cyberattacks [1]. Any successful cyberattack has resulted in physical disruptions to the ICSs [2]. Stuxnet in 2009 and the Ukrainian blackout in 2015 are examples of publicized cyberattacks on the ICSs [3].

One of the crucial procedures to overcome these concerns is identifying potential cyberattacks and evaluating the potential loss they cause, which is known as ICS cyber-risk assessment. Generally speaking, risk is defined as the possibility of an event that will adversely affect the achievement of the organization's objectives [1]. In information security, risk is defined as the possible damage associated with the realization of a specific threat exploiting the weaknesses of an asset or a group of assets (often referred to as vulnerabilities) [1]. From the latter risk definition viewpoint, the identified cyberattacks and the loss evaluation facilitate the design of cyberattack detection schemes and decision-making around risk mitigation. However, the lack of sufficient attack data in ICSs challenges the development of effective cyberattack detection schemes. Therefore, understanding the attacker's behavior is useful in designing robust detection schemes and securing the ICSs. Current attack generation methods do not consider the structure of the ICSs holistically and only focus on the physical systems [4, 5, 6], which motivates one of the underlying research question in this paper. Therefore, risk assessment for the ICSs involves identifying potential cyberattacks interacting with an ICS's cyber and physical systems and evaluating their physical outcomes.

The ICSs are complex heterogeneous systems incorporating networking, physical system, and computations that require a multi-paradigm framework to assess their cybersecurity risk [7]. Therefore, risk assessment for the ICSs requires models that capture the physics and network structure of the systems. A generic cyber-physical structure of an ICS incorporates both the cyber network and the physical system. A cyberattack usually intrudes into the system from initial vulnerable cyber components on a computer network [1, 8]. It exploits a sequence of vulnerabilities that requires limited effort to gain access to some sensors and/or controllers that allows imposing adversarial impact on the physical system if it successful. This behavior of the cyberattacks is similar to finding a *shortest path* through the cyber system of an ICS while satisfying preconditions on the current vulnerability exploitation and, then, manipulating the physical system to impose catastrophic damages.

Many of the existing cyber-risk assessment methods for ICSs focus on cyber risk assessment [9, 10] and do not account for

*This work was funded by National Science Foundation Award 2119654.

†`naftabi@clemson.edu`
‡`dli4@clemson.edu`
§`tcshark@clemson.edu`

the potential physical outcomes of the attack. In contrast, others only focus on physical impacts without considering the connectivity and vulnerabilities in the cyber network [11, 12, 13, 5]. In general, there lacks a comprehensive methodology investigating how cyberattacks can strategically compromise the cyber network to maximize the impact on the physical system [1, 2, 7], which motivates our second underlying research question in this paper.

This paper presents a novel optimization framework for identifying the most physically impactful attack intending to intrude on the system by exploiting a sequence of vulnerabilities in the cyber system with limited effort to reach the sensors and/or controllers that interact with the physical system (see Figure 1). Gaining control of sensors and/or controllers allows the attacker to manipulate the physical system and bring it to an undesirable state while ensuring the existing cyberattack detection mechanism does not detect the manipulating actions. We consider the cyberattacks that affect the availability and integrity of operational data and control in the ICS, which can cause physical damage to the ICSs since these attacks can tamper with both cyber and physical systems. Since the successful execution of a cyberattack is contingent upon both the system's vulnerabilities and the attackers' ability to utilize them, it is useful to understand them quantitatively for the worst-case risk assessment. Therefore, we assume the attacker can determine where to invest its resources while executing a cyberattack. Accordingly, this assumption allows us to focus on the worst-case cyberattack that differs from probabilistic risk assessment approaches, distinguishing our method from other cyber risk assessment methods in the literature. Similar to our work, Biehler et al. [5] propose an optimization framework to understand the mechanics of insider attacks (e.g., stealthy attacks) that need to have complete information about the system. Due to the type of attack, they only focus on the physical system. In contrast, we assume a general attacker who intends to impose damage to the physical system by passing through a computer network and may not have complete knowledge about the system.

To verify the effectiveness, flexibility, and applicability of our proposed framework, we generate cyberattacks on instances of synthetic ICSs containing both cyber and physical components and also consider a case study on a boiling water power plant ICS. This analysis allows us to investigate how a sequence of strategic attack actions can cause significant physical impact while remain undetected during the entire attack horizon.

The contributions of this paper include the following:

1. We propose a novel framework for generating the most critical cyberattack in an ICS that will have the most impact on physical outcomes with limited effort in the cyber system by integrating the cyber and physical constraints.

2. We leverage the physical system dynamics of the ICS and its degradation to quantify the cyber-to-physical impact in terms of the failure time of the ICS. More specifically, we propose a physics-based objective function for the attacker based on the mean failure time (MFT) of the ICS intended to be minimized for imposing physical impact.

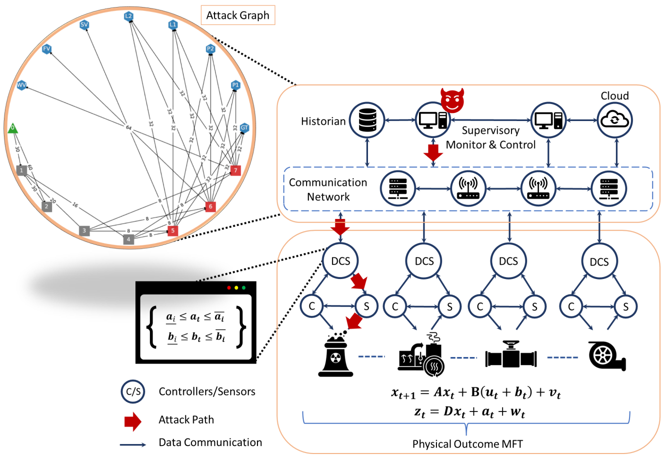

Attack Graph
Historian
Supervisory Monitor & Control
Cloud
Communication Network
DCS
C
S
$\underline{a_i} \le a_t \le \overline{a_i}$
$\underline{b_i} \le b_t \le \overline{b_i}$
C/S
Controllers/Sensors
Attack Path
Data Communication
$x_{t+1} = Ax_t + B(u_t + b_t) + v_t$
$z_t = Dx_t + a_t + w_t$
Physical Outcome MFT

Figure 1: Proposed cyber-risk assessment framework

The remainder of this paper is organized as follows. The literature is reviewed in Section 2, and the research gaps and our contribution are thoroughly discussed. We build the required background for our framework in Section 3. The proposed framework is introduced in Section 4. The experimental results of a numerical study and a real-world case are reported in Section 5. Finally, the paper concludes by discussing the results in Section 6.

2 Related Work

In this section, we briefly review the related literature on the cyber-risk assessment of ICSs in both the cyber and physical systems and discuss the relevant work supporting formalizing the cyber-aware physical impact of a cyberattack on the ICSs.

2.1 *Traditional Cyber-Risk Assessment*

From an optimization modeling perspective, several methods have been used to formalize the industrial cyber-physical systems as a framework and use them for threat identification, vulnerability analysis, cyber-risk assessment, and risk mitigation [14, 15]. Stochastic programming [16, 17, 18, 19], multi-objective optimization [20, 21], network interdiction and multilevel optimization [19, 22], and nonlinear programming [23] are some of the methods have been applied in the literature. The main focus of these methods is to increase the the required effort needed for cyberattacks to be complete. In this line of research, the focus is on understanding the types of attacks that can potentially be conducted within the cyberinfrastructure of a physical system. However, they do not model the impact of how a successful attack will then change the operations of the physical system.

2.2 *Physics-based Risk Assessment*

From a physical perspective, many physics-based methods have been proposed for detecting various kinds of attacks from the anomalies in the system, based on domain knowledge of the physical system of the ICSs [24, 25, 26, 27]. Traditional state estimation detection schemes apply statistical process monitoring methods to the state estimation residuals. The state estimator can also be designed based on the original differential equation models of the physical system, where statistical monitoring methods are used for detection. Most physics-based methods focus on a physical system rather than the entire ICS and do not consider the strategic cyber intrusion process. Furthermore, they usually focus on a specific type of attack and aim to minimize the detection delay.

2.3 Cross-Layer Risk Assessment

A noticeable limitation of physics-based and traditional research on the cyber-risk assessment of ICSs is that they tend to model the cyber network and physical system independently. However, the interaction between them plays an essential role in the operations of ICSs, especially when facing cyberattacks that aim to cause physical damage. Recently, efforts have been made for cyber-risk assessment and establishing risk mitigation strategies while considering both cyber and physical systems. The multi-paradigm frameworks [11, 13], Markov decision process [28], and game theory [29] were used to capture the structural characteristics of both of the systems. These studies mainly focus on the disruptions of physical systems instead of the entire ICS and do not consider the connectivity and vulnerabilities in the cyber system and the strategic cyber intrusion process.

2.4 Degradation and Remaining Useful Life

An essential step in the cyber-risk assessment of the ICSs is to evaluate the physical outcome of the malicious actions of a cyberattack on both the cyber and physical systems. The most traditional risk assessment approaches evaluate the potential loss of a cyberattack in the monetary value. Even though this approach is sensible, it requires assumptions on the values of assets of ICSs. An appropriate alternative for evaluating the physical outcome is to assess how cyberattacks accelerate system failure. Studying the degradation process of the physical system provides a tool to introduce a new security metric to quantify the risk associated with cyberattacks on the ICSs. The literature on the system failure time, degradation, and the remaining useful life of the physical systems is extensive. For instance, [30, 31, 32] modeled the degradation signal of a physical process as a drifted Brownian motion (BM) that provide insights about the physical system lifetime.

3 System Modeling

As discussed earlier, a general ICS consists of two interconnected networks, cyber and physical systems. This section describes the models and methods used to characterize each of the network and the cyber-to-physical outcomes.

3.1 Cyber System Model

Among the attack modeling techniques, in the cyber system of an ICS, an attack graph or tree is one of the popular methods that mathematically and visually represents the sequence of events that lead to a successful cyberattack [8]. On this graph, nodes represent the attack states or security pre-/post-conditions, and arcs correspond to the transition of states fulfilled by an attack action. An attack state becomes active if its required attack state(s) are activated before. Consequently, a successful attack is a *feasible path* from initial vulnerable states to a target state on the attack graph while satisfying the *precondition dependencies*. A *critical path* on this graph is a feasible path that requires least effort. The effort or difficulty of a vulnerability exploitation is usually represented by assigning a quantity on each arc. Depending on the range of values for this quantity, the cyber system can be represented as a probabilistic attack graph [20] or exploitation-time attack graph [19].

Mathematically speaking, we define the attack graph as a directed network $\mathbf{G} = (\mathbf{N}, \mathbf{A})$ where $\mathbf{N}$ and $\mathbf{A}$ are sets of it nodes and arcs representing attack states and exploitations respectively. To reach an attack state $j \in \mathbf{N}$, the dependent preconditions $i \in \mathbf{N}$ must be satisfied such that $(i, j) \in \mathbf{A}$. Without loss of generality, we consider target nodes to be attack states where only a particular sensor or controller is being compromised. We define $\mathbf{N}_S \subset \mathbf{N}$ and $\mathbf{N}_C \subset \mathbf{N}$ as the set of states where sensors and controllers are compromised, respectively. Therefore, $|\mathbf{N}_S|$ and $|\mathbf{N}_C|$ are equal to the number of sensors and controllers in the physical system respectively.

3.2 Physical System Model

The physical system being controlled is modeled as a multivariate discrete-time linear time-invariant system of equations. This formulation of the physical process allows us to: (*i*) quantify the cyber-to-physical impact, and (*ii*) it is easily generalizable to formulate many physical systems. We consider a multi-component physical system whose dynamics are modeled as

$$\mathbf{x}_{t+1} = A\mathbf{x}_t + B\mathbf{u}_t + \mathbf{v}_t \tag{1a}$$

$$\mathbf{z}_t = D\mathbf{x}_t + \mathbf{w}_t \tag{1b}$$

In above equations, $\mathbf{x}_t \in \mathbb{R}^{|M|}$ represents the system states, and M denotes the set of these states; $\mathbf{u}_t \in \mathbb{R}^{|\mathbf{N}_C|}$ and $\mathbf{z}_t \in \mathbb{R}^{|\mathbf{N}_S|}$ are the control actions and sensor measurements respectively; $\mathbf{v}_t \in \mathbb{R}^{|M|}$ and $\mathbf{w}_t \in \mathbb{R}^{|\mathbf{N}_S|}$ are the process and measurement noises respectively; and, A, B, and D are matrices of appropriate dimensions.

3.3 Degradation Model

The physical system state variable $\mathbf{x}_t$ is used to model physical component degradation. The degradation signal $\mathbf{S}(t)$ is typically modeled as a drifted BM, $\mathbf{S}(t) = \mu t + \sigma_s \mathbf{B}(t)$ where $\mathbf{B}(t)$ is an standard BM such that $\mathbf{B}(t) \sim \mathcal{N}(0, t)$ that implies $\mathbf{S}(t) \sim \mathcal{N}(\mu t, \sigma_s^2 t)$. The degradation rate of the physical system is a function of the system state $\mathbf{x}_t$, representing the working condition of the system in practice. Given a sequence $\{\mathbf{x}_t\}$ and observing the value of the logged signal in the discrete time intervals, the discretized model of the degradation signal $\mathbf{S}_t$ of the physical system is defined as (a drifted Gaussian random walk)

$$\mathbf{S}_t = \mathbf{S}_{t-1} + \theta(\mathbf{x}_t) + e_t \tag{2}$$

where $e_t \sim \mathcal{N}(0, \sigma_s^2)$ are independent and identically distributed Gaussian white noise for each t, and, the degradation rate $\theta(\mathbf{x}_t)$ is a function of $\mathbf{x}_t$. To determine the distribution of $\mathbf{S}_t$, we define the random variable $\mathbf{J}_t$ to be the jump in the degradation signal from time $t-1$ to t for $t \geq 1$, i.e., $\mathbf{J}_t = \mathbf{S}_t - \mathbf{S}_{t-1}$. Since $\{e_t\}$ is a sequence of independent random variables, $\{\mathbf{J}_t\}$ is also a sequence of independent random variables, such that $\mathbf{J}_t \sim \mathcal{N}(\theta(\mathbf{x}_t), \sigma_s^2)$. Furthermore, the degradation signal can be written as the sum of independent jumps up to time t, i.e., $\mathbf{S}_t = \sum_{\tau=1}^{t} \mathbf{J}_\tau$, that implies

$$\mathbf{S}_t \sim \mathcal{N}\left(\sum\nolimits_{\tau=1}^{t} \theta(\mathbf{x}_\tau), \sigma_s^2 t\right) \tag{3}$$

4 Risk Assessment Methodology

In this section, we first discuss the structural representation of the adversarial model built upon the system models we discussed in Section 3 and the assumptions that affect the security of an ICS. Then, using the degradation model and its distribution (Section 3.3), we quantify the cyber-to-physical impact of the cyberattack. Afterward, we present the risk assessment formulation.

4.1 Adversarial Model

To model the worst-case attacker for a defender in an ICS, we assume a rational attacker defined as follows:

Definition 4.1 (Rational Attacker)**.** *A rational attacker intrudes into the physical system through the cyber system with limited effort and imposes physical damage to the physical system by injecting malicious data through compromised sensors and controllers using their knowledge and expertise.*

This definition does not mean that we are assuming either a specific behavior of an attacker or their amount of knowledge about the system configuration. However, it helps us to study the worst-case scenario, which really matters in designing a detection scheme.

Thinking like a rational attacker, they intend to intrude the cyber system with limited effort. To implement this intention, similar to [19], we model the cyber system as an exploitation-time attack graph (Section 3.1) where each arc $(i, j) \in \mathbf{A}$ corresponds to a vulnerability associated with a required time $t_{ij} \in \mathbb{R}_+$ to be exploited. A rational attacker on the cyber system wants to complete an attack as soon as possible, which corresponds to finding a *critical path* to the target states on $\mathbf{G}$ by choosing a proper set of vulnerabilities that requires limited exploitation time and satisfies preconditions dependencies. Therefore, as preconditions dependencies, we assume an exploit can start if at least one of its predecessor exploits has been completed, i.e., OR-type dependency. On this path, if an exploitation is successful, it can be used to reach multiple targets. However, due to the attacker's limited resources, cyberattacks on all sensors and controllers cannot be executed. To capture the attacker's resource limitation, we consider (*i*) $T \in \mathbb{Z}_+$ as the attacker's overall available time, and (*ii*) $K \in \{1, \ldots, |\mathbf{N}_S| + |\mathbf{N}_C|\}$ as the number of targets states the attacker intends to compromise. We are considering a time constraint because many attackers have limited time to access different sensors and controllers, and by expanding the time constraint, we also consider a long-term attacker. The constraint on T and K allows us to consider different types of attackers with different levels of expertise to access the sensors and controllers.

Afterwards, once a set of sensors and/or controllers are compromised, attacker starts manipulating sensor measurements and/or control action in the physical system. To incorporate this intention into our formulation, we utilize the state-space model (1) under attack that is defined as

$$\mathbf{x}_{t+1} = A\mathbf{x}_t + B[\mathbf{u}_t + \mathbf{b}_t] + \mathbf{v}_t \tag{4a}$$
$$\mathbf{z}_t = D\mathbf{x}_t + \mathbf{a}_t + \mathbf{w}_t \tag{4b}$$

where $\mathbf{a}_t \in \mathbb{R}^{|\mathbf{N}_S|}$ and $\mathbf{b}_t \in \mathbb{R}^{|\mathbf{N}_C|}$ are the attacker's manipulative actions through the compromised sensors or controllers. Since the attacker cannot perceive the exact behavior of the physical system, they alter the expected evolution of the physical system. Therefore, in building the framework, we ignore the Gaussian noises in (4). The matrices A, B, and D determine the attacker's knowledge about the physical system's state evolution, actuation, and measurement, respectively. Based on this knowledge, the attacker may have different attack strategies to impose various levels of physical impact.

Moreover, the state $\mathbf{x}_t$ is not directly observable to the attacker. Instead, the sensor measurements $\mathbf{z}_t$ are partially observable to the attacker and determine how the physical system evolves. The attacker's restriction on realizing the sensor measurements is assumed to be

$$\|\mathbf{z}_t\|_\infty \leq \delta. \tag{5}$$

Also, given the observed measurements $\mathbf{z}_t$, the attacker estimates the current state of the physical system $\mathbf{x}_t$ and the control actions $\mathbf{u}_t$ to decide on the attack actions. To quantify this behavior of the attacker, notice the relationship between measurements and states is linear in (1) without any attack actions. Therefore, using linear regression on sensor data, the estimation of the physical system states is determined as $\hat{\mathbf{x}}_t = \left(D^T D\right)^{-1} D^T \mathbf{z}_t$. Then, the control actions $\mathbf{u}_t$ are estimated based on $\hat{\mathbf{x}}_t$ and $\mathbf{z}_t$. We assume the attacker estimates the control actions as a direct perception of the measurements, i.e.,

$$\mathbf{u}_t = E\mathbf{z}_t \tag{6}$$

For the attacker, E should be determined such that $\mathbf{u}_t$ preserves the system state close to some constant target value $\tilde{\mathbf{x}}$. Therefore, $\mathbf{u}_t$ is determined by minimizing $\mathbb{E}(\hat{\mathbf{x}}_t - \tilde{\mathbf{x}})$ that implies $B\mathbf{u}_t = \tilde{\mathbf{x}} - A\hat{\mathbf{x}}_t$. Thus, assuming B and D have full column rank, we obtain

$$E = -B^\dagger A D^\dagger \tag{7}$$

where $X^\dagger$ is the pseudo-inverse of X. In addition to the attacker's physical limitations stated above, their aim is to alter the physical process in such a way that the attack actions $\mathbf{a}_t$ and $\mathbf{b}_t$ bypass the existing intrusion detection mechanism. To quantify the attacker's objective of avoiding from being detected, we consider a detection threshold on the attack actions $\mathbf{a}_t$ and $\mathbf{b}_t$, and we consider the detection algorithm as

$$\underline{\mathbf{a}}_i \leq \mathbf{a}_{it} \leq \overline{\mathbf{a}}_i, \quad i \in \mathbf{N}_S \tag{8a}$$
$$\underline{\mathbf{b}}_i \leq \mathbf{b}_{it} \leq \overline{\mathbf{b}}_i, \quad i \in \mathbf{N}_C \tag{8b}$$

The most important goal of an attacker is to impose physical impact to the physical system of an ICS by considering the cyber and physical systems and their limits. In the next sub-section, we elaborate on this objective.

4.2 Cyber-to-Physical Impact Quantification

On a physical system, the attack actions manipulate the sensor and controller data that essentially alter $\mathbf{x}_t$, which accelerates the unexpected system failure. The cyber-to-physical outcome can be quantified by the unexpected failures brought on by the cyberattack. Let the random variable $\mathbf{T}_f$ denote the system failure time. Since the attacker perceives a realization of the system in discrete time intervals, considering the degradation model (2), the failure time is defined as

$$\mathbf{T}_f = \inf\{t > 0 : \mathbf{S}_t \geq \lambda\} \tag{9}$$

where λ is the known failure threshold. The attacker's aim is to maximize the physical impact by minimizing the mean failure time (MFT), i.e., MFT $= \mathbb{E}(\mathbf{T}_f)$. To characterize MFT from the attacker's viewpoint, notice that

$$\mathbb{E}(\mathbf{T}_f) = \sum_{\tau=1}^{\infty} \mathbb{P}(\mathbf{T}_f \geq \tau) \tag{10}$$

where $\mathbb{P}(\mathbf{T}_f \geq \tau) = \mathbb{P}(\mathbf{T}_f \geq \tau, \mathbf{S}_\tau \geq \lambda) + \mathbb{P}(\mathbf{T}_f \geq \tau, \mathbf{S}_\tau < \lambda)$. From the attacker's perspective, if the system is in a failure state, attacking the system is not beneficial that implies $\{\mathbf{T}_f \geq \tau, \mathbf{S}_\tau \geq \lambda\} = \emptyset$. Considering the attacker's restriction on the available time for executing a cyberattack on an ICS, T, the MFT, for the attacker, is defined as $\mathbb{E}(\mathbf{T}_f) = \sum_{\tau=1}^{T} \mathbb{P}(\mathbf{S}_\tau < \lambda)$ since $\{\mathbf{T}_f \geq \tau, \mathbf{S}_\tau < \lambda\} = \{\mathbf{S}_\tau < \lambda\}$. Now, by degradation signal distribution (3), the MFT for the attacker can be re-written as

$$\mathbb{E}(\mathbf{T}_f) = \sum_{\tau=1}^{T} \Phi(z_\tau) \tag{11}$$

where $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution, and

$$z_\tau = \frac{\lambda - \sum_{t=1}^{\tau} \theta(\mathbf{x}_t)}{\sigma_s \sqrt{\tau}} \tag{12}$$

Since the attacker's aim is to minimize MFT, by monotonicity of the c.d.f., minimizing $\mathbb{E}(\mathbf{T}_f)$ in (11) is equivalent to minimizing $\sum_{\tau=1}^{T} z_\tau$. We assume the degradation rate for the attacker is a direct perception of $\mathbf{x}_t$ [31], i.e., $\theta(\mathbf{x}_t) = \kappa + \gamma^T \mathbf{x}_t$ where κ is constant drift, and γ represents correlation between system states. Hence, (12) can be re-written as

$$z_\tau = \frac{\lambda - \kappa\tau - \gamma^T \sum_{t=1}^{\tau} \mathbf{x}_t}{\sigma_s \sqrt{\tau}} \tag{13}$$

4.3 Risk Assessment Model - Attack Generation

Our proposed framework is formulated in (14), which consists of two interconnected networks and their constraints while optimizing attacker's objective. In our formulation, we define the decision variables:

- $h_i \in \mathbb{R}_+$ as the start time of attack state $i \in \mathbf{N}$ on $\mathbf{G}$,
- $y_{ij} \in \{0, 1\}$ to take 1 if vulnerability $(i, j) \in \mathbf{A}$ on $\mathbf{G}$ is exploited to execute the cyberattack and 0 otherwise,
- $f_{ij} \in \mathbb{Z}_+$ as the number of target states attacked through the exploited vulnerability $(i, j) \in \mathbf{A}$,
- $\alpha_{it} \in \{0, 1\}$ and $\beta_{jt} \in \{0, 1\}$ to take 1 if the target states $i \in \mathbf{N}_S$ and/or $j \in \mathbf{N}_C$ be compromised by time $t \in T$,
- $\mathbf{x}_t \in \mathbb{R}^{|M|}$, $\mathbf{u}_t \in \mathbb{R}^{|N_C|}$, and $\mathbf{z}_t \in \mathbb{R}^{|N_S|}$ to be the physical system states, control actions, and sensor measurement at time $t \in T$,
- $\mathbf{a}_t \in \mathbb{R}^{|\mathbf{N}_S|}$ and $\mathbf{b}_t \in \mathbb{R}^{|\mathbf{N}_C|}$ as the attacker's action on the physical system through the compromised sensors and controllers at time $t \in T$ respectively.

$$\begin{aligned}
&\min \sum_{(i,j)\in\mathbf{A}} y_{ij} + \sum_{\tau=1}^{T} z_\tau && && \text{(14a)}\\
&\text{s.t. relation (13)} && && \text{(14b)}\\
&h_j \geq h_i + t_{ij} - T(1 - y_{ij}) && (i, j) \in \mathbf{A} && \text{(14c)}\\
&f_{ij} \leq K y_{ij} && (i, j) \in \mathbf{A}' && \text{(14d)}\\
&\sum_{j:(i,j)\in\mathbf{A}'} f_{ij} - \sum_{j:(j,i)\in\mathbf{A}'} f_{ji} = \begin{cases} K & i = 0 \\ -K & i = \nu \\ 0 & i \in \mathbf{N} \setminus \{0\} \end{cases} && i \in \mathbf{N} \cup \{\nu\} && \text{(14e)}\\
&\alpha_{it} \leq \sum_{j:(j,i)\in\mathbf{A}} f_{ji} && i \in \mathbf{N}_S, t \in T && \text{(14f)}\\
&h_i - 1 \leq \sum_{t\in T} (1 - \alpha_{it}) && i \in \mathbf{N}_S && \text{(14g)}\\
&\alpha_{i,t-1} \leq \alpha_{it} && i \in \mathbf{N}_S, t \in T_1 && \text{(14h)}\\
&\beta_{it} \leq \sum_{j:(j,i)\in\mathbf{A}} f_{ji} && i \in \mathbf{N}_C, t \in T && \text{(14i)}\\
&h_i - 1 \leq \sum_{t\in T} (1 - \beta_{it}) && i \in \mathbf{N}_C && \text{(14j)}\\
&\beta_{i,t-1} \leq \beta_{it} && i \in \mathbf{N}_C, t \in T_1 && \text{(14k)}\\
&\mathbf{x}_{t+1} = A\mathbf{x}_t + B\left[\mathbf{u}_t + \mathbf{b}_t\right] && t \in T && \text{(14l)}\\
&\mathbf{z}_t = D\mathbf{x}_t + \mathbf{a}_t && t \in T && \text{(14m)}\\
&\mathbf{u}_t = E\mathbf{z}_t && t \in T && \text{(14n)}\\
&-\delta \leq \mathbf{z}_{it} \leq \delta && i \in \mathbf{N}_S, t \in T && \text{(14o)}\\
&\underline{\mathbf{a}}_i \alpha_{it} \leq \mathbf{a}_{it} \leq \overline{\mathbf{a}}_i \alpha_{it} && i \in \mathbf{N}_S, t \in T && \text{(14p)}\\
&\underline{\mathbf{b}}_i \beta_{it} \leq \mathbf{b}_{it} \leq \overline{\mathbf{b}}_i \beta_{it} && i \in \mathbf{N}_C, t \in T && \text{(14q)}\\
&h_i \in \mathbb{R}_+ && i \in \mathbf{N} && \text{(14r)}\\
&f_{ij} \in \mathbb{Z}_+, y_{ij} \in \{0, 1\} && (i, j) \in \mathbf{A}' && \text{(14s)}\\
&\alpha_{it} \in \{0, 1\} && i \in \mathbf{N}_S, t \in T && \text{(14t)}\\
&\beta_{it} \in \{0, 1\} && i \in \mathbf{N}_C, t \in T && \text{(14u)}\\
&\mathbf{x}_t \in \mathbb{R}^{|M|}, \mathbf{a}_t, \mathbf{z}_t \in \mathbb{R}^{|\mathbf{N}_S|} && t \in T && \text{(14v)}\\
&\mathbf{b}_t, \mathbf{u}_t \in \mathbb{R}^{|\mathbf{N}_C|} && t \in T && \text{(14w)}
\end{aligned}$$

As discussed in Section 4.1, the objective function (14a) and (14b) incorporates the attacker's objective, which is to reach the target sensors and/or controllers by exploiting a limited number of vulnerabilities in the cyber system and minimize the MFT to impose physical damage corresponding to the set of compromised sensors and controllers in the cyber system. The detailed explanation of each constraint in the framework (14) according to their association with each cyber or physical system is as follows:

4.3.1 Cyber System Constraints

We assume $\mathbf{G}$ has a single initial vulnerable state with $h_0 = 0$. This assumption is not restrictive. If there exists multiple initial vulnerable states, we consider their start time as zero, and we add an auxiliary node to $\mathbf{G}$ with $h_{aux} = 0$ and add directed arcs from auxiliary node to all initial vulnerable states with zero exploitation time.

- On attack graph $\mathbf{G}$, since all states are of OR-type, in each state $i \in \mathbf{N}$, the attacker exploits the vulnerability that requires limited exploitation time to reach state j, i.e., $h_j = \min_{i:(i,j)\in\mathbf{A}} h_i + t_{ij}$. This non-linear relation can be linearized using integer variables as in (14c).

- The attacker pays the required effort to exploit a vulnerability once but may benefit from that to reach multiple targets. This means if the attacker does not exploit a vulnerability, no goal node is attacked through that exploitation. This limitation is modeled in (14d).
- At each attack state $i \in \mathbf{N}$, the attacker hack a set of devices on cyber system to reach attack state $j \in \mathbf{N}$, such that $(i, j) \in \mathbf{A}$. Since the attacker has limited resources on attacking the targets, K, and any exploitation can be used to attack different targets, they can hack the devices based their remaining resources at each attack state $i \in \mathbf{N}$. This cyber system constraint is formulated in (14e). As this constraint is a network-flow constraint, we add an auxiliary node v (super sink node) to $\mathbf{G}$ and add directed arcs from all target states to this node with zero exploitation time. In other words, we define the arcs set $R = \{(u, v) : u \in \mathbf{N}_S \cup \mathbf{N}_C\}$ with $t_{ij} = 0$, for $(i, j) \in R$, and $\mathbf{A}' = \mathbf{A} \cup R$. This node can be interpreted as the attacker's imaginary target state where the physical state is under control.

4.3.2 *Cross-layer Constraints*

The constraints sets (14f- 14k) aim at modeling the connection between the cyber and the physical systems.

- The attacker can start malicious action on the physical system after the time (s)he follows a sequence of exploits in the cyber system that leads to an attack state where a particular sensor and/or controller is being compromised. This constraint is modeled in (14f) and (14i) for sensors and controllers respectively.
- The time the attacker starts the malicious action on the physical system is after the time the target states be activated on $\mathbf{G}$. This constraint is formulated in (14g) and (14j) for sensors and controllers respectively.
- The attacker preserves her/his control on the compromised sensor and/or controller until the end of attack execution time T. This limitation is denoted in (14h) and (14k) for sensors and controllers respectively. In these constraints $T_1 = T \setminus \{0\}$.

4.3.3 *Physical System Constraints*

The constraints sets (14l- 14q) formulate the attacker's perception and restrictions on the physical system and avoiding from being detected.

- The evolution of the physical system from attacker's view point (i.e., attacker's knowledge of the physical system) is captured in constraints (14l), (14m), and (14n).
- The constraint (14o) is the linearization of the attackers restriction on realizing the sensor measurements.
- The attacker's malicious actions on the physical system need to bypass the existing cyberattack intrusion detection mechanism that is captured in (14p) and (14q) where $\underline{\mathbf{a}}, \overline{\mathbf{a}} \in \mathbb{R}^{|\mathbf{N}_S|}$ and $\underline{\mathbf{b}}, \overline{\mathbf{b}} \in \mathbb{R}^{|\mathbf{N}_C|}$ are lower and upper bounds of attacker's action on measurements and controllers, respectively. The attack actions are forced by the cross-layer decision variables and constraints to remain zero until the corresponding sensor and/or controller be compromised.

There might exist an attacker who is an expert in either the cyber or the physical system. Since these attackers are not as harmful as the attacker with knowledge and expertise on both systems, it is worth mentioning that our proposed framework is designed for attackers with various levels of expertise and knowledge on both cyber and physical systems.

Since our proposed framework is a mixed-integer linear program (MILP), current powerful general solvers, such as Gurobi, can find the optimal solution. To evaluate the effectiveness of the proposed model, the optimization model is programmed in Python using Gurobi 10.0.1 and setting its parameters, *Heuristics* to 0, *IntFeasTol* to $1e-09$, *IntegralityFocus* to 1, and *Presolve* to 0. All the experiments are conducted on a machine with CPU model 11th Gen Intel(R) Core(TM) i7-1165G7 2.80GHz and 16GB RAM, and all the experiments are solved to optimality.

5 Computational Experiments and Results

We conduct a numerical and case study to illustrate and validate the risk assessment methodology proposed in Section 4. The numerical study is intended to show the efficiency of the proposed framework. We validate our methodology on a hardware-in-the-loop (HIL) simulation test bed to demonstrate the applicability of our method in practice. For both studies, we design experiment scenarios based on the attacker's available time to execute the cyberattack T and the attacker's available resources to conduct the cyberattack K. This way, we generate the most critical attack for attackers with different knowledge and expertise. In order to have a fair comparison of the normal system versus the system under the optimal attack, in both numerical and case studies, we simulate the physical system (4) for 1000 time units given the optimal attack actions ($\mathbf{a}_t$ and $\mathbf{b}_t$) obtained from solving (14) and the same Gaussian noises in the process. We provide a pseudo-code with the respective inputs in Pseudo-Code 1 to summarize the procedure.

5.1 *Numerical Study*

5.1.1 *Experiment Setup*

We consider an ICS where a physical system with $|M| = 3$, $|\mathbf{N}_S| = 3$, and $|\mathbf{N}_C| = 3$ (2nd scenario in [31]). A communication network, as the cyber network, is connected to these sensors and controllers. Random cyber networks are generated with a single initial vulnerability point, 12 security states, three states for sensors and controllers representing each of them are being compromised, and $t_{ij} \sim U(1, 50)$. Three types of random networks are generated with these settings, (*i*) all sensors' and controllers' states are connected to the same precondition states (Figure 2a), (*ii*) sensors' and controllers' states are connected to two separate precondition states (Figure 2b), and, (*iii*) three random combinations of sensor-controller states are connected to three separate precondition states (Figure 2c). The physical system dynamics is given by, $B = D = \mathbf{I}_3$, $\sigma_{\mathbf{v}} = 0.1$, $\sigma_{\mathbf{w}} = \sqrt{0.001}$, $\mathbf{v}_t \sim \mathcal{N}(0, \sigma_{\mathbf{v}}^2 \mathbf{I}_3)$, $\mathbf{w}_t \sim \mathcal{N}(0, \sigma_{\mathbf{2}}^2 \mathbf{I}_3)$, $\kappa = 0.4713\sqrt{2}$, $\gamma = \begin{bmatrix} 0.058 & 0.058 & 0.996 \end{bmatrix}^T$,

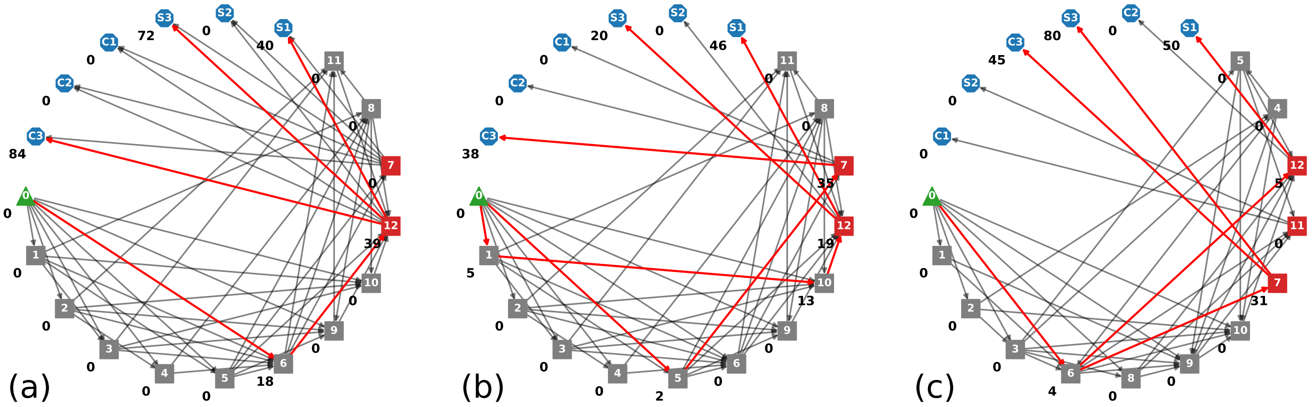


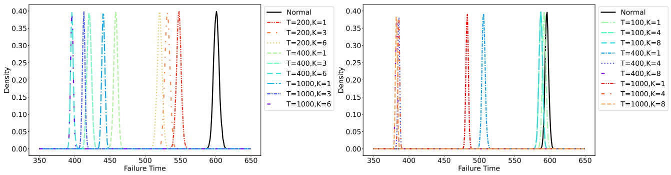
Figure 2: Attack graph topologies and optimal solutions for $K = 3$. Red edges are optimal attack actions y_{ij} and the numbers next to each node are h_i.

```
Pseudo-Code 1 Risk Assessment Procedure
  Inputs:
      • Attacker's limitations: $T, K$ and $\delta$ in (5)
      • Attack Graph $\mathbf{G} = (\mathbf{N}, \mathbf{A})$
      • Exploitation times $t_{ij}$, $\forall (i, j) \in \mathbf{A}$
      • State-space model dynamics in (1): $A, B, D, \sigma_{\mathbf{v}}, \sigma_{\mathbf{w}}$
      • Degradation model parameters in (13): $\kappa, \gamma, \sigma_s$
      • Data control system parameters in (8): $\underline{\mathbf{a}}, \overline{\mathbf{a}}, \underline{\mathbf{b}}, \overline{\mathbf{b}}$
  procedure Attack Generation(Inputs)
      1. Compute Matrix $E$ in (7)
      2. Solve model (14) for attacker's action on both cyber and physical systems using a general purpose solver such as Gurobi
      3. return Attacker's optimal actions:
             Cyber system: $y^*_{ij}, f^*_{ij}$
             Physical system: $\mathbf{a}^*_t, \mathbf{b}^*_t$
  end procedure
  procedure Risk Assessment($y^*_{ij}, f^*_{ij}, \mathbf{a}^*_t, \mathbf{b}^*_t$)
      1. Deploy the attacker's optimal actions $y^*_{ij}, f^*_{ij}, \mathbf{a}^*_t, \mathbf{b}^*_t$ into the ICS to identify the critical vulnerabilities in the cyber system and evaluate MFT of the physical system using (11) and (13)
  end procedure
```

$\sigma_s = 0.1$, E is computed by the relation (7), and

$$A = \begin{bmatrix} 0.12 & 0.30 & 0.06 \\ 0.06 & 0.48 & 0.06 \\ 0.30 & 0.12 & 0.54 \end{bmatrix}$$

Given these data, we simulate the normal system for 50 replications with 1000 time units and compute $\mathbf{S}_t$. We set $T_\lambda = 600$ to be the time the system fails under normal operation and obtain λ by taking the average of the degradation signals at this time. Also, we compute the average standard deviation of the sensor reading over the replications, $\sigma_{\mathbf{z}}$, and we set $\delta = 3\sigma_{\mathbf{z}}$, $\underline{\mathbf{a}}_i = -3\sigma_{\mathbf{w}}$, $\overline{\mathbf{a}}_i = 3\sigma_{\mathbf{w}}$, $\underline{\mathbf{b}}_i = -3\sigma_{\mathbf{v}}$, and $\overline{\mathbf{b}}_i = 3\sigma_{\mathbf{v}}$. Given these settings, we design scenarios based on the attacker's available (*i*) resources for exploiting vulnerability (K), and (*ii*) time on executing attacks (T).

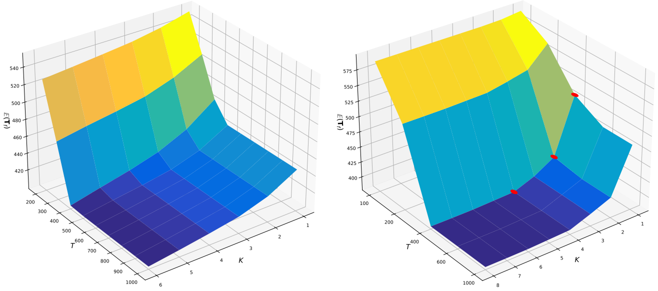


(a) Numerical study (b) Case study

Figure 3: Failure time $\mathbf{T}_f$ distribution.

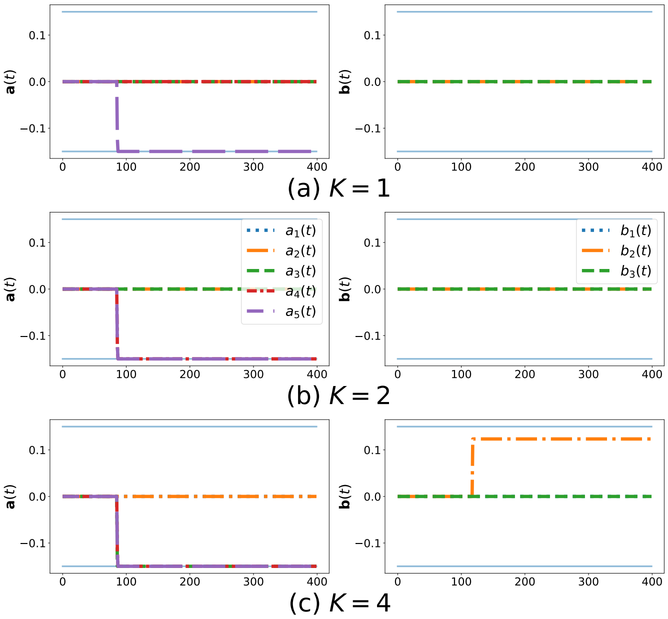
5.1.2 Results

For each random attack graph type, the optimal attack actions on the cyber system y_{ij} and the attack state start times h_i are denoted in Figure 2 when $K = 3$. Malicious data injection starts when an attacker compromises sensors or controllers. Considering attack graph (a) in Figure 2, the attacker compromises the controller C_3 at $t = 84$; Therefore, $\mathbf{b}_{C_3,t} = 0$ for $t < 84$, and, after that, $\mathbf{b}_{C_3,t}$ can be determined by the attacker. Despite existing sensors and controllers that were easier to reach regarding exploitation time, the attacker targeted those that led to a noticeable physical impact. This observation denotes that, from the attacker's perspective, an ICS's cyber and physical systems are not independent, and the attacker's decision to exploit the vulnerabilities in the cyber system is correlated with the physical impact that will be imposed on the physical system. Furthermore, from Figure 2, we observe that the different topology of the attack graph leads to different times that the attacker can reach the



(a) Numerical study (b) Case study

Figure 4: Mean failure time $\mathbb{E}(\mathbf{T}_f)$.



(a) $K = 1$

(b) $K = 2$

(c) $K = 4$

Figure 5: Attacker's optimal actions on sensors and controllers for $T = 600$.

physical system. Hence, implementing the mitigations in the cyber system should not only focus on increasing the difficulty of reaching the most accessible targets, but it also is essential to focus on the impacts that a cyberattack can impose on the physical system.

Figure 3(a) denotes the distribution of the failure time for the physical system under normal operation and the system under worst-case attack when $K = 1, 3, 6$ and $T = 200, 400, 1000$. We observe that the failure rate increases as K and T increase, which shows how the worst-case malicious actions accelerate the failure time. However, the increase in K has more effect on increasing the failure rate. Figure 4(a) elaborates further on this observation. Regardless of the attacker's available time T, as K increases, the failure rate increases and, consequently, the attacker's MFT decreases. However, when $T \geq 500$, the failure time acceleration remains almost the same as K increases. This observation denotes that if the attacker has limitations on compromising the sensors and controllers but can execute a cyberattack for a longer time, the physical consequences will not be significant, and, accordingly, the chance of being detected by the intrusion detection system will increase.

5.2 Case Study

5.2.1 Experiment Setup

The HIL case study described in [11] is investigated here as an ICS comprising both cyber and physical systems. As the physical system, a boiling water power plant (BWPP) contains three states to be controlled (drum pressure, electric output, and fluid density), three controllers, a water valve (WV), a fuel valve (FV), a steam valve (SV), and five sensors, two pressure sensors (P1 and P2), two water level sensors (L1 and L2), and a sensor (GT) for reading the generated electricity in the field area. The cyber system consists of an administrator host (AH) and two PCs in a corporate network, a human-machine interface (HM) and a data server (DS) in a supervisory network, and three embedded controllers in a control network. Given the vulnerability list and their exploitation times for this case, we construct the corresponding attack graph of the cyber system.

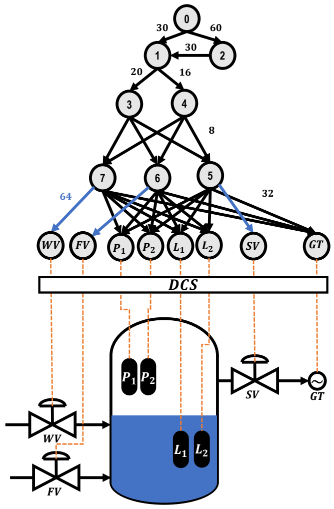


Figure 6: Cyber-physical system for the case study. Blue edges on the attack graph have the same exploitation time 64.

Figure 6 denotes the cyber-physical structure of this case. The graph represents different states, starting from initial access ('0') to acquiring user privilege on AH and PCs ('1' and '2'), root privilege on HM and DS ('3' and '4'), and successfully acting as HM and sending forged Modbus commands ('5', '6', and '7'). The remaining states correspond to the compromised sensors and controllers. Huang et al. [11] reported the same vulnerability and exploitation time for both sensors and controllers. However, compromising controllers requires more effort, so we consider their exploitation time to be twice as long. The shortest path length to compromised sensors is 86 time units, while for controllers it is 118 time units.

Given the physical system dynamics, we define $\tilde{\mathbf{x}}_t = \mathbf{x}_t - \bar{\mathbf{x}}$ where $\bar{\mathbf{x}}$ is the reported steady state of the system. Since $\tilde{\mathbf{x}}_t$ exists, and it must be that $\mathbb{E}(\tilde{\mathbf{x}}_t) = 0$, any linear function $f(\tilde{\mathbf{x}}_t) = \beta_0 + \beta^T \tilde{\mathbf{x}}_t$ exists such that $\mathbb{E}(f(\tilde{\mathbf{x}}_t)) = \beta_0$. As discussed in Section 3.3, the degradation rate is a function of $\mathbf{x}_t$. Therefore, we assume exact values of κ and γ of this case depend on the plant type, and further investigations are out of the scope of this study. Hence, for simplicity, and without losing generality, we generate random values between 0 and 1 as degradation parameters. Also, we consider $\sigma_s^2 = 0.01$. In the same way as the numerical study, we obtain λ, δ, $\underline{\mathbf{a}}_i$, $\bar{\mathbf{a}}_i$, $\underline{\mathbf{b}}_i$, and $\bar{\mathbf{b}}_i$. It is worth mentioning that we do not conduct physical experiments but we use the dynamics provided in [11] in our risk assessment model in (14).

5.2.2 Results

Figure 3(b) denotes the p.d.f. the failure time for the normal system operation and the system under worst-case attack when $K = 1, 4, 8$ and $T = 100, 400, 1000$. We observe the same results as the numerical study, and it validates our observations on a real system. However, when $T = 100$, the failure rate distribution is close to the normal system regardless of increase in K. Since the shortest path length to the controllers is higher than the attacker's available time, the attacker could only compromise the sensors (depending on K). This observation denotes the impor-

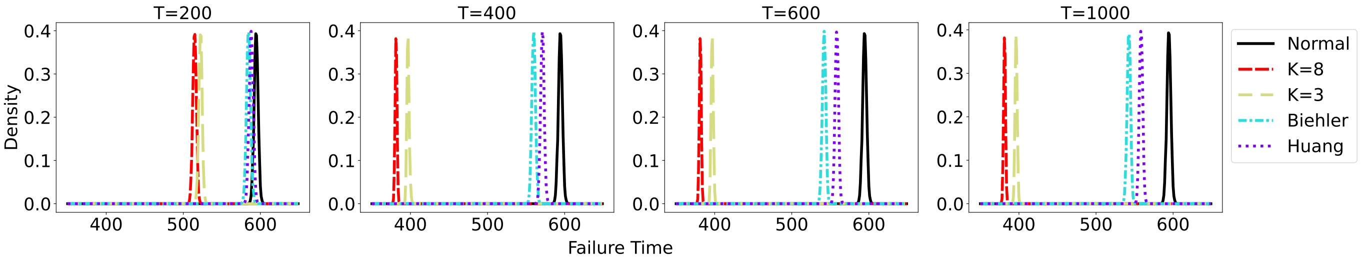


Figure 7: Failure time distribution of attack actions: Huang et al. [11] attack scenario vs. worst-case attack obtained by our framework ($K = 3$); Biehler et al. [5] attack generation model vs. worst-case attack obtained by our framework ($K = 8$).

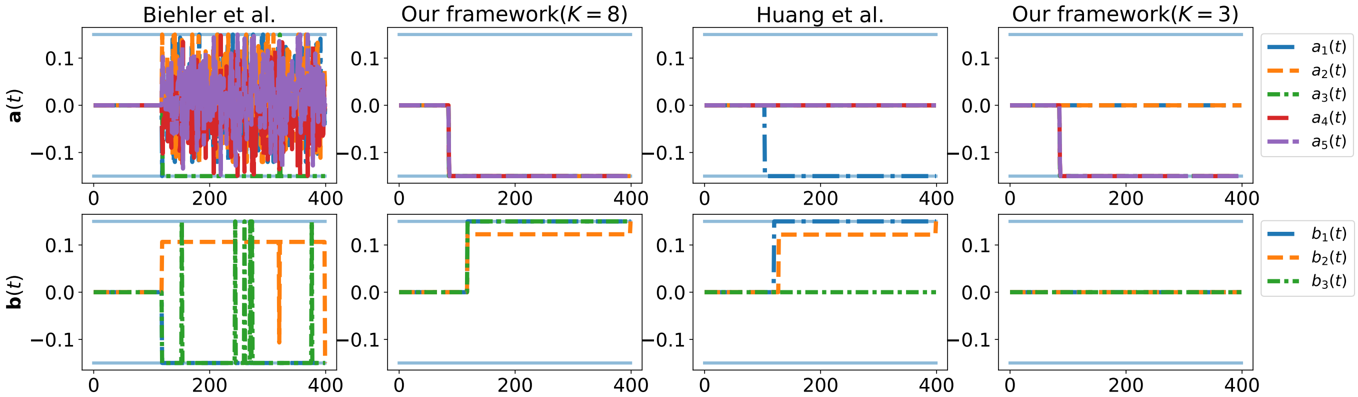


Figure 8: Attack actions: Huang et al. [11] attack scenario vs. worst-case attack obtained by our framework ($K = 3$); Biehler et al. [5] attack generation model vs. worst-case attack obtained by our framework ($K = 8$).

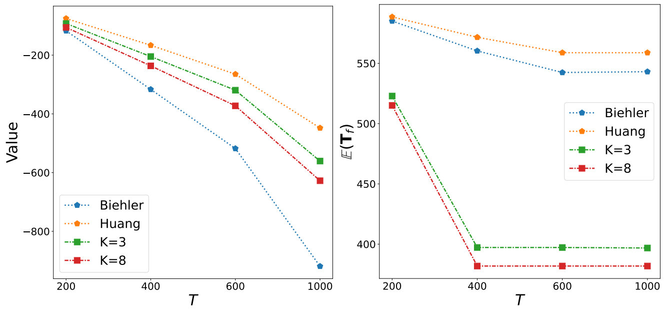


Figure 9: Damage function (15) value (left) and $E(\mathbf{T}_f)$ (right) of attack actions: Huang et al. [11] attack scenario vs. worst-case attack obtained by our framework ($K = 3$); Biehler et al. [5] attack generation model vs. worst-case attack obtained by our framework ($K = 8$).

tance of protecting controllers from cyberattacks. Even though when the attacker has high resources ($K \geq 6$), the controllers remain secured, and consequently, the imposed physical impact is insignificant.

Figure 4(b) elaborates more on this observation. We observe a noticeable jump in MFT when T increases from 200 to 400 as K increases. Figure 5 magnifies the worst-case attack actions that led to this observation. We observe that as K increases, the attacker could compromise the controller SV, leading to aggressive actions on the sensors GT, L_1, and L_2. Furthermore, from $K = 2$ to $K = 4$, the attacker causes significant reduction in MFT by compromising SV and GT and manipulating L_1 and L_2 in the same way as $K = 2$.

5.2.3 *Comparative Analysis*

Due to the different nature of the cyber-risk assessment methods in the literature, it is hard to find an ICS containing their required cyber and physical system information in order to establish a fair way to compare our framework with them. Therefore, our intention in this section is not to show that our framework outperforms the other methods. Instead, we point out the significant characteristics of an ICS that are incorporated in our formulation.

Huang et al. [11] designed an attack scenario to validate their proposed method. In this scenario, the attacker follows the path $0 \rightarrow 1 \rightarrow 3 \rightarrow (5, 6) \rightarrow (P_1, SV, FV)$ on the attack graph (Figure 6) and compromises P_1, FV, and SV at time 104, 120, 128 respectively. Then, the attacker starts attack actions on the BWPP through P_1, FV, and SV. This scenario is equivalent to a case where $K = 3$ for some attack execution time T. For $T \in \{200, 400, 600, 100\}$, we implement this attack scenario on our framework in (14). Also, we solve (14) for the worst-case attack when $K = 3$. The worst-case attack found by our framework follows the path $0 \rightarrow 1 \rightarrow 4 \rightarrow 5 \rightarrow (L_1, L_2, GT)$ on the attack graph (Figure 6) and compromises L_1, L_2, and GT at time 86. Figures 7 and 9 denote the failure time distribution and the MFT of BWPP under Huang's attack actions vs. the worst-case attack obtained from our framework when $K = 3$. We observe that the MFT of the BWPP is further decreased through the attack actions obtained from our framework. Figure 8 magnifies this observation by looking at the attack actions on BWPP. We observe that the worst-case attack with $K = 3$ can significantly reduce MFT by only compromising the sensors L_1, L_2, GT that require less effort than controllers.

Bielher et al. [5] proposed a generic attack generation framework on physical system without considering the cyber system and assuming the attacker has full access to the physical system. Their approach is equivalent to the case where $K = 8$ and some value of T in our formulation. Our physical constraints (14l)-(14q) correspond to the constraints in their formulation with special case functions (i.e., linear functions) and no cost on physical attack actions. To elaborate more on this, we formulate the BWPP as (1) that can be re-written as $\mathbf{z}_t = DB\mathbf{u}_t + DA\mathbf{x}_t + \mathbf{v}'_t$. By (4), the BWPP process output under attack is $\mathbf{z}_t^A = \mathbf{z}_t + DB\mathbf{b}_t + \mathbf{a}_t + \mathbf{v}'_t$. Therefore, the objective function (damage function) in their formulation is

$$\min_{\mathbf{a}_t,\mathbf{b}_t} -\sum_{t=1}^{T}\left\|\mathbf{z}_t^{ref} - DB\mathbf{b}_t - \mathbf{a}_t\right\|_1 \qquad (15)$$

where $\mathbf{z}_t^{ref}$ is the sensor measurements under normal operations. The constraint of avoiding from being detected in their formulation is equivalent to (8). The physical limits constraint in their formulation is equivalently modeled in our framework in (5) and (6).

Showing the equivalence between our physical system constraints and the attack generation model in [5], we solve our framework (14) for $K = 8$ and $T \in \{200, 400, 600, 1000\}$ by replacing (15) with (14b) and (14a) to generate Biehler's attack actions. Figures 7 and 9 denote the failure time distribution and the MFT of BWPP under Biehler's attack actions vs. the worst-case attack obtained from our framework when $K = 8$. We observe that the MFT of the BWPP is further decreased through the attack actions obtained from our framework. Figure 8 elaborates more on this observation by looking at the attack actions on BWPP. The pattern of the attack actions generated by our framework and Biehler's attack actions are noticeably different. This variation is due to considering different physical dynamics to evaluate the physical impact of an attack action. Figure 9 (left) denotes the value of the objective function (15) in Biehlers's formulation for the attack actions obtained by solving our framework for $K = 3$ and $K = 8$, attack actions obtained from Biehlers's formulation, and the Huang's attack scenario. The objective function (15) basically determines attack actions that has high l_1−norm distance with the reference sensors measurement. We observe that Biehler's attack actions generate attacks that alter the sensors measurement more than the other approaches even though these attack actions do not affect MFT significantly.

6 Conclusion

In this article, we present a holistic optimization framework for identifying the most critical cyberattacks on ICSs by jointly modeling the cyber and physical systems of the ICSs to assess and infer their physical outcome. The attacker's problem is formulated as a MILP where the cyber system is modeled as an attack graph with multiple target states, each corresponding to a state where a particular sensor or controller is compromised. The physical system is formulated as a state-space LTI model subject to sensor and control data manipulation. To assess the impact of a cyberattack on the physical system, we have developed a novel objective function based on the degradation of the physical system. Designing robust cyberattack detection schemes and mitigation strategies in ICSs is challenging due to the need for real-case attack data. Our framework serves as a tool for generating the worst-case attack with various levels of expertise and resources to manipulate sensor measurements and control actions while remaining undetected by the detection mechanisms. These worst-case attacks can be used to assess the risk while developing the detection scheme for the ICSs.

The experimental results validate our framework's efficiency and applicability. The worst-case cyberattack generated by our framework can have more devastating physical consequences than a random cyberattack, thus directly highlighting its effectiveness in developing mitigating strategies. Nevertheless, our current proposed framework lacks the inclusion of uncertainties present in both cyber and physical systems. The stochastic nature of these systems will be incorporated into our future analysis.

References

[1] Manuel Cheminod, Luca Durante, and Adriano Valenzano. Review of security issues in industrial networks. *IEEE transactions on industrial informatics*, 9(1):277–293, 2012.

[2] Kunwu Zhang, Yang Shi, Stamatis Karnouskos, Thilo Sauter, Huazhen Fang, and Armando Walter Colombo. Advancements in industrial cyber-physical systems: an overview and perspectives. *IEEE Transactions on Industrial Informatics*, 2022.

[3] Xirong Ning and Jin Jiang. Design, analysis and implementation of a security assessment/enhancement platform for cyber-physical systems. *IEEE Transactions on Industrial Informatics*, 18(2):1154–1164, 2021.

[4] Heegyun Jeon and Yongsoon Eun. A stealthy sensor attack for uncertain cyber-physical systems. *IEEE Internet of Things Journal*, 6(4):6345–6352, 2019.

[5] Michael Biehler, Zhen Zhong, and Jianjun Shi. Sage: Stealthy a ttack ge neration in cyber-physical systems. *IISE Transactions*, pages 1–15, 2023.

[6] Kangkang Zhang, Christodoulos Keliris, Thomas Parisini, and Marios M Polycarpou. Stealthy integrity attacks for a class of nonlinear cyber-physical systems. *IEEE Transactions on Automatic Control*, 67(12):6723–6730, 2021.

[7] Ankica Barišić, Ivan Ruchkin, Dušan Savić, Mustafa Abshir Mohamed, Rima Al-Ali, Letitia W Li, Hana Mkaouar, Raheleh Eslampanah, Moharram Challenger, Dominique Blouin, et al. Multi-paradigm modeling for cyber–physical systems: A systematic mapping review. *Journal of Systems and Software*, 183:111081, 2022.

[8] Harjinder Singh Lallie, Kurt Debattista, and Jay Bal. A review of attack graph and attack tree visual syntax in cyber security. *Computer Science Review*, 35:100219, 2020.

[9] Matthias Eckhart, Andreas Ekelhart, Stefan Biffl, Arndt Lüder, and Edgar Weippl. Qualsec: An automated quality-driven approach for security risk identification in cyber-physical production systems. *IEEE Transactions on Industrial Informatics*, 2022.

[10] Huy Nguyen and Thomas C Sharkey. A computational approach to determine damage in infrastructure networks from outage reports. *Optimization Letters*, 11:753–770, 2017.

[11] Kaixing Huang, Chunjie Zhou, Yu-Chu Tian, Shuanghua Yang, and Yuanqing Qin. Assessing the physical impact of cyberattacks on industrial cyber-physical systems. *IEEE Transactions on Industrial Electronics*, 65(10):8153–8162, 2018.

[12] Xuan Li, Chunjie Zhou, Yu-Chu Tian, Naixue Xiong, and Yuanqing Qin. Asset-based dynamic impact assessment of cyberattacks for risk analysis in industrial control systems. *IEEE Transactions on Industrial Informatics*, 14(2):608–618, 2017.

[13] Song Deng, Jiantang Zhang, Di Wu, Yi He, Xiangpeng Xie, and Xindong Wu. A quantitative risk assessment model for distribution cyber physical system under cyber attack. *IEEE Transactions on Industrial Informatics*, 2022.

[14] Thomas C Sharkey, Sarah G Nurre Pinkley, Daniel A Eisenberg, and David L Alderson. In search of network resilience: An optimization-based view. *Networks*, 77(2): 225–254, 2021.

[15] Forough Enayaty-Ahangar, Laura A Albert, and Eric DuBois. A survey of optimization models and methods for cyberinfrastructure security. *IISE Transactions*, 53(2): 182–198, 2020.

[16] Kaiyue Zheng, Laura A Albert, James R Luedtke, and Eli Towle. A budgeted maximum multiple coverage model for cybersecurity planning and management. *IISE Transactions*, 51(12):1303–1317, 2019.

[17] Kaiyue Zheng and Laura A Albert. A robust approach for mitigating risks in cyber supply chains. *Risk Analysis*, 39 (9):2076–2092, 2019.

[18] Adam Schmidt, Laura A Albert, and Kaiyue Zheng. Risk management for cyber-infrastructure protection: A bi-objective integer programming approach. *Reliability Engineering & System Safety*, 205:107093, 2021.

[19] Kaiyue Zheng and Laura A Albert. Interdiction models for delaying adversarial attacks against critical information technology infrastructure. *Naval Research Logistics (NRL)*, 66(5):411–429, 2019.

[20] MHR Khouzani, Zhengliang Liu, and Pasquale Malacaria. Scalable min-max multi-objective cyber-security optimisation over probabilistic attack graphs. *European Journal of Operational Research*, 278(3):894–903, 2019.

[21] Xuan Li, Chunjie Zhou, Yu-Chu Tian, and Yuanqing Qin. A dynamic decision-making approach for intrusion response in industrial control systems. *IEEE Transactions on Industrial Informatics*, 15(5):2544–2554, 2018.

[22] Nail Orkun Baycik, Thomas C Sharkey, and Chase E Rainwater. Interdicting layered physical and information flow networks. *IISE Transactions*, 50(4):316–331, 2018.

[23] Huaizhi Wang, Jiaqi Ruan, Guibin Wang, Bin Zhou, Yitao Liu, Xueqian Fu, and Jianchun Peng. Deep learning-based interval state estimation of ac smart grids against sparse cyber attacks. *IEEE Transactions on Industrial Informatics*, 14(11):4766–4778, 2018.

[24] Dan Li, Nagi Gebraeel, Kamran Paynabar, and A. P. Sakis Meliopoulos. An online approach to covert attack detection and identification in power systems. *IEEE Transactions on Power Systems*, 38(1):267–277, 2023. doi: 10.1109/TPWRS.2022.3167024.

[25] Dan Li, Nagi Gebraeel, and Kamran Paynabar. Detection and differentiation of replay attack and equipment faults in scada systems. *IEEE Transactions on Automation Science and Engineering*, 18(4):1626–1639, 2021. doi: 10.1109/TASE.2020.3013760.

[26] Weiming Fu, Jiahu Qin, Yang Shi, Wei Xing Zheng, and Yu Kang. Resilient consensus of discrete-time complex cyber-physical networks under deception attacks. *IEEE Transactions on Industrial Informatics*, 16(7):4868–4877, 2019.

[27] Alan Oliveira de Sá, Luiz F Rust da Costa Carmo, and Raphael CS Machado. Covert attacks in cyber-physical control systems. *IEEE Transactions on Industrial Informatics*, 13(4):1641–1651, 2017.

[28] Ge Cao, Wei Gu, Peixin Li, Wanxing Sheng, Keyan Liu, Lijing Sun, Zhihuang Cao, and Jing Pan. Operational risk evaluation of active distribution networks considering cyber contingencies. *IEEE Transactions on Industrial Informatics*, 16(6):3849–3861, 2019.

[29] Kaixing Huang, Chunjie Zhou, Yuanqing Qin, and Weixun Tu. A game-theoretic approach to cross-layer security decision-making in industrial cyber-physical systems. *IEEE Transactions on Industrial Electronics*, 67(3):2371–2379, 2019.

[30] Naipeng Li, Yaguo Lei, Nagi Gebraeel, Zhijian Wang, Xiao Cai, Pengcheng Xu, and Biao Wang. Multi-sensor data-driven remaining useful life prediction of semi-observable systems. *IEEE Transactions on Industrial Electronics*, 68(11):11482–11491, 2020.

[31] Dan Li, Kamran Paynabar, and Nagi Gebraeel. A degradation-based detection framework against covert cyberattacks on scada systems. *IISE Transactions*, 53(7): 812–829, 2021.

[32] Yizhen Peng, Yu Wang, and Yanyang Zi. Switching state-space degradation model with recursive filter/smoother for prognostics of remaining useful life. *IEEE Transactions on Industrial Informatics*, 15(2):822–832, 2018.