Ensuring Resilience Against Stealthy Attacks on Cyber-Physical Systems

Paul Griffioen ©, Member, IEEE, Bruce H. Krogh ©, Life Fellow, IEEE, and Bruno Sinopoli ©, Fellow, IEEE

Abstract—This article provides a tool for analyzing mechanisms that aim to achieve resilience against stealthy, or undetectable, attacks on cyber-physical systems. We consider attackers who are able to corrupt all of the inputs and outputs of the system. To counter such attackers, a response scheme must be implemented that keeps the attacker from corrupting the inputs and outputs of the system for certain periods of time. To aid in the design of such a response scheme, our tool provides sufficient lengths for these periods of time in order to ensure safety with a particular probability. We provide an upper bound on how long the system can remain under stealthy attack before the safety constraints are violated. Furthermore, we show how a detector limits the set of biases an attacker can exert on the system while still remaining stealthy, aiding a system operator in the design of the detector. Our contributions are demonstrated with an illustrative example.

Index Terms—Cyber-physical systems (CPSs), fault detection, fault-tolerant systems, robust control.

I. INTRODUCTION

ECURING cyber-physical systems (CPSs), engineered systems that include sensing, processing, control, and communication in physical spaces, is essential in today’s society. CPSs are ubiquitous in modern critical infrastructures including manufacturing, transportation systems, energy delivery, health care, water management, and smart grid. The presence of heterogeneous components and devices creates numerous attack surfaces in these large-scale, highly-connected systems. Consequently, these systems are attractive targets for adversaries and are essential to protect.

Due to the strong coupling between cyber and physical domains, the tools and methodologies developed to ensure security in the cyber domain alone are insufficient to secure CPSs. Techniques within cybersecurity such as authenticated encryption, message authentication codes (MACs), and digital signatures are often computationally expensive to implement and fail to recognize purely physical attacks. For example, the integrity of sensor measurements can be modified by changing a sensor’s local environment while control inputs can be changed by directly manipulating system actuators. In such a scenario, MACs or digital signatures fail to recognize an attack.

CPS vulnerabilities have culminated in several effective attacks from highly resourceful and knowledgeable adversaries. In the Maroochy Shire incident [1], a malicious insider was able to utilize detailed system knowledge to attack a waste management system in Queensland, Australia, resulting in the leakage of millions of liters of sewage. With the Stuxnet attack [2], a nation state adversary was able to compromise a uranium enrichment facility in Iran, leading to the destruction of a thousand centrifuges. In 2015, hackers were able to remotely compromise a supervisory control and data acquisition system in Ukraine [3], allowing them to cause widespread blackouts.

Motivated by the threat of such sophisticated attackers, we aim to design resilient CPSs by providing a set of mechanisms and tools that can be used to preserve safety while functionality is restored in the presence of attacks. In this work, we present a tool that provides an upper bound on the amount of time a system can be under stealthy attack while still remaining probabilistically safe. This tool is general enough that it can be used in the design and analysis of any resilience mechanisms. The importance of this tool lies in its ability to analyze stealthy attacks, which are by definition very unlikely to be detected. Unlike other attacks where mitigation strategies are deployed once an attack is detected, for stealthy attacks, these mitigation strategies are rarely deployed because the attack is rarely detected. Consequently, it is very important to understand how long it will take for stealthy attacks to render the system unsafe. Once this time period has been determined, prevention mechanisms can be employed frequently enough to ensure that potential stealthy attacks never have enough time to render the system unsafe.

A number of works have developed similar metrics and tools for resilience analysis and security [4], [5], [6], [7], [8]. In contrast to these works, we assume that adversaries can modify all of the actuators and sensor measurements while Murguia et al. [4] only consider attacks on the sensor measurements, and the authors in [7] and [8] only consider attacks on the actuators. We also assume an adversary can read all the inputs and outputs of the system while the authors in [6] and [8] assume...
that an adversary cannot read any of the control inputs or sensor measurements. Furthermore, Perodou et al. [7] do not address stealthy attacks, and Teixeira [5] does not analyze the amount of time it would take for the safety constraints to be violated under a stealthy attack.

The rest of this article is organized as follows. Section II introduces the problem being solved, and Section III surveys a variety of mechanisms used to achieve resilience, specifically those used for detection and response. Section IV introduces the system model used in analyzing these response mechanisms, along with the control, estimation, and detection schemes. Section V describes stealthy adversaries, analyzing the set of biases they are able to exert on the system without being detected. Section VI provides sufficient conditions for ensuring the probabilistic safety of the overall system, describing how the length of time under which the components can be trusted directly affects whether or not these conditions are satisfied. Section VII includes results from an example illustration. Finally, Section VIII concludes this article.

**Notation:** We let $\mathcal{N}(\mu, \Sigma)$ denote a Gaussian distribution with mean vector $\mu$ and covariance matrix $\Sigma$. We let $\chi^2(n, \lambda)$ denote a noncentral $\chi^2$ distribution with $n$ degrees of freedom and with noncentrality parameter $\lambda$. We let $\mathbb{E}[X]$ and $\text{Cov}(X)$ denote the expected value and covariance, respectively, of a random variable $X$. We let $p(Y)$ represent the probability of event $Y$. We let $I$ represent the identity matrix, and we let $\text{BlkDiag}(A_1, \ldots, A_N)$ denote a block diagonal matrix with matrices $A_1, \ldots, A_N$ on the diagonal.

## II. PROBLEM FORMULATION

In this article, we address the problem of designing resilient CPSs by providing system operators with a general framework that can be used to analyze the resilience of any design. In designing resilient CPSs, two necessary components include detection and response. When designing detection and response mechanisms, two important degrees of freedom are usually leveraged: 1) trust and 2) time. First, there must exist a component of the system that serves as the root of trust. Second, the time periods where certain components and information may be trusted need to be identified and used to the defender’s advantage.

In designing resilient detection and response mechanisms, we want to create time periods where we can trust certain components and information and use those components and information to recover the system from any potential attacks. Consequently, this article seeks to analyze those periods of time, investigating how long the components need to be restored to full functionality to ensure safety against stealthy attacks. Since a response scheme can be implemented once an attack is detected, we are interested in understanding when a response scheme should be implemented to mitigate attacks that are not detected. More specifically, we investigate how long the system can tolerate all the components malfunctioning or being manipulated by a stealthy adversary before restoration of those components must take place. This analysis will aid system operators as they design particular response strategies, no matter what the response strategy is. We provide sufficient conditions that the response strategy should meet in order to ensure the probabilistic safety of the overall system against stealthy attacks.

More formally, the problem addressed in this article can be stated as follows.

**Problem 1:** Given a set of safety constraints $X$, an initial state $x_0 \in X$, and a particular response strategy, how often does that response strategy need to be implemented in order to ensure that $\Pr(x_k \in X) \geq p \ \forall k$ when the system is under stealthy attack?

In summary, then, this article answers the following question: For a system designer who wants to ensure that the state remains within the safety constraints with a particular probability, what is the maximum amount of time the system can be under a stealthy attack before the safety constraints are violated? The response strategy can then be designed and implemented frequently enough to ensure that the system will never be under stealthy attack for any longer than this amount of time.

## III. PREVIOUS WORK

Since this article provides a general framework to aid system operators in analyzing the resilience of any design they use, we first provide a brief overview of a variety of detection and response mechanisms that may be used in designing resilient CPSs.

### A. Detection

There are two main forms of detection: 1) passive and 2) active. Passive detection techniques process the defender’s information to make a binary decision about the system, outputting either the null hypothesis (normal system operation) or the alternative hypothesis (the system is under attack). While passive detection techniques are effective against benign faults, bad data detectors can oftentimes be bypassed by powerful adversaries who leverage access to system channels and/or model knowledge to construct attacks so that the outputs received by a system operator are statistically consistent with expected output behavior. Since passive detection techniques provably fail in these instances, active detection techniques must be used to detect malicious attacks, where a defender intelligently changes the policy online by adding perturbations to the system [9], [10].

---

**Fig. 1.** Active detection mechanism for attacks on (a) the sensor measurements and (b) the control inputs and sensor measurements. (a) Physical watermarking. (b) Moving target defense.
Software rejuvenation is an established method for dealing with so-called software aging in traditional computing systems or failures that occur when a running program encounters a state that was not anticipated when the software was designed [20], [21]. In CPSs, software rejuvenation consists of periodically refreshing the run-time system with a secure and trusted copy of the control software, recovering the system from attacks during time periods called secure execution intervals when the control software can be trusted [22], [23], [24], [25], [26].

Overlay networks can be used to ensure CPS resilience against man-in-the-middle attacks and denial-of-service (DoS) attacks [27], [28], [29], [30], [31], [32]. This is accomplished by routing around malicious nodes and links in a network, specifying the pathway that is used each time data are sent between the controller and the plant. Reachability analysis can be used to ensure that the safety constraints are not violated, even when a large portion of the pathways across the network have been compromised by adversaries [33].

Reconfigurable control [34], [35] and resilient estimation [36], [37] utilize functional redundancy in the system, ensuring resilience when up to a particular subset of actuator commands and sensor measurements are compromised. Reconfigurable control adapts the control policy to use those actuators that have not been compromised to preserve functionality while resilient estimation constructs an accurate state estimate from those sensor measurements that have not yet been compromised.

IV. SYSTEM MODEL

A. System Dynamics

To begin, we introduce the model for the system under consideration. We model the CPS as a discrete-time linear-time invariant system where the system dynamics are given by

\[
\begin{align*}
x_{k+1} &= \begin{cases} 
    Ax_k + Bu_k + w_k, & \text{normal operation} \\
    Ax_k + B(u_k + w_k^a) + w_k, & \text{attack}
\end{cases} \\
y_k &= \begin{cases} 
    Cx_k + v_k, & \text{normal operation} \\
    Cx_k + y_k^a + v_k, & \text{attack}
\end{cases}
\end{align*}
\]

where normal operation is guaranteed when the system operates securely, free from attacks. Here, \( x_k \in \mathbb{R}^n \) represents the system state at time step \( k \), \( y_k \in \mathbb{R}^m \) denotes the sensor measurements, and \( u_k \in \mathbb{R}^l \) is the control input vector. To capture uncertainty, we consider independent and identically distributed (i.i.d.) Gaussian process noise \( w_k \sim \mathcal{N}(0, Q) \) and i.i.d. Gaussian sensor noise \( v_k \sim \mathcal{N}(0, R) \). We assume that \( (A, B) \) and \( (A, Q^\frac{1}{2}) \) are stabilizable, \( (A, C) \) is detectable, and \( R > 0 \).

In this article, we consider an adversary who can perform integrity attacks, where the attacker is able to corrupt all of the inputs and outputs, denoted by adding a bias of \( w_k^a \) and \( y_k^a \) to the control inputs and sensor measurements, respectively [38], [39]. Without loss of generality, an attack is assumed to begin at time step \( k = 0 \). This can be carried out for example through a memory corruption attack where the adversary corrupts the control software and injects malicious code [40]. Alternatively, this can be carried out through a man-in-the-middle attack where the adversary intercepts and modifies the data passing over
the network. This can also be carried out by modifying the physical actuators and sensors themselves, where an adversary may change the settings of the actuators or the environment surrounding the system sensors. Motivated by the potential resources of sophisticated malicious adversaries, we consider an attacker who can modify all the inputs and outputs.

**Remark 1:** In the case where an adversary carries out a memory corruption attack, the results presented in this article are only applicable when the state estimator and detector are not corrupted under attack. This can be obtained, for instance, by implementing them within a secure hypervisor. However, no such qualification needs to be made if the adversary carries out the attack through other means, including man-in-the-middle attacks and attacks on the physical actuators and sensors themselves.

We will also consider the possibility that an attacker has detailed system knowledge, including knowledge of the plant as well as the ability to read all of the inputs and outputs. This can be obtained through knowledge of the first principles, which govern the dynamics or through system identification stemming from passive observation of the inputs and outputs. This knowledge, when combined with the ability to corrupt all the inputs and outputs, can lead to powerful undetectable attacks [41]. For example, an attacker can subtract his or her effect on the system dynamics by choosing an arbitrary sequence of attack inputs \( \{u_k\} \) and then constructing the sensor measurement bias according to

\[
y_k^a = -Cx_k^a, \quad x_{k+1}^a = Ax_k^a + Bu_k^a, \quad x_0 = 0.
\]

In this way, an attacker can arbitrarily perturb the system along the controllable subspace \((A,B)\). Due to the linearity of the system, the probability distribution of the outputs under such an attack is identical to the probability distribution of the outputs under normal operation. Consequently, no standard bad data detector can recognize this adversarial behavior, and as a result, this behavior is perfectly stealthy. In this article, we will analyze how resilient a system is to these stealthy attacks.

**B. State Estimation**

A Kalman filter is a linear estimator, which can be used to compute the minimum mean-squared error state estimate \( \hat{x}_{k|k} \) given the set of previous measurements up through \( y_k \). We assume that the system has been running for a long time so that the Kalman filter has converged to a fixed gain linear estimator given by

\[
\hat{x}_{k+1|k} = A\hat{x}_{k|k} + Bu_k \tag{4}
\]

\[
\hat{x}_{k|k} = \hat{x}_{k|k-1} + K(y_k - C\hat{x}_{k|k-1}) \tag{5}
\]

where \( \hat{x}_{k+1|k} \) is the a priori state estimate and \( \hat{x}_{k|k} \) is the a posteriori state estimate. The Kalman gain \( K \) and error covariance matrix \( P \) are given by

\[
K = PCT(CPC^T + R)^{-1} \tag{6}
\]

\[
P = APAT - APC^T(CPC^T + R)^{-1}CPAT + Q. \tag{7}
\]

The control input is given by \( u_k = L\hat{x}_{k|k} \), where \( L \) is the control gain matrix. One possible way to design \( L \) is to minimize a quadratic function of the states and inputs according to

\[
J = \lim_{T \to \infty} \frac{1}{T+1} \mathbb{E} \left[ \sum_{k=0}^{T} x_k^TWx_k + u_k^TVu_k \right] \tag{8}
\]

where \( W \) and \( V \) are positive-definite matrices that, respectively, denote the relative cost of each state and control input. Then, the optimal \( L \) is given by

\[
L = -(B^TSB + V)^{-1}B^TS \tag{9}
\]

\[
S = ATSA + W - ATSB(B^TSB + V)^{-1}B^TS \tag{10}
\]

where \( S \) satisfies the Riccati equation in (10) [42].

The residue \( z_k \) is a function of the a priori state estimate and the Kalman filter has converged to a fixed gain linear estimator

\[
z_k \triangleq y_k - C\hat{x}_{k|k-1}. \tag{11}
\]

Depending on whether or not the system is under attack, the residue takes on different values, given by

\[
z_k = \begin{cases} 
  z_k^n, & \text{normal operation} \\
  z_k^n + \Delta z_k, & \text{attack} 
\end{cases} \tag{12}
\]

where \( z_k^n \) represents the residue under normal operation, which follows a normal distribution according to \( z_k^n \sim \mathcal{N}(0, \Sigma) \) with \( \Sigma \triangleq CPC^T + R \). When the system is under attack, the cumulative set of input and output biases \( \{u_k^n, y_k^n\} \) results in the attacker exerting a bias \( \Delta z_k \) on the residue, given by

\[
\Delta z_k = y_k^n + \sum_{j=0}^{k-1} C(A(I - KC))^{k-1-j}(Bu_j^n - AKy_j^n). \tag{13}
\]

According to the system model in (1) and (2) and the Kalman filter in (4) and (5), the error dynamics are given by

\[
e_{k+1|k} = \begin{cases} 
  Ae_{k|k} + w_k - Kz_{k+1}^n, & \text{normal operation} \\
  Ae_{k|k} + w_k - K(z_{k+1}^n + \Delta z_{k+1}) + Bu_k^n, & \text{attack} 
\end{cases} \tag{14}
\]

where \( e_{k+1|k} \triangleq x_{k+1} - \hat{x}_{k+1|k} \) is the a posteriori state estimation error.

The overall system dynamics are then given by

\[
\bar{x}_{k+1} = \begin{cases} 
  A\bar{x}_k + Iw_k + Kz_{k+1}^n, & \text{normal operation} \\
  A\bar{x}_k + Iw_k + K(z_{k+1}^n + \Delta z_{k+1}) + Bu_k^n, & \text{attack} 
\end{cases} \tag{15}
\]

where \( \bar{x}_k \triangleq \left[ x_k^T \quad e_{k|k}^T \right]^T \), \( A \triangleq \begin{bmatrix} A & BL \\ 0 & A \end{bmatrix}, \quad I \triangleq \begin{bmatrix} I \\ I \end{bmatrix}, \quad K \triangleq \begin{bmatrix} 0 \\ -K \end{bmatrix}, \quad B \triangleq \begin{bmatrix} B \\ 0 \end{bmatrix} \).

**C. Detection**

To detect attacks on the CPS, a residue-based detector called the \( \chi^2 \) detector is utilized and is given by

\[
g(z_{k-T+1:k}) = \sum_{i=k-T+1}^{k} z_i^T \Sigma^{-1} z_i \geq \eta \tag{16}
\]
where $T$ represents the detector window that considers past measurements, and the detection statistic $g$ follows a $\chi^2$ distribution with $mT$ degrees of freedom under normal operation. The $\chi^2$ detector attempts to exploit this fact by testing to see if the residues follow the correct distribution. Here, $\eta$ represents the threshold of the bad data detector, $H_0$ is the null hypothesis that represents normal system operation, and $H_1$ is the alternative hypothesis that denotes that the system has deviated from normal operation. Measurements that are in close agreement with expected values generate small detection statistics and, thus, raise no alarm. Large deviations between measured and expected behavior will lead to a large detection statistic, thus causing an alarm.

When the system is under attack, the detection statistic $g$ follows a noncentral $\chi^2$ distribution, as shown in Lemmas 1 and 2.

**Lemma 1:** Let $a \in \mathbb{R}^{n_a} \sim \mathcal{N}(\mu, \Sigma)$. Then, $a^T \Omega^{-1} a \sim \chi^2(n_a, \mu^T \Omega^{-1} \mu)$, where $n_a$ represents the degrees of freedom and $\mu^T \Omega^{-1} \mu$ is the noncentrality parameter.

**Proof:** Let $b \triangleq Y^{-1} a$, where $a \in \mathbb{R}^{n_a} \sim \mathcal{N}(\mu, \Omega)$ and $YY^T = \Omega$. Then, $b \sim \mathcal{N}(Y^{-1} \mu, I)$, and $a^T \Omega^{-1} a = b^T b = \sum_{i=1}^{n_a} b_i^2$ is distributed according to the noncentral $\chi^2$ distribution with $n_a$ degrees of freedom and a noncentrality parameter given by $\mu^T \Omega^{-1} \mu$.

**Lemma 2:** When the system is under attack

$$g(z_{k-T+1:k}) \sim \chi^2(mT, \sum_{i=k-T+1}^{k} \Delta z_i^T \Sigma^{-1} \Delta z_i)$$

(17)

$$\mathbb{E}[g(z_{k-T+1:k})] = mT + \sum_{i=k-T+1}^{k} \Delta z_i^T \Sigma^{-1} \Delta z_i$$

(18)

where $mT$ represents the degrees of freedom and $\sum_{i=k-T+1}^{k} \Delta z_i^T \Sigma^{-1} \Delta z_i$ is the noncentrality parameter.

**Proof:** When the system is under attack, $z_k \sim \mathcal{N}(\Delta z_k, \Sigma)$.

Then

$$z_k^T \Sigma^{-1} z_k \sim \chi^2(m, \Delta z_k^T \Sigma^{-1} \Delta z_k)$$

according to Lemma 1. Extending the result of Lemma 1 to

$$g(z_{k-T+1:k}) = \begin{bmatrix} z_k \\ \vdots \\ z_{k-T+1} \end{bmatrix}^T \begin{bmatrix} \Sigma & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma \end{bmatrix}^{-1} \begin{bmatrix} z_k \\ \vdots \\ z_{k-T+1} \end{bmatrix}$$

yields that $g(z_{k-T+1:k})$ is distributed according to the noncentral $\chi^2$ distribution with $mT$ degrees of freedom and a noncentrality parameter given by $\sum_{i=k-T+1}^{k} \Delta z_i^T \Sigma^{-1} \Delta z_i$. The expected value then follows directly from the properties of the noncentral $\chi^2$ distribution.

**D. Saturation Limits**

We assume the controller $u_k = L \hat{x}_{k|k}$ is designed so that the physical saturation limits of the actuators are rarely reached under normal operation. For example, this can be accomplished for LQG control using the techniques presented in [43]. The physical saturation limits are given by

$$U \triangleq \{ \bar{u}_k | \bar{u}_k U \bar{u}_k \leq 1 \}$$

(19)

where $U > 0$ and $\bar{u}_k$ represents the vector of applied control inputs, given by

$$\bar{u}_k \triangleq \begin{cases} u_k, & \text{normal operation} \\ u_k + u_k^a, & \text{attack} \end{cases}$$

(20)

Whenever saturation occurs under normal operation, it is equivalent to a control input bias that results in the control input lying on the boundary of the region $U$. In other words, it is equivalent to an attacker adding a bias of $u_k^a$ on the control inputs such that $\bar{u}_k U \bar{u}_k = 1$. In this way, control input saturation may be treated as a type of attack.

When the system is under attack, the saturation limits $U$ will constrain the set of possible control input biases an attacker can exert on the system so that $u_k^a$ will lie within the region $u_k^a \in U^a(\bar{x}_k)$, given by

$$U^a(\bar{x}_k) \triangleq \{ u_k^a | (u_k^a + \bar{L} \bar{x}_k)^T U (u_k^a + \bar{L} \bar{x}_k) \leq 1 \}$$

(21)

where $\bar{L} \triangleq [ L - L ]$.

**V. STEALTHY ADVERSARIES**

Since an attack can be mitigated if it is detected, we are interested in mitigating attacks that are not detected. More specifically, we would like to ensure resilience against stealthy adversaries, which we define as adversaries that can bypass the $\chi^2$ detector. This definition is expressed as follows.

**Definition 1:** A $(p_d, k')$-stealthy adversary designs $\Delta z_{0:k'}$ such that $p(g(z_{k-T+1:k}) > \eta | H_1) \leq p_d \forall k \in \{0, \ldots, k'\}$.

In other words, a $(p_d, k')$-stealthy adversary designs the biases on the residues such that the probability of detection is less than or equal to $p_d$ through time step $k'$.

To analyze the resilience of the system against stealthy adversaries, we first quantify the set of biases an adversary can exert on the residues that bypass the $\chi^2$ detector. Lemma 3 describes this set.

**Lemma 3:** The full set of biases $\Delta z_{0:k'}$ that are $(p_d, k')$-stealthy is given by

$$Z_{0:k'} \triangleq \bigcap_{k=0}^{k'} \left\{ \Delta z_{0:k'} | \sum_{i=k-T+1}^{k} \Delta z_i^T \Sigma^{-1} \Delta z_i \leq \lambda \right\}$$

(22)

where

$$\lambda \triangleq \max_{\lambda} \ s.t. \ Q_{\frac{1}{2} mT} \left( \sqrt{\lambda}, \sqrt{\eta} \right) \leq p_d$$

(23)

and $Q_{\frac{1}{2} mT}(\sqrt{\lambda}, \sqrt{\eta})$ is the generalized Marcum Q-function, given by

$$Q_{\frac{1}{2} mT}(\sqrt{\lambda}, \sqrt{\eta}) \triangleq 1 - e^{-\frac{1}{2} \lambda} \sum_{j=0}^{\infty} \left( \frac{1}{2} \lambda \right)^j \frac{\eta^{\frac{1}{2} j}}{j!} \int_{0}^{\infty} e^{-\frac{1}{2} t} t^{\frac{1}{2} mT + j - 1} e^{-\frac{1}{2} t} dt$$

(24)

**Proof:** Lemma 2 states that when the system is under attack, the detection statistic follows a noncentral $\chi^2$ distribution with a noncentrality parameter that is a quadratic function of the
set of biases $\Delta z_{k-T+1:k}$. Consequently, the full set of biases that are $(p_d, k')$-stealthy is given by maximizing the noncentrality parameter subject to the conditions of $(p_d, k')$-stealthiness provided in Definition 1. These conditions are set forth in the constraint of (23) and the intersection of the sets in (22), where the relationship between the generalized Marcum Q-function and the cumulative distribution function of the noncentral $\chi^2$ distribution is given by [44]

$$Q_{1/mT}(\sqrt{\lambda}, \sqrt{\eta}) = p (g(z_{k-T+1:k}) > \eta | H_1).$$

Letting $\lambda$ represent the noncentrality parameter, (23) maximizes this parameter and uses this maximum value to define the set in (22).

Lemmas 4 and 5 show that the optimization problem in (23) can be solved with a unique solution for $\lambda$, allowing us to quantify the full set of $(p_d, k')$-stealthy biases $Z_{0:k'}$.

**Lemma 4 (see [45]):** The generalized Marcum Q-function $Q_{1/mT}(\sqrt{\lambda}, \sqrt{\eta})$ is strictly increasing in $\frac{1}{mT}$ and $\sqrt{\lambda}$ for all $\sqrt{\lambda} \geq 0, \sqrt{\eta} > 0$, and $\frac{1}{mT} > 0$. It is strictly decreasing in $\sqrt{\eta}$ for all $\sqrt{\lambda} \geq 0, \sqrt{\eta} \geq 0$, and $\frac{1}{mT} > 0$.

**Lemma 5:** The solution to the optimization problem in (23) is

$$\lambda \triangleq \lambda \text{ s.t. } Q_{1/mT}(\sqrt{\lambda}, \sqrt{\eta}) = p_d. \quad (25)$$

**Proof:** This follows directly from the fact that according to Lemma 4 the constraint of (23), given by the generalized Marcum Q-function, is strictly increasing in $\lambda$ for all $\lambda \geq 0$, $\eta > 0$, and $mT > 0$.

In other words, Lemma 5 allows us to compute an exact value for $\lambda$ by successively increasing $\lambda$ until $Q_{1/mT}(\sqrt{\lambda}, \sqrt{\eta}) = p_d$.

**Corollary 1:** If an adversary designs $Z_{0:k'}$, more explicitly showing the relationship between the size of the stealthy bias set and the parameters $p_d, \eta$, and $T$.

Letting $Z_{0:k'} = \cap_{k=0}^{k'} \{ \Delta z_{0:k'} | \sum_{i=k-T+1}^{k} \Delta z_i^T \Sigma^{-1} \Delta z_i \leq \eta p_d - mT \}$.

then the adversary will remain $(p_d, k')$-stealthy.

**Proof:** The proof is provided in the Appendix.

**Note:** Note that the set $Z_{0:k'}$ grows as $p_d$ increases, as $\eta$ increases, and as $T$ decreases. Since $Z_{0:k'} \subseteq Z_{n:k'}$, the set $Z_{0:k'}$ will also grow as $p_d$ increases, as $\eta$ increases, and as $T$ decreases.

This knowledge can be used when designing the detector since choosing values of the threshold $\eta$ that are large or values of the time window $T$ that are small will result in an adversary being able to use a large set of stealthy biases to perturb the system.

Having now exactly quantified the set of $(p_d, k')$-stealthy biases $Z_{0:k'}$, Lemma 6 projects this set onto $\Delta z_k$ to decouple the correlation of the biases over time, in turn overapproximating the set of $(p_d, k')$-stealthy biases.

**Lemma 6:** If an adversary remains $(p_d, k')$-stealthy, then $\forall k \in \{0, \ldots, k'\}, \Delta z_k$ will lie within the region $\Delta z_k \in \mathbb{Z}$, given by

$$Z \triangleq \{ \Delta z_k | \Delta z_k^T \Sigma^{-1} \Delta z_k \leq \lambda \}. \quad (27)$$

**Proof:** Follows directly from projecting the set $Z_{0:k'}$ onto $\Delta z_k$. Projecting $\{ \Delta z_{0:k'} | \sum_{j=z-T+1}^{z} \Delta z_i^T \Sigma^{-1} \Delta z_i \leq \lambda \}$ onto $\Delta z_k$ yields $\Delta z_k \in Z \forall j = \{k, \ldots, k' \}$ and $\Delta z_k \in \mathbb{R}^m$ otherwise. Taking the intersection of all these projections over $j = \{0, \ldots, k'\}$ yields $Z$.

**Remark 2:** Note that in the case of a covert attack such as the one described in (3), there will be no bias on the residue so $\Delta z_k = 0$. Consequently, covert attack biases always belong to the set $Z$.

**Remark 3:** Note that some conservatism in this article is introduced by using the set $Z$ of stealthy biases, a projection of the set $Z_{0:k'}$ onto time step $k$. The amount of conservativeness is directly dependent upon $T$, the time window of the detector. To illustrate this, let the detector window be $T = 1$. In this case, $Z_{0:k} = \cap_{k=0}^{k'} \{ \Delta z_{0:k'} | \Delta z_k^T \Sigma^{-1} \Delta z_k \leq \lambda \}$, so no conservatism is introduced when projecting this set onto time step $k$ (regardless of how large $k$ is). As the detector window $T$ is increased, more conservativeness is introduced in the projection because more dependency is introduced between the sets being intersected in the definition of $Z_{0:k'}$. Since the time window for the detector is usually chosen to be small, the amount of conservativeness introduced in Lemma 6 will usually be quite small.

**Remark 4:** Note that Lemma 6 states that for a $(p_d, k')$-stealthy adversary, $\Delta z_k \in Z \forall k \in \{0, \ldots, k'\}$. Since $Z$ is time-invariant, $\Delta z_k$ will still lie within this region as the attacker’s stealthy time horizon $k' \to \infty$. Consequently, the region $Z$ is an overapproximation of the set of biases that are $(p_d)$-stealthy, where $p_d$-stealthiness is defined as follows.

**Definition 2:** A $p_d$-stealthy adversary designs $\Delta z_{0:k'}$ such that $p (g(z_{k-T+1:k}) > \eta | H_1) \leq p_d \forall k$.

In other words, a $p_d$-stealthy adversary designs the biases on the residues such that the probability of detection is less than or equal to $p_d$ for all time.

**VI. SAFETY**

We now investigate the amount of time the overall system can be under stealthy attack and still remain safe with a particular probability. For ease of analysis, we divide up the dynamics in (15) according to the bounded uncertainties $\{ \Delta z_{k+1}, u^a_k \}$ and the stochastic uncertainties $\{ w_k, z^a_{k+1} \}$. Letting $\bar{x}^1_k$ and $\bar{x}^2_k$ be defined such that $\bar{x}^1_k + \bar{x}^2_k \triangleq \bar{x}_k$, the overall dynamics in (15) can be represented as

$$\bar{x}^1_k = \begin{bmatrix} A \bar{x}^1_k + u^a_k & 1 \end{bmatrix} \quad \text{normal operation}$$

$$\bar{x}^2_k = A \bar{x}^2_k + \begin{bmatrix} I & \begin{bmatrix} K & \begin{bmatrix} u^a_k \\ z^a_{k+1} \\ w_k \\ \end{bmatrix} \end{bmatrix} \end{bmatrix} \quad \text{attack}$$

(28) (29)

In the next two sections, we provide sufficient conditions for ensuring that $\bar{x}^1_k$ and $\bar{x}^2_k$, respectively, remain within particular ellipsoidal sets when under normal operation and attack.
A. Bounded Uncertainties

Lemmas 7 and 8 quantify how quickly the Lyapunov function $\bar{x}_k^T \bar{P}_1 \bar{x}_k^1$ associated with the bounded uncertainties system in (28) either decreases or increases when the system is under normal operation or attack.

**Lemma 7:** For the system in (28) under normal operation

\[
\gamma \bar{P}_1 - A^T \bar{P}_1 A \succeq 0 \iff \bar{x}_{k+1}^T \bar{P}_1 \bar{x}_{k+1}^1 \leq \gamma \bar{x}_k^T \bar{P}_1 \bar{x}_k^1 \tag{30}
\]

where $\gamma \geq 0$ and $\bar{P}_1$ is a symmetric positive-definite Lyapunov matrix.

**Proof:**

\[
\gamma \bar{P}_1 - A^T \bar{P}_1 A \succeq 0 \iff \bar{x}_k^T A^T \bar{P}_1 A \bar{x}_k^1 \leq \gamma \bar{x}_k^T \bar{P}_1 \bar{x}_k^1
\]

\[
\iff \bar{x}_{k+1}^T \bar{P}_1 \bar{x}_{k+1}^1 \leq \gamma \bar{x}_k^T \bar{P}_1 \bar{x}_k^1.
\]

**Lemma 8:** If $\exists \alpha_1 \geq 0$ such that

\[
\begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{T2} & \Gamma_{22} \end{bmatrix} \succeq 0
\tag{31}
\]

where

\[
\Gamma_{11} \triangleq \begin{bmatrix} (\gamma_a - 2\alpha_1) \bar{P}_1 - A^T \bar{P}_1 A + \alpha_1 \bar{L}^T \bar{L} \bar{U} \bar{L} & \alpha_1 \bar{L}^T \bar{U} \\ \alpha_1 \bar{L}^T \bar{L} & \alpha_1 \bar{L}^T \bar{U} \end{bmatrix}
\]

\[
\Gamma_{12} \triangleq \begin{bmatrix} -A^T \bar{P}_1 \bar{K} & \alpha_1 \bar{L}^T \bar{U} - A^T \bar{P}_1 \bar{B} \\ 0 & \alpha_1 \bar{L}^T \bar{U} \end{bmatrix}
\]

\[
\Gamma_{22} \triangleq \begin{bmatrix} \frac{\bar{K} \Sigma^{-1} - \bar{K} \bar{P}_1 \bar{K}}{\alpha_1 \bar{L}^T \bar{U}} & -\bar{K} \bar{P}_1 \bar{B} \\ -\bar{B} \bar{P}_1 \bar{K} & \alpha_1 \bar{U} - \bar{B}^T \bar{P}_1 \bar{B} \end{bmatrix}
\]

then for system in (28) when under attack, where $\gamma_a \geq 0$ and $\bar{P}_1$ is a symmetric positive-definite Lyapunov matrix.

**Proof:** By using the S-procedure [46], (31) is equivalent to

\[
\xi_k \triangleq \begin{bmatrix} -2 \bar{P}_1 + \bar{L} \bar{L}^T \bar{U} & \bar{L} \bar{L}^T \bar{U} & 0 & \bar{L} \bar{U}^T \\ \bar{L} \bar{L}^T \bar{U} & \bar{L} \bar{L}^T \bar{U} & 0 & \bar{U} \bar{L} \\ 0 & 0 & \frac{1}{\alpha_1} \Sigma^{-1} & 0 \\ \bar{U} \bar{L} & \bar{U} \bar{L} & 0 & \bar{U} \end{bmatrix} \xi_k \leq 0
\]

\[
\Rightarrow \xi_k \triangleq \begin{bmatrix} A^T \bar{P}_1 A - \gamma_a \bar{P}_1 & A^T \bar{P}_1 \bar{K} & A^T \bar{P}_1 \bar{B} \\ 0 & 0 & 0 & 0 \\ \bar{K}^T \bar{P}_1 \bar{A} & \bar{K}^T \bar{P}_1 \bar{K} & \bar{K}^T \bar{P}_1 \bar{B} \\ \bar{B}^T \bar{P}_1 \bar{A} & \bar{B}^T \bar{P}_1 \bar{K} & \bar{B}^T \bar{P}_1 \bar{B} \end{bmatrix} \xi_k \leq 0
\]

where $\xi_k \triangleq \begin{bmatrix} \bar{x}_k^T \bar{x}_k^1 & \bar{x}_k^T \Delta z_{k+1} & u_k^T \end{bmatrix}^T$. This, in turn, is equivalent to

\[
-2 \bar{x}_k^T \bar{P}_1 \bar{x}_k^1 + \Delta z_{k+1}^T \frac{1}{\alpha_1} \Sigma^{-1} \Delta z_{k+1} + (u_k^a + \bar{L} \bar{x}_k^1 + \bar{L} \bar{x}_k^2)^T(U(u_k^a + \bar{L} \bar{x}_k^1 + \bar{L} \bar{x}_k^2) \leq 0 \tag{33}
\]

\[
\Rightarrow \bar{x}_{k+1}^T \bar{P}_1 \bar{x}_{k+1}^1 \leq \gamma_a \bar{x}_k^T \bar{P}_1 \bar{x}_k^1
\]

for system in (28) when under attack. Note that

\[
\left\{ \begin{array}{l}
\frac{\bar{x}_k^T \bar{P}_1 \bar{x}_k^1}{\Delta z_{k+1}^T \frac{1}{\alpha_1} \Sigma^{-1} \Delta z_{k+1}} \geq \lambda \\
(u_k^a + \bar{L} \bar{x}_k^1 + \bar{L} \bar{x}_k^2)^T(U(u_k^a + \bar{L} \bar{x}_k^1 + \bar{L} \bar{x}_k^2) \leq 1 \end{array} \right. \tag{34}
\]

Taking (33) and (34) in conjunction with one another yields that

\[
\forall u_k^a \in \mathcal{U}(\bar{x}_k) \text{ and } \forall \Delta z_{k+1} \in \mathcal{Z}
\]

\[
\bar{x}_k^T \bar{P}_1 \bar{x}_k^1 \geq 1 \Rightarrow \bar{x}_{k+1}^T \bar{P}_1 \bar{x}_{k+1}^1 \leq \gamma_a \bar{x}_k^T \bar{P}_1 \bar{x}_k^1
\]

for system in (28) when under attack.

**Lemma 8** provides a sufficient condition guaranteeing that the Lyapunov function associated with the bounded uncertainties system in (28) will decrease at a rate of at least $\gamma_a$ if $\gamma_a \geq 0$, or will increase at a rate of at most $\gamma_a$ if $\gamma_a \geq 1$ when the system is under $p_d$-stealthy attack, where the adversary exerts biases $u_k^a$ and $\Delta z_{k+1}$ within the saturation and stealthiness limits of $\mathcal{U}(\bar{x}_k)$ and $\mathcal{Z}$, respectively.

Having now quantified the rate at which the Lyapunov function decreases or increases for the system in (28), Theorem 1 quantifies the size of the ellipsoid in which $\bar{x}_k^1$ is guaranteed to lie. The size of this ellipsoid is a function of the Lyapunov rate parameters $\gamma_a$ and $\gamma_b$ as well as the amount of time the system has been vulnerable to attacks.

**Theorem 1:** If $\gamma \bar{P}_1 - A^T \bar{P}_1 A \succeq 0$, if $\exists \alpha_1 \geq 0$ such that (31) is satisfied, and if $\bar{x}_k^1 \in \mathcal{E}_1(0, 0)$, then

\[
\bar{x}_k^1 \in \mathcal{E}_1(k, T_a) \triangleq \left\{ \bar{x}_k^1 | \bar{x}_k^T \bar{P}_1 \bar{x}_k^1 \leq \max(1, \gamma^k \gamma_{1+a} T_a) \right\}
\tag{35}
\]

where $T_a$ represents the total number of time steps the system has been under $p_d$-stealthy attack.

**Proof:** If $\gamma \bar{P}_1 - A^T \bar{P}_1 A \geq 0$ and if $\exists \alpha_1 \geq 0$ such that (31) is satisfied, then it follows directly from Lemmas 7 and 8 that

\[
\bar{x}_{k+1}^T \bar{P}_1 \bar{x}_{k+1}^1 \leq \left\{ \begin{array}{ll}
\gamma \bar{x}_k^T \bar{P}_1 \bar{x}_k^1, & \text{normal operation} \\
\gamma_a \bar{x}_k^T \bar{P}_1 \bar{x}_k^1, & \text{attack when } \bar{x}_k^T \bar{P}_1 \bar{x}_k^1 \geq 1
\end{array} \right.
\]

for the system in (28). Consequently

\[
\bar{x}_k^T \bar{P}_1 \bar{x}_k^1 \leq \max(1, \gamma^k \gamma_{1+a} T_a)
\]

If $\bar{x}_0^1 \in \mathcal{E}_1(0, 0)$, then max $\bar{x}_{k+1}^T \bar{P}_1 \bar{x}_{k+1}^1 = 1$, implying that

\[
\bar{x}_k^T \bar{P}_1 \bar{x}_k^1 \leq \max(1, \gamma^k \gamma_{1+a} T_a)
\]

Theorem 1 shows that when $\gamma \in [0, 1)$ and $\gamma_a \geq 1$, the ellipsoid $\mathcal{E}_1(k, T_a)$ grows by a factor of $\gamma_a$ each time step the system is under attack and shrinks by a factor of $\gamma$ each time step the
system is not under attack. Corollary 2 describes the volume of this ellipsoid with respect to $\gamma, \gamma_a, k,$ and $T_a$.

**Corollary 2:** When $\gamma^{k-T_a} \leq 1$, the volume of $E_1(k, T_a)$
1. grows logarithmically with $\gamma$ and $\gamma_a$;
2. is negatively proportional to $k$ if $\gamma \in (0, 1)$;
3. is proportional to $T_a$ if $\gamma_a > \gamma$.
When $\gamma^{k-T_a} \geq 1$, the volume of $E_1(k, T_a)$ is not dependent on $\gamma, \gamma_a, k,$ or $T_a$.

**Proof:** The volume of $E_1(k, T_a)$ is proportional to

\[ -\log \det \left( \frac{1}{\max(1, \gamma^{k-T_a} \gamma_a)} P_1 \right) \]

\[ = -\log \left( \frac{1}{\max(1, \gamma^{k-T_a} \gamma_a) + 2n} \det(P_1) \right) \]

\[ = 2n \left( (k - T_a) \log(\gamma) + T_a \log(\gamma_a) \right) - \log \det(P_1) \]

\[ = \begin{cases} 
2n \left( (k - T_a) \log(\gamma) + T_a \log(\gamma_a) \right) - \log \det(P_1), & \text{if } \gamma^{k-T_a} \gamma_a > 1 \\
- \log \det(P_1), & \text{if } \gamma^{k-T_a} \gamma_a \leq 1 
\end{cases} \]

\[ = \begin{cases} 
2n \left( (k - T_a) \log(\gamma) + T_a \log(\gamma_a) \right) - \log \det(P_1), & \text{if } \gamma^{k-T_a} \gamma_a > 1 \\
- \log \det(P_1), & \text{if } \gamma^{k-T_a} \gamma_a \leq 1 
\end{cases} \]

**B. Stochastic Uncertainties**

Having quantified the size of the ellipsoid in which $x_k^1$ lies, we now quantify the ellipsoid in which $x_k^2$ lies with at least probability $p$ for the stochastic uncertainties system in (29). To do so, we leverage the notions of quadratic boundedness, robust positive invariance, and probabilistic positive invariance, which are described in Definitions 3–5. These definitions, along with Lemmas 9 and 10, have been modified and adapted from the work in [47] and [48] in order to arrive at the result presented in Theorem 2.

**Definition 3 (see [47]):** Let $s_k \in \mathbb{R}^{n_x}$ and $d_k \in \mathbb{R}^{n_d}$ represent state and disturbance vectors, respectively, and let $D$ be a compact set. A system of the form

\[ s_{k+1} = \bar{A}s_k + \bar{B}d_k \]

is quadratically bounded with symmetric positive-definite Lyapunov matrix $P$ if and only if

\[ s_k^T P s_k \geq 1 \Rightarrow s_{k+1}^T P s_{k+1} \leq s_k^T P s_k \quad \forall d_k \in D. \]  

**Definition 4:** (see [47]) The set $\mathcal{S}$ is a robustly positively invariant set for (36) if and only if $s_{k+1} \in \mathcal{S} \forall s_k \in \mathcal{S} \forall d_k \in D$.

**Definition 5:** (see [48]) Let $d_k \sim \mathcal{Q}^d$ indicate that $d_k$ is a random variable of distribution $\mathcal{Q}^d$. The set $\mathcal{S}$ is a probabilistically positively invariant set of probability level $p$ for (36) with $d_k \sim \mathcal{Q}^d$ if and only if $s_0 \in \mathcal{S}$ implies that $Pr(s_{k+1} \in \mathcal{S}) \geq p \forall k \geq 0$.

Given these definitions, Lemma 9 provides a condition that is equivalent to robust positive invariance. Lemma 10 shows the correspondence between robust positive invariance and probabilistic positive invariance, and Lemma 11 combines these results to provide a sufficient condition for probabilistic positive invariance.

**Lemma 9:** (see [47]) Let $d_k \in \mathcal{D} = \{d_k | d_k^T P^{-1} d_k \leq 1, k \geq 0 \}$. $\mathcal{S} \triangleq \{s_k | s_k^T P s_k \leq 1\}$ is a robustly positively invariant set for (36) and the system in (36) is quadratically bounded with symmetric positive-definite Lyapunov matrix $P$ if and only if $\exists \alpha \geq 0$ such that

\[ \left[ (\alpha - 1)P + \bar{A}^T P \bar{A} \quad \bar{A}^T P \bar{B} \right] \]

\[ \left[ \begin{array}{c} B^T P \bar{A} \\ B^T P \bar{B} \end{array} \right] \leq 0. \]

**Lemma 10:** (see [48]) If $\mathcal{S}$ is a robustly positively invariant set for (36) with $d_k \in \mathcal{D} = \{d_k | d_k^T P^{-1} d_k \leq 1 \}$, then $\mathcal{S}$ is also a probabilistically positively invariant set of probability level $p$ for (36) with $d_k \sim \mathcal{Q}^d$, $E[d_k] = 0$, and $\text{Cov}(d_k) = D$. If $\exists \alpha \geq 0, P > 0$ such that

\[ \left[ (\alpha - 1)P + \bar{A}^T P \bar{A} \quad \bar{A}^T P \bar{B} \right] \]

\[ \left[ \begin{array}{c} B^T P \bar{A} \\ B^T P \bar{B} \end{array} \right] \leq 0 \]

then $\mathcal{S} \triangleq \{s_k | s_k^T P s_k \leq 1\}$ is a probabilistically positively invariant set of probability level $p$ for (36).

**Proof:** Follows directly from Lemmas 9 and 10.

Given these Lemmas, Theorem 2 sets forth the ellipsoid in which $x_k^2$ lies with at least probability $p$ for the system in (29), and Corollary 3 describes the volume of this ellipsoid with respect to $p$.

**Theorem 2:** Let $R \triangleq \text{BlkDiag}(Q, \Sigma)$ so that $x_k \in \mathbb{R}^{n+m} \sim \mathcal{N}(0, R)$. If $\exists \alpha \geq 0, P > 0$ such that

\[ \left[ (\alpha - 1)P + A^T P A \quad A^T P \bar{K} \right] \]

\[ \left[ \begin{array}{c} \bar{K}^T P A \\ K^T P A \end{array} \right] \leq 0 \]

then $E_2(p) \triangleq \{x_k^2 | x_k^2 P x_k^2 \leq 1/(1 - p)\}$ is a probabilistically positively invariant set of probability level $p$ for (29).

**Proof:** Applying Lemma 11 to (29) yields that if $\exists \alpha \geq 0, P > 0$ such that

\[ \left[ (\alpha - 1)P + A^T P A \quad A^T P \bar{K} \right] \]

\[ \left[ \begin{array}{c} \bar{K}^T P A \\ K^T P A \end{array} \right] \leq 0 \]

then $E_2(p) = \{x_k^2 | x_k^2 P x_k^2 \leq 1\}$ is a probabilistically positively invariant set of probability level $p$ for (29). Letting $P \triangleq (1 - p)P_2$ yields the desired result.

**Corollary 3:** The volume of $E_2(p)$ grows logarithmically with $p$.

**Proof:** The volume of $E_2(p)$ is proportional to

\[ -\log \det((1 - p)P_2) = -\log((1 - p)2n \det(P_2)) \]

\[ = -2n \log(1 - p) - \log \det(P_2). \]
C. Safety Set

We have now quantified the ellipsoid $E_1(k, T_a)$ in which $\bar{x}_k^1$ is guaranteed to lie as well as the ellipsoid $E_2(p)$ in which $\bar{x}_k^2$ lies with at least probability $p$. Theorem 3 uses these results to quantify the region within which the overall state $\bar{x}_k$ lies, and this region can be used to determine whether or not the safety constraints have been violated.

**Theorem 3:** If $\bar{x}_0 \in \{ E_1(0,0) \oplus E_2(p) \}$ and if $\exists \alpha_1, \alpha_2, \gamma, \gamma_a \geq 0$ and $P_1, P_2 \succ 0$ such that $\gamma P_1 - A^T P_1 A \succeq 0$, (31), and (41) are satisfied, then

$$\Pr(\bar{x}_k \in \{ E_1(k, T_a) \oplus E_2(p) \}) \geq p \quad \forall k \geq 0$$

(43)

when a $p_d$-stealthy adversary attacks the system for a total of $T_a$ time steps.

**Proof:**

If $\bar{x}_0 \in \{ E_1(0,0) \oplus E_2(p) \}$ and if $\exists \alpha_1, \alpha_2, \gamma, \gamma_a \geq 0$ and $P_1, P_2 \succ 0$ such that $\gamma P_1 - A^T P_1 A \succeq 0$, (31), and (41) are satisfied, then it follows directly from Theorems 1 and 2 that

$$\bar{x}_k^1 \in E_1(k, T_a), \quad \Pr(\bar{x}_k^2 \in E_2(p)) \geq p \quad \forall k \geq 0.$$

Since $\bar{x}_k = \bar{x}_k^1 + \bar{x}_k^2$, $\bar{x}_k \in \{ E_1(k, T_a) \oplus E_2(p) \}$ with probability greater than or equal to $p \forall k \geq 0$.

Given a set of safe state constraints, Theorem 3 can be used to provide a conservative estimate of how long the system can remain under $p_d$-stealthy attack before the safe state constraints are violated. It provides a tool to analyze the response mechanism being used, providing the probability of safety given the number of time steps the system has been vulnerable to attack.

In executing this analysis, the variables $\gamma$, $\gamma_a$, $P_1$, and $P_2$ are all parameters that can be chosen by the system operator as long as $\gamma P_1 - A^T P_1 A \succeq 0$, (31), and (41) are satisfied. The meanings of these parameters are presented as follows.

1) $\gamma$: The rate at which the Lyapunov function decreases ($\gamma \in [0,1]$) or increases ($\gamma \geq 1$) under normal operation for the system in (28).

2) $\gamma_a$: The rate at which the Lyapunov function decreases ($\gamma_a \in [0,1]$) or increases ($\gamma_a \geq 1$) under attack for the system in (28).

3) $P_1$: The shape matrix for $E_1(k, T_a)$.

4) $P_2$: The shape matrix for $E_2(p)$.

To minimize the rate parameters $\gamma$ and $\gamma_a$ as well as minimize the volume of the shape matrix for the ellipsoid $E_1(k, T_a)$, the parameters $\gamma$, $\gamma_a$, and $P_1$ can be designed according to the following optimization problem:

$$\arg \min_{\omega_1, \omega_2, \omega_3} \omega_1 \gamma + \omega_2 \gamma_a - \omega_3 \log \det P_1 \quad \text{s.t.} \quad P_1 \succ 0$$

$$\alpha_1, \alpha_2, \gamma, \gamma_a \geq 0, \quad \gamma P_1 - A^T P_1 A \succeq 0, \quad (31) \text{ is satisfied}$$

(44)

where $\omega_1$, $\omega_2$, and $\omega_3$ are nonnegative constants weighting the rate at which the Lyapunov function for (28) decreases/increases under normal operation, the rate at which the Lyapunov function for (28) decreases/increases under attack, and the volume of the shape matrix for $E_1(k, T_a)$, respectively. To minimize the volume of the ellipsoid $E_2(p)$, the parameter $P_2$ can be designed according to the following optimization problem:

$$\arg \max_{\alpha_2, \alpha_3} \log \det P_2 \quad \text{s.t.} \quad P_2 \succ 0, \quad \alpha_2 \geq 0, \quad (41) \text{ is satisfied}.$$  

(45)

VII. Example

To demonstrate the effectiveness of this analysis for understanding resilience against stealthy attacks, we consider the quadruple tank process, a multivariable laboratory process that consists of four interconnected water tanks [49]. This process is an example of a CPS that can be subject to stealthy attacks where an adversary corrupts the inputs and outputs through a memory corruption attack, a man-in-the-middle attack, or an attack on the physical actuators and sensors. The objective of the quadruple tank process is to control the water level of the first two tanks using two pumps. The system has four states (water level for each tank), two inputs (voltages applied to the pumps), and two outputs (voltages from level measurement devices for the first two tanks).

We use the system model in [49] at the minimum-phase operating point with a 1.5 cm² cross-section for each tank’s outlet hole. We discretize the system with a sampling rate of 2 s and use an LQG controller with weights $W = I$ and $V = 100I$. To ensure an appropriate noise magnitude, $Q$ and $R$ are created by generating a matrix from a uniform distribution, multiplying it by its transpose, and dividing by 100. The saturation limits for the actuators are given by $0 \leq \bar{v}_i \leq 6$ so that $U = \text{BlkDiag}(1/9,1/9)$ when $\bar{v}_i = 2$ V, where $\bar{v}_i$ represents the voltage applied to pump $i$ and $\bar{v}_i^0$ represents the operating point of pump $i$. A window size of $T = 10$ is used for the $\chi^2$ detector, and the threshold $\eta$ is chosen so that the desired false alarm rate is $1\%$. This system enters an unsafe region of operation when the tanks overflow, so the set of safe states is given by $h_i \leq 30$ cm so that $h_i - h_i^0 \leq 30 - h_i^0$, where $h_i$ represents the water level of tank $i$ and $h_i^0$ represents the operating point of tank $i$.

We solve for the maximum noncentrality parameter $\tilde{\lambda}$ according to (25) for a $p_d$-stealthy adversary with $p_d = 99\%$. Since the optimization problem in (44) is not convex, we design the parameters $\gamma$, $\gamma_a$, and $P_1$ in a suboptimal manner by first finding a $P_1$ that satisfies $\gamma P_1 - A^T P_1 A \succeq 0$ and then finding the minimal nonnegative value for $\gamma$ such that $\gamma P_1 - A^T P_1 A \succeq 0$ is satisfied. After that, we solve for $\alpha_1$ and $\gamma_a$ by minimizing $\gamma_a$ subject to (31). Since minimizing $E_2(p)$ jointly over $\alpha_2$ and $P_2$ in (45) is not convex, we find a suboptimal solution for $\alpha_2$ and $P_2$ by restricting the possible values of $\alpha_2$ to a finite set and maximizing $\max_{P_2} \log \det P_2$ subject to (41) over that finite set.

Figs. 3–5 depict the volume of $E_1(k, T_a)$, the ellipsoid in which $\bar{x}_k^1$ lies. As stated in Corollary 2, Fig. 3 shows that the volume of $E_1(k, T_a)$ is proportional to $T_a$, growing linearly with the total amount of time the system has been under $p_d$-stealthy attack. Fig. 4 also confirms the results of Corollary 2, showing that the volume of $E_1(k, T_a)$ is negatively proportional to $k$, decreasing linearly with the total amount of time the system has been under normal operation. Fig. 5 similarly shows how the volume of $E_1(k, T_a)$ grows linearly with the percentage of time the system is under $p_d$-stealthy attack. In Figs. 3–5, it can be seen that there is a minimal volume for $E_1(k, T_a)$, which is due to the appearance of the term $\max(1, \gamma^{1-k^2} T_a^{\gamma^2})$ in the definition of $E_1(k, T_a)$ in (35).
Fig. 3. Volume of $E_1(k, T_a)$ as a function of the total number of time steps $T_a$ under $p_d$-stealthy attack.

Fig. 4. Volume of $E_1(k, T_a)$ as a function of the total number of time steps $k$.

Fig. 5. Volume of $E_1(k, T_a)$ as a function of the percentage of time under $p_d$-stealthy attack.

Fig. 6. Volume of $E_2(p)$ as a function of the probability level $p$.

Fig. 7 and 8 leverage the safety set presented in Theorem 3 to indicate whether or not specific response mechanisms will result in maintaining resilience against stealthy adversaries. Fig. 7 shows an upper bound on the total number of time steps $T_a$ under $p_d$-stealthy attack that the system can remain safe with at least 99% probability. This number of time steps is proportional to the total number of time steps, indicating that the system can handle being under $p_d$-stealthy attack for more time when it has been under normal operation for a larger amount of time. Let $\tau_a^k$ represent the total number of time steps the response mechanism has allowed the system to be vulnerable to attacks at time step $k$. Any response mechanism that ensures $\tau_a^k$ lies below this upper bound will guarantee that the system remains safe against $p_d$-stealthy attacks with at least 99% probability. Fig. 7 compares $\tau_a^k$ for two response mechanisms to this upper bound. Response mechanism 1 ensures that every four time steps, the system is not under attack for the first two time steps but then is vulnerable to attacks for the next two time steps. This may be achieved, for example, through software rejuvenation [26]. Response mechanism 2 ensures that every 10 time steps, the system is not under attack for the first 2 time steps but then is vulnerable to attacks for the next 8 time steps. As seen in Fig. 7, $\tau_a^k$ for response mechanism 1 lies below an upper bound for safety, guaranteeing that the system will remain safe against $p_d$-stealthy adversaries with at least 99% probability. However, safety cannot be guaranteed when response mechanism 2 is implemented since $\tau_a^k$ lies above an upper bound for safety. The exact same results are shown in Fig. 8 except that the $y$-axis depicts the percentage of time under $p_d$-stealthy attack instead of the total number of time steps $T_a$ under $p_d$-stealthy attack.
We simulated the quadruple tank process over 1000 trials for a period of 100 time steps, allowing the system to be under $p_d$-stealthy attack for the maximum number of time steps according to the upper bound in Fig. 7. Over these 1000 trials, the tanks never overflowed, confirming that the upper bound in Fig. 7 ensures that the system remains safe against stealthy adversaries with at least 99% probability. Note that these results are a bit conservative since safety violations are less frequent in practice than the upper bound in Fig. 7 would indicate. One area of future work includes investigating this conservativeness and finding less conservative upper bounds.

**VIII. CONCLUSION**

In this article, we have shown that when analyzing attacks on CPSs and designing associated response countermeasures, particular attention must be given to how long the system can remain under attack and still remain safe. We show how a detector limits the set of stealthy biases an adversary can exert on the system, and we use this fact to produce an upper bound on the amount of time the system can be under stealthy attack without violating the safety constraints. This timing analysis can then be used in the design of the response mechanism to thwart adversaries. We illustrate our results with the example of the quadruple tanks process. While this work demonstrates that response mechanisms can be developed to prevent stealthy attacks from violating the safety constraints, future work includes ensuring resilience against attacks that are not stealthy, particularly when there may be missed detections as well as a nonnegligible time to detection that causes a delay in initiating the response mechanism. Furthermore, future work includes leveraging online information from the sensor measurements to produce a less conservative upper bound on the amount of time that the system can remain resilient against stealthy attacks.

**APPENDIX**

We first introduce Lemma 12, which is used in the proof of Corollary 1.

Lemma 12: The probability of detection is upper bounded according to

$$p \left( g(z_{k-T+1:k}) \geq \eta \mid \mathcal{H}_1 \right) \leq \frac{1}{\eta} \left( mT + \sum_{i=k-T+1}^{k} \Delta z_i^T \Sigma^{-1} \Delta z_i \right).$$

(46)

Proof: According to the Markov inequality and the expected value of the noncentral $\chi^2$ distribution given in (18)

$$p \left( g(z_{k-T+1:k}) \geq \eta \mid \mathcal{H}_1 \right) \leq \mathbb{E}[g(z_{k-T+1:k}) \mid \mathcal{H}_1] / \eta$$

$$= \frac{1}{\eta} \left( mT + \sum_{i=k-T+1}^{k} \Delta z_i^T \Sigma^{-1} \Delta z_i \right).$$

Proof of Corollary 1

Proof:

$$\bigcap_{k=0}^{k'} \left\{ \Delta z_{0:k} \Bigg| \sum_{i=k-T+1}^{k} \Delta z_i^T \Sigma^{-1} \Delta z_i \leq \eta p_d - mT \right\} \subseteq \bigcap_{k=0}^{k'} \left\{ \Delta z_{0:k} \Bigg| p_d \geq \frac{1}{\eta} \left( mT + \sum_{i=k-T+1}^{k} \Delta z_i^T \Sigma^{-1} \Delta z_i \right) \right\} \oplus \left\{ \Delta z_{0:k} \Bigg| p \left( g(z_{k-T+1:k}) \geq \eta \mid \mathcal{H}_1 \right) \leq p_d \wedge \right. \left. p_d < \frac{1}{\eta} \left( mT + \sum_{i=k-T+1}^{k} \Delta z_i^T \Sigma^{-1} \Delta z_i \right) \right\}$$

$$= \bigcap_{k=0}^{k'} \left\{ \Delta z_{0:k} \Bigg| p_d \geq \frac{1}{\eta} \left( mT + \sum_{i=k-T+1}^{k} \Delta z_i^T \Sigma^{-1} \Delta z_i \right) \wedge \right.$$

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
\[ p\left(g(x_{k-T+1:k}) \geq \eta \right| H_1) \leq \frac{1}{\eta} \left( mT + \sum_{i=k-T+1}^{k} \Delta z_i^T \Sigma^{-1} \Delta z_i \right) \]
\[ + \left\{ \Delta z_{0:k} \mid p\left(g(x_{k-T+1:k}) \geq \eta \right| H_1) \leq p_d \wedge \right. \]
\[ p_d < \frac{1}{\eta} \left( mT + \sum_{i=k-T+1}^{k} \Delta z_i^T \Sigma^{-1} \Delta z_i \right) \]\n\[ = \sum_{k=0}^{k'} \left\{ \Delta z_{0:k} \mid p\left(g(x_{k-T+1:k}) \geq \eta \right| H_1) \leq p_d \right. \]
\[ p\left(g(x_{k-T+1:k}) \geq \eta \right| H_1) \leq \frac{1}{\eta} \left( mT + \sum_{i=k-T+1}^{k} \Delta z_i^T \Sigma^{-1} \Delta z_i \right) \]
\[ = \sum_{k=0}^{k'} \left\{ \Delta z_{0:k'} \mid p\left(g(x_{k-T+1:k}) \geq \eta \right| H_1) \leq p_d \wedge \right. \]
\[ p\left(g(x_{k-T+1:k}) \geq \eta \right| H_1) \leq \frac{1}{\eta} \left( mT + \sum_{i=k-T+1}^{k} \Delta z_i^T \Sigma^{-1} \Delta z_i \right) \]
\[ = \sum_{k=0}^{k'} \left\{ \Delta z_{0:k'} \mid p\left(g(x_{k-T+1:k}) \geq \eta \right| H_1) \leq p_d \wedge \right. \]
\[ p\left(g(x_{k-T+1:k}) \geq \eta \right| H_1) \leq \frac{1}{\eta} \left( mT + \sum_{i=k-T+1}^{k} \Delta z_i^T \Sigma^{-1} \Delta z_i \right) \]

where the first and fourth equalities follow directly from Lemma 12.
[37] M. Pajic, I. Lee, and G. J. Pappas, “Attack-resilient state estimation for noisy dynamical systems,” IEEE Trans. Control Netw. Syst., vol. 4, no. 1, pp. 82–92, Mar. 2017.

[38] A. A. Cardenas, S. Amin, and S. Sastry, “Secure control: Towards survivable cyber-physical systems,” in Proc. 28th Int. Conf. Distrib. Comput. Syst. Workshops, 2008, pp. 495–500.

[39] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, “A secure control framework for resource-limited adversaries,” Automatica, vol. 51, pp. 135–148, 2015.

[40] L. Szekeres, M. Payer, T. Wei, and D. Song, “SoK: Eternal war in memory,” in Proc. IEEE Symp. Secur. Privacy, 2013, pp. 48–62.

[41] R. S. Smith, “Covert misappropriation of networked control systems: Presenting a feedback structure,” IEEE Control Syst. Mag., vol. 35, no. 1, pp. 82–92, Feb. 2015.

[42] P. R. Kumar and P. Varaiya, Stochastic Systems: Estimation, Identification, and Adaptive Control. Philadelphia, PA, USA: SIAM, 2015.

[43] C. Gokcek, P. T. Kabamba, and S. M. Meerkov, “An LQR/LQG theory for systems with saturating actuators,” IEEE Trans. Autom. Control, vol. 46, no. 10, pp. 1529–1542, Oct. 2001.

[44] A. Nuttall, “Some integrals involving the QM function (Corresp.),” IEEE Trans. Inf. Theory, vol. TIT-21, no. 1, pp. 95–96, Jan. 1975.

[45] Y. Sun, A. Baricz, and S. Zhou, “On the monotonicity, log-concavity, and tight bounds of the generalized marcum and nuttall Q-functions,” IEEE Trans. Inf. Theory, vol. 56, no. 3, pp. 1166–1186, Mar. 2010.

[46] S. Boyd, L. E. Ghaoui, E. Feron, and V. Balakrishnan, Linear Matrix Inequalities in System and Control Theory. Philadelphia, PA, USA: SIAM, 1994.

[47] A. Alessandri, M. Baglietto, and G. Battistelli, “On estimation error bounds for receding-horizon filters using quadratic boundedness,” IEEE Trans. Autom. Control, vol. 49, no. 8, pp. 1350–1355, Aug. 2004.

[48] L. Hewing, A. Carron, K. P. Wabersich, and M. N. Zeilinger, “On a correspondence between probabilistic and robust invariant sets for linear systems,” in Proc. Eur. Control Conf., 2018, pp. 1642–1647.

[49] K. H. Johansson, “The quadruple-tank process: A multivariable laboratory process with an adjustable zero,” IEEE Trans. Control Syst. Technol., vol. 8, no. 3, pp. 456–465, May 2000.

Paul Griffioen (Member, IEEE) received the B.S. degree in electrical and computer engineering from Calvin College, Grand Rapids, MI, USA, in 2016, and the M.S. and Ph.D. degrees in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2018 and 2022, respectively.

He is currently a Postdoctoral Researcher with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA, working with Murat Arkac. His research interests include modeling, analysis, and design of high-performance cyber-physical systems that ensure safety while operating under computational constraints. His current research interests include the modeling, analysis, and design of active detection techniques and response mechanisms for ensuring resilient and secure cyber-physical systems.

Bruno Sinopoli (Fellow, IEEE) received the Dr.Eng. degree from the University of Padova, Padua, Italy, in 1998, and the M.S. and Ph.D. degrees in electrical engineering from the University of California at Berkeley, Berkeley, CA, USA, in 2003 and 2005, respectively.

He is currently the Das Family Distinguished Professor with Washington University in St. Louis, St. Louis, MO, USA, where he is also the founding Director of the Center for Trustworthy AI in Cyber-Physical Systems and the Chair of the Electrical and Systems Engineering Department. He was a Postdoctoral with Stanford University. Then, from 2007 to 2019, he was a member of the faculty with Carnegie Mellon University, where he was a Professor with the Department of Electrical and Computer Engineering with courtesy appointments in mechanical engineering and with Robotics Institute and a Co-Director of the Smart Infrastructure Institute. His research interests include modeling, analysis, and design of resilient cyber-physical systems with applications to smart interdependent infrastructure systems, such as energy and transportation, Internet of Things, and control of computing systems.