A Moving Target Defense to Detect Stealthy Attacks in Cyber-Physical Systems

J. Giraldo[1], A. Cardenas[2], and R. G. Sanfelice[2]

***Abstract*— Cyber-Physical Systems (CPS) have traditionally been considered more static, with regular communication patterns when compared to classical information technology networks. Because the structure of most CPS remains unchanged during long periods of time, they become vulnerable to adversaries who can tailor their attacks based on their precise knowledge of the system dynamics, communications, and control. Moving Target Defense (MTD) has emerged as a strategy to add uncertainty about the state and execution of a system in order to prevent adversaries from having predictable effects with their attacks. In this work we propose a novel type of MTD strategy that randomly changes the availability of the sensor data, so that it is harder for adversaries to tailor stealthy attacks and at the same time it can minimize the impact of false-data injection attacks. Using tools from switched control systems we formulate an optimization problem to find the probability of the switching signals that increase the visibility of stealthy attacks while decreasing the deviation caused by false data injection attacks.**

I. Introduction

One of the traditional problems in security is that if the adversary can predict the behavior of the system under attack, then it is very likely that attacks will be successful. Moving Target Defense (MTD) has emerged as a strategy to add uncertainty about the state and execution of a system in order to prevent adversaries from having predictable effects with their attacks [1]. According to the National Science and Technology Council, MTD *"enables us to create, analyze, evaluate, and deploy mechanisms and strategies that are diverse and that continually shift and change over time to increase complexity and cost for attackers, limit the exposure of vulnerabilities and opportunities for attack, and increase system resiliency.... The characteristics of an MTD system are dynamically altered in ways that are manageable by the defender yet make the attack space appear unpredictable to the attacker." [2].* Several authors have proposed MTD approaches for state estimation in the smart grid [3]–[5], where the main idea consists on changing the physical topology of the power grid in order to reveal false data injection attacks. Watermarking uses the addition of an external system with linear time-varying dynamics that depend on the system states and that is unknown to the attacker [6]. The detection strategy measures those external states, making it harder for an adversary to design stealthy attacks.

In this paper we propose the use of MTD to achieve **two security properties**:

1) *Detect attacks with high accuracy*; i.e., it should be hard for the attacker to evade an intrusion detection system (IDS). In Section IV we show how our MTD algorithm can help us detect a strong type of stealthy attacks introduced in [7], [8], even when the adversary knows the system dynamics, the detection strategy, has access to all control inputs, and all sensor readings.
2) *Minimize the impact of a sensor compromise*; i.e., even if the attacker compromises a sensor, once the MTD defense is activated, the impact of the attack can be attenuated. We present an optimization problem in Section V to design an MTD algorithm that minimizes the impact of attacks.

In addition, we want to prevent our MTD algorithm from degrading (significantly) the behavior of the original control system (when the system operates without MTD and without attacks). To address this performance goal we first show the conditions under which the new MTD system is stable (Section III) and in Section V we include the performance of the system as one of the constrained variables in the optimization problem.

We note that our MTD defense does not have to be active at all times. In fact, it might be activated only when there are other external indications of an attack, and therefore by activating the MTD defense, it can help us reveal previously undetected attacks.

II. Problem Statement

We consider the control system depicted in Fig. 1 which consists of a physical process, a moving target defense (MTD) mechanism that randomizes which sensor values the controller uses at a given time, an observer-based controller, and an intrusion detection system (IDS). The main goal of the MTD mechanism is to add uncertainty to the system so it is harder for the attacker to hide its attacks and simultaneously limit the impact of the attack; namely, how much control the adversary gets over the plant.

A. System Description

We consider a continuous-time linear time-invariant systems of the form

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + \delta^a(t) \\ \widetilde{y}(t) &= \Theta(t)y(t) \end{aligned} \tag{1}$$

where $x(t) \in \mathbb{R}^n, u(t) \in \mathbb{R}^m, y(t) \in \mathbb{R}^q$ are the states, input, and output vectors, respectively. The signal $\delta^a(t) \in \mathbb{R}^q$ denotes the attack vector injected to the sensors. The signal $\widetilde{y}(t) \in \mathbb{R}^q$ is the output received by the estimator, where $\Theta(t)$ denotes the MTD mechanism.

*This work is partially supported by the Air Force Office of Scientific Research under award number FA9550-17-1-0135, the Department of Commerce by Nist Award 70NANB17H282. Research by R. G. Sanfelice partially supported by NSF Grants no. ECS-1710621 and CNS-1544396, by AFOSR Grants no. FA9550-16-1-0015 and FA9550-19-1-0053, and by CITRIS and the Banatao Institute at the University of California.

[1] J. Giraldo is with the Computer Science Department at the University of Texas at Dallas, TX, 75080, USA. jairo.giraldo @ utdallas.edu.

[2] A. Cardenas and R. G. Sanfelice are with the Department of Electrical and Computer Engineering, University of California, Santa Cruz, CA 95064, USA. {alvaro.cardenas, ricardo}@ucsc.edu

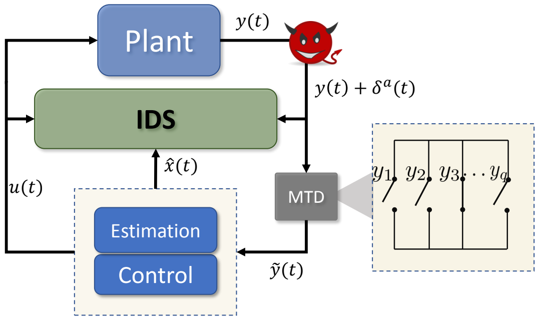

Fig. 1. Proposed MTD mechanism.

B. Moving Target Defense Mechanism

We propose an MTD approach that randomly changes the availability of the sensors as depicted in Fig. 1. Let $\Theta(t)$ be a diagonal matrix of the form $\Theta(t) = diag(\theta_1(t), \theta_2(t), \dots, \theta_q(t))$ and let $\mathcal{S} = \{1, 2, \dots, q\}$ be the sensor index set. Therefore,

$$\begin{aligned}\tilde{y}_i(t) &= \theta_i(t) y_i(t) \\ &= \theta_i(t)(C_i x(t) + \delta_i(t)),\end{aligned} \tag{2}$$

for all $i \in \mathcal{S}$, where $C_i \in \mathbb{R}^{1\times n}$ denotes the i^{th} row of matrix C and $\theta_i(t) \in \{0,1\}$ is a piecewise binary signal. In particular, we focus our attention on *random switching*, where we only need to define the probability distribution of a group of Bernoulli random variables.

Let $\mathcal{T} = \{t_0, t_1, \dots, t_k, \dots\}$ denote the set of time points where the MTD strategy is updated with $0 < T_{min} < t_k - t_{k-1} < T_{max}$ and let $(S_k)_{k\in\mathbb{Z}_+}$ be the sequence of holding times where $S_{k+1} = t_{k+1} - t_k$. We can define $\beta_j(t_k) \sim \mathcal{B}(p_j)$ as a random variable drawn from a Bernoulli distribution such that $\beta_j(t_k) = 1$ with probability p_j (and zero otherwise) for all $t_k \in \mathcal{T}$. Therefore, we have that on each time interval $[t_k, t_{k+1})$, $k \in \mathbb{Z}_+$,

$$\theta_j(t) = \beta_j(t_k), \quad \forall t \in [t_k, t_{k+1}).$$

Remark 2.1: We will see later that the sequence $(S_k)_{k\in\mathbb{Z}_+} = t_{k+1} - t_k$ is considered random, which adds an extra level of uncertainty to the MTD strategy, making even harder for an adversary to predict the system's behavior.

C. State-Observer and Control with MTD Measurements

We assume that the pair (C, A) is detectable. Since the elements of $\tilde{y}(t)$ are switching over time, we propose a state observer described by

$$\dot{\hat{x}}(t) = A\hat{x}(t) + Bu(t) + L\Theta(t)(\tilde{y}(t) - C\hat{x}(t)), \tag{3}$$

where $L \in \mathbb{R}^{n\times q}$. This observer knows which sensor readings are active (which can be easily guaranteed by sending $\theta_j(t)$ along with $y(t)$), and updates its estimation using only those active readings. We define $e(t) = x(t) - \hat{x}(t)$ as the estimation error. Combining (1) and (3), and since $\Theta(t)\Theta(t) = \Theta(t)$ we obtain

$$\dot{e}(t) = (A - L\Theta(t)C)e(t) - L\Theta(t)\delta^a(t), \tag{4}$$

and the observer design becomes a stabilization problem where L and $\Theta(t)$ have to be chosen in such a way that the switched system in (4) has $e = 0$ globally asymptotically stable when $\delta^a(t) = 0$ for all $t \geq 0$.

Finally, we consider that the pair (A, B) is controllable and the output-feedback controller of the form

$$u(t) = K\hat{x}(t). \tag{5}$$

D. Intrusion Detection

Taking advantage of the state estimator in (3), we can construct anomaly detection modules that compare the estimated sensor readings with the real ones to determine the presence of an attack. Therefore, we define the residuals as

$$\begin{aligned}r(t) &= y(t) - C\hat{x}(t) \\ &= Ce(t) + \delta^a(t).\end{aligned} \tag{6}$$

The anomaly detection then takes the vector of residuals $r(t)$ and computes a measure of how deviated the sensor readings are from the estimation. There are different types of anomaly detection strategies, such as the χ^2-test, distributed bad-data detection, and CUSUM [8]. For simplicity, we will focus our attention on the distributed bad-data detection with the detection statistic given by

$$h(t) = |r(t)|, \tag{7}$$

where $|\cdot|$ is evaluated component-wise. If any $h_i(t) > \tau_i$, for some fixed detection threshold $\tau_i > 0$, then an alarm is triggered.

E. Adversary Model

Capabilities and goals: the attacker has compromised a set of sensors and wants to send false signals δ^a to drive the system away from the operational states.

Knowledge: the attacker knows the non-MTD system model; i.e., the attacker knows A, B, C, K, the estimation gain L, and the detection threshold τ but does not know the MTD mechanism $\Theta(t)$. These strong assumptions allow us to consider worst-case scenarios, implying that our MTD will be effective against weaker adversaries.

F. Motivational Example

To illustrate why our MTD approach can make difficult for an adversary to design strong attacks and also minimize the impact of an attack in the system states, we consider the simple example where $A = -0.1, B = 1, L = 0.2, K = -0.3, C = 1$, with an intrusion detection threshold of $\tau_i = \tau = 0.1$.

*a) **Without MTD**:* $\Theta(t) = 1$, and if $\delta^a = 0.3$, then the detection statistic in the limit converges to

$$\lim_{t\to\infty} h(t) = |C(A - LC)^{-1}L\delta^a + \delta^a| = 0.1,$$

and therefore the attack remains stealthy (undetected by our residual-based IDS). We can measure the impact of the attack in terms of how much the system state is deviated from the origin. Without MTD, we have that

$$\lim_{t\to\infty} x(t) = (A + BK)^{-1}BK(A - LC)^{-1}L\delta^a = -0.15$$

*b) **With MTD**:* With the proposed MTD mechanism where $\Theta(t) = \theta(t)$ with $p = 0.3$, and $\delta^a = 0.3$, we have that

$$\lim_{t\to\infty} E[h(t)] = |C(A - LpC)^{-1}Lp\delta^a + \delta^a| = 0.1875,$$

and *the same attack is no longer stealthy*. Furthermore, since the IDS knows the MTD realization, the addition of MTD does not increase the false positive rate.

Now, with the MTD mechanism, the expected state converges to

$$\lim_{t\to\infty} E[x(t)] = (A+BK)^{-1}BK(A-LpC)^{-1}Lp\delta^a = -0.084.$$

Note that for this particular example the random MTD mechanism causes the residuals to increase while the state deviation decreases which illustrates the double benefit that can be achieved with our approach.

The cost of MTD can be observed in terms of the convergence speed. In our example, the slowest (and only) eigenvalue of the expected estimation error is $\lambda_{MTD} = A - LpC$ and without MTD is $\lambda_{noMTD} = A - LC$. Clearly $\lambda_{MTD} > \lambda_{noMTD}$ for any $0 \leq p < 1$ and therefore the observer convergence is degraded when p is small.

In the next section we formulate our problem as a switched system and derive conditions for stability.

III. Stability of the MTD System

A. Switched System

In order to formulate our problem as a switched system and exploit some existing tools, we define the family of non-identical diagonal binary matrices $\{\Theta_1, \Theta_2, \dots, \Theta_s\}$, and the finite index set $\Sigma = \{1, 2, \dots, s\}$, where $s = 2^q$. Each $\Theta_i \in \mathbb{R}^{q\times q}$ describes one possible combination of $\{1, 0\}$ for each $\theta_1, \theta_2, \dots, \theta_q$, where $i \in \Sigma$. We also define the piecewise switching signal $\sigma : [0, \infty) \to \Sigma$, which is updated at the time points $t_k \in \mathcal{T}$ and remains constant in the time interval (t_k, t_{k+1}). The signal $\sigma(t)$ is used to specify, at each time instant t, the index $i \in \Sigma$ of each active subsystem.

Then, our MTD approach in (2) can be rewritten as

$$\widetilde{y}(t) = \Theta_{\sigma(t)}y(t), \tag{8}$$

where $\sigma(t)$ randomly chooses among the index set Σ, according to the probability mass function $\Omega : \Sigma \to [0, 1]$, where, for each $i \in \Sigma$,

$$\begin{aligned}\Omega(i) = \widetilde{p}_i &= \prod\nolimits_{j\in\Sigma} [\Theta_i]_j p_j + (1 - [\Theta_i]_j)(1 - p_j) \\ &= \prod\nolimits_{j\in\Sigma} (1 - p_j - [\Theta_i]_j + 2[\Theta_i]_j p_j), \end{aligned} \tag{9}$$

for $[\Theta_i]_j$ refers to the j^{th} diagonal element of matrix Θ_i.

Example: Let $p_i = p$, and the number of sensors is $q = 2$. Then there exist 4 possible matrices Θ_i, given by $\Theta_1 = diag(0, 0)$, $\Theta_2 = diag(1, 0)$, $\Theta_3 = diag(0, 1)$, $\Theta_4 = diag(1, 1)$, with a probability mass function $\Omega(i) = \{(1 - p)^2, p(1 - p), p(1 - p), p^2\}$. for all $i \in \Sigma$

Having formulated our MTD strategy as a switched system, we can rewrite the observer in (3) as follows

$$\dot{\hat{x}}(t) = A\hat{x}(t) + Bu(t) + L\Theta_{\sigma(t)}(\widetilde{y}(t) - C\hat{x}(t)), \tag{10}$$

and the estimation error can be described by

$$\dot{e} = (A - L\Theta_{\sigma(t)}C)e - L\Theta_{\sigma(t)}\delta^a(t). \tag{11}$$

Let us define $F_{E,\sigma(t)} = A - L\Theta_{\sigma(t)}C$, and let $z(t) = [x^\top(t), e^\top(t)]^\top$ denote the extended state vector, such that

$$\begin{aligned}\dot{z} &= \begin{bmatrix} A + BK & -BK \\ 0 & F_{E,\sigma(t)} \end{bmatrix} z + \begin{bmatrix} 0 \\ -L\Theta_{\sigma(t)} \end{bmatrix} \delta^a \\ &=: F_{\sigma(t)}z + G_{\sigma(t)}\delta^a. \end{aligned} \tag{12}$$

Thanks to the separation principle, we can design K independently of the observer gain or the switching signal (e.g., an LQR that satisfies specific performance conditions). Therefore, if K is such that $A + BK$ is stable, the stability of (12) is dictated by $F_{E,\sigma(t)}$.

B. Stability Conditions

Assume $\delta^a(t) = 0$ for $t \geq 0$. Recall that we have the family of matrices $F_{E,i} = A - L\Theta_i C$ for all $i \in \Sigma$. When $\Theta_1 = diag(0, 0, \dots, 0)$ (which indicates the case when all sensors are off at the same time), then $F_{E,1} = A$. Since A is not necessarily Hurwitz, we need to define a general stability condition for switched systems in the presence of unstable subsystems and random switching.

We use the results in [9], where the authors establish globally asymptotic stability conditions (GAS) for switched systems with stable and unstable subsystems, where the switching signal has specific random properties. In fact, it only requires that the probability the unstable subsystems are active to be small.

We are interested in the following definition of stability introduced in [9].

Definition 3.1: The system (12) is said to be globally asymptotically stable almost surely (GAS a.s.) if the following two properties are simultaneously verified:

$$\Pr\left(\forall\epsilon > 0\ \exists\beta > 0,\ \text{such that } \|x_0\| < \beta \implies \sup_{t\geq 0}\|x(t)\| < \epsilon)\right) = 1.$$

$$\Pr\left(\forall r, \epsilon' > 0\ \exists T \geq 0 \text{ such that } \|x_0\| < r \implies \sup_{t\geq T}\|x(t)\| < \epsilon'\right) = 1$$

Definition 3.1 indicates that the trajectories of $x(t)$ converge to an equilibrium with probability 1 from any bounded initial condition x_0.

The conditions for stability under random switching introduced in [9] employ a family of Lyapunov functions, one for each subsystems $F_{E,i}$ for $i \in \Sigma$, that possesses the following properties.

Assumption A1: There exist a family of continuously differentiable real-valued functions $V_i(x) \in \mathbb{R}$ for all $i \in \Sigma$, functions $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$, numbers $\mu \geq 1, \lambda_i \in \mathbb{R}$ such that

(A1.1): $\alpha_1(\|x\|) \leq V_i(x) \leq \alpha_2(\|x\|),\ \ \forall x \in \mathbb{R}^n, \forall i \in \Sigma.$
(A1.2): $\dot{V}_i(x) \leq -\lambda_i V_i(x),\ \ \forall x \in \mathbb{R}^n, \forall i \in \Sigma,$
(A1.3): $V_i(x) \leq \mu V_j(x),\ \ \forall x \in \mathbb{R}^n, \forall i, j \in \Sigma.$

We also impose some assumptions on the switching signal.

Assumption A2: The switching signal $\sigma(t)$ satisfies the following properties:

- The sequence $(S_k)_{k\in\mathbb{N}}$, $S_{k+1} = t_{k+1} - t_k$ of holding times is a sequence of i.i.d. uniform random variables with parameter $T_{max} > 0$ and $T_{min} = 0$.
- The probability that the i^{th} subsystem is active is $\Pr(\sigma(t_k) = i) = \widetilde{p}_i$.

- S_k and σ_i are mutually independent.

The following theorem follows [9, Theorem 3.4], and introduces sufficient conditions for GAS a.s. of the switched system in (12) according to Definition 3.1.

Theorem 3.1: Suppose that Assumptions **A1** and **A2** hold where $\sigma(t)$ has parameter T_{max} and probabilities $\widetilde{p}_i = \Omega(i)$ for all $i \in \Sigma$ according to (9). If

$$\mu \sum_{i\in\Sigma} \widetilde{p}_i \left(\frac{1 - e^{-\lambda_i T_{max}}}{\lambda_i T_{max}} \right) < 1, \tag{13}$$

then the switched system is GAS a.s..

Theorem 3.1 is an adaptation of [9, Theorem 3.4] for linear systems, where $\sigma_i(t_k)$ follows the probability distribution Ω defined in (9). The proof is omitted due to space constraints.

Remark 3.1: If $\lambda_i < 0$ (which is related to the unstable matrices) and it has large magnitude, the term $\left(\frac{1-e^{-\lambda_i T_{max}}}{\lambda_i T_{max}}\right)$ may be greater than one, such that the probability associated to that term has to be small enough; on the other hand, if $\lambda_i > 0$, the term $\left(\frac{1-e^{-\lambda_i T_{max}}}{\lambda_i T_{max}}\right)$ is positive and it gets closer to zero when the magnitude of λ_i increases. Therefore, in order to guarantee that (13) holds, $\widetilde{p}_i$ has to be chosen such that the unstable subsystems are selected with low probability.

IV. Detecting Stealthy Sensor Attacks

One of the main advantages of MTD is that it makes harder for an adversary to tailor stealthy attacks due to the uncertainty added by the MTD mechanism. In particular, with our proposed sensor MTD, the adversary fails to predict how his attack affects the IDS, such that the attacks that are stealthy under normal conditions, are visible with the MTD strategy.

We will focus our attention on a very powerful type of stealthy attack that has been introduced in [7], [8], [10]. Then, we will show how, by appropriately selecting the probabilities p_i, it is possible to make these attacks visible, even when the adversary has access to the control inputs, all sensor readings, knows A, B, L, C, K, and knows the thresholds τ of the detection mechanism.

While we could try to define a similar non-stochastic defense by changing C deterministically, this would give the adversary more chances of finding the deterministic changes and adapt its attack accordingly. The uncertainty presented to the adversary is one of the advantages of MTD.

A. Construction of Stealthy Attacks

Suppose the attacker has access to all sensor readings and computes its own estimation of the system states $\hat{x}_a(t)$ in order to forge powerful cyber-attacks. The attacker's estimator is described by

$$\dot{\hat{x}}_a(t) = A\hat{x}_a(t) + Bu(t) + L(Cx(t) - C\hat{x}_a(t) + \delta^a(t)). \tag{14}$$

Let $s(t) = \hat{x}(t) - \hat{x}_a(t)$ denote the error between the system estimation used by the controller and the attacker estimation. We introduce the following lemma.

Lemma 4.1: Suppose there is no MTD mechanism, i.e., $\Theta_{\sigma(t)} = I$, and L is such that $A - LC$ is Hurwitz. Then, the error $s(t)$ converges in the limit to $\lim_{t\to\infty} s(t) = 0$ and the attacker is able to compute an estimation that converges to the one used by the IDS.

Proof: Notice that $\dot{s}(t) = \dot{\hat{x}}(t) - \dot{\hat{x}}_a(t)$. Combining (10) and (14) we get

$$\begin{aligned} \dot{s}(t) &= F_{E,\sigma(t)} s(t) + L(\Theta_{\sigma(t)} - I)Ce(t) \\ &\quad - L(\Theta_{\sigma(t)} - I)Cs(t) + L(\Theta_{\sigma(t)} - I)\delta^a(t). \end{aligned} \tag{15}$$

Since $\Theta_{\sigma(t)} = I$, we have that $\dot{s}(t) = (A - LC)s(t)$, which is stable independently of $\delta^a(t)$ and the trajectories will always converge to 0. ■

W.l.o.g., in the reminder of this section we will assume that the system is in steady state before the attack, such that $x(0) = 0, s(0) = 0$.

In the following lemma we will introduce a type of stealthy attack that uses $\hat{x}_a(t)$ to bypass the IDS algorithm. This attack does not depend on the zero-dynamics, which makes it suitable for more general applications.

Lemma 4.2: Suppose that the detection strategy corresponds to the bad-data detection introduced in (7) with detection thresholds $\tau = [\tau_1, \ldots, \tau_q]^\top$. If there is no MTD mechanism and the adversary launches an attack of the form

$$\delta^a(t) = -y(t) + C\hat{x}_a(t) + \tau \tag{16}$$

then the attack remains stealthy.

Proof: Replacing (16) in (6), we obtain

$$\begin{aligned} r(t) &= C(x(t) - \hat{x}(t)) - C(x(t) - \hat{x}_a(t)) + \tau \\ &= -Cs(t) + \tau \end{aligned} \tag{17}$$

Without MTD, $s(t) = 0$ and the residuals are then $r(t) = \tau$. As a consequence, $h(t) = |\tau|$ and the alarm is never triggered. ■

Remark 4.1: This type of attack is very powerful when the matrix A is not stable. If we apply the attack in (16) to (12), the dynamics of the estimation error become $\dot{e}(t) = Ae(t) + L\Theta_{\sigma(t)}Cs(t) + L_{\sigma(t)}\tau$. If we define the extended state $w = [x^\top, e^\top, s^\top]^\top$, it is easy to see from $\dot{w} = Hw + J\tau$ that part of the eigenvalues of H correspond to the eigenvalues of A. If A is not stable, the attack causes that the entire system becomes unstable without being detected.

B. Revealing Stealthy Attacks

We assume that the adversary does not know the MTD mechanism, such that he launches the stealthy attack in (16). The following theorem introduces the conditions to reveal the stealthy attack.

Theorem 4.1: Suppose that the conditions in Theorem 3.1 are satisfied and an adversary launches the stealthy attack described in (16) for the bad-data detection strategy. Let $E[\Theta_{\sigma(t)}] = \boldsymbol{P} = diag(p_1 \ldots, p_q)$ such that $\bar{F}_E = A - L\boldsymbol{PC}$ is Hurwitz. The stealthy attack is revealed if any of the following conditions holds for *at least one* $j \in \mathcal{S}$,

$$\begin{aligned} &C_j \bar{F}_E^{-1} L(\boldsymbol{P} - I)\tau > 0, \\ &C_j \bar{F}_E^{-1} L(\boldsymbol{P} - I)\tau < -2\tau_j. \end{aligned} \tag{18}$$

Proof: Replacing the attack in (16) with the dynamics of the error in (15), we obtain

$$\dot{s}(t) = (A - L\Theta_{\sigma(t)}C)s(t) + L(\Theta_{\sigma(t)} - I)\tau. \tag{19}$$

Since τ is finite and constant, and since when $\delta^a(t) = 0$, (12) is GAS a.s. according to Theorem 3.1, then the term $L(\Theta_{\sigma(t)} - I)\tau$ will cause an accumulation of the error between the real effect of the attack and the effect estimated by the attacker. To facilitate the analysis, and since $s(t)$ is independent of $\Theta_{\sigma(t)}$, we define $E[s(t)] = \bar{s}(t)$. Then, $\dot{\bar{s}}(t) = \bar{F}_E \bar{s}(t) + L(\boldsymbol{P} - I)\tau$. Therefore, the following limit exists

$$\lim_{t\to\infty} \bar{s}(t) = -\bar{F}_E^{-1} L(\boldsymbol{P} - I)\tau,$$

which with $E[r(t)] = \bar{r}(t)$ and (17) leads to $\lim_{t\to\infty} \bar{r}(t) = \left(C\bar{F}_E^{-1} L(\boldsymbol{P} - I) + I\right)\tau$. Applying the absolute value, with $E[h(t)] = \bar{h}(t)$ leads to

$$\lim_{t\to\infty} \bar{h}(t) = |C\bar{F}_E^{-1} L(\boldsymbol{P} - I)\tau + \tau|. \tag{20}$$

such that the attack is reveled when at least one $h_j(t) > \tau_j$, which is ensured if any of the conditions in (18) hold. ■.

Notice that the conditions for revealing the attack depend on $\boldsymbol{P}$. As a consequence, in the next section we show how to impose constraints on $\boldsymbol{P}$ during the design process to guarantee that this type of strong stealthy attacks are always revealed.

Remark 4.2: With our proposed MTD, the attacker is not able to estimate the system states subject to his own attack. Therefore, *any attack that depends on the attackers estimation can be potentially revealed.* Furthermore, as it was shown in the Motivation example in Section II-F, other types of stealthy attacks can be also revealed.

V. MTD Design

So far, Theorem 3.1 provides conditions for almost sure asymptotic stability of the system subject to the proposed MTD strategy. In Theorem 4.1 we derived conditions for revealing a powerful type of stealthy attacks. In this section we introduce a methodology to design the probability matrix $\boldsymbol{P}$ that satisfies the conditions introduced in Theorems 3.1 and 4.1, while increasing the system resiliency to any type of attack (not only stealthy attacks).

Recall that $\boldsymbol{P} = E[\Theta(t)] = diag(p_1, p_2, \ldots, p_q)$, and let $\bar{z}(t) = E[z(t)] = [\bar{x}^\top(t), \bar{e}^\top(t)]^\top$. Applying the expectation operator $E[\cdot]$ to (12), we obtain

$$\dot{\bar{z}}(t) = \bar{F}\bar{z}(t) + \bar{G}\bar{\delta}^a(t), \tag{21}$$

where $E[\delta^a(t)] = \bar{\delta}^a(t)$, and

$$\bar{F} = \begin{bmatrix} A + BK & -BK \\ 0 & A - L\boldsymbol{P}C \end{bmatrix}, \quad \bar{G} = \begin{bmatrix} 0 \\ -L\boldsymbol{P} \end{bmatrix}.$$

Impact of the Attack: We can define the impact of the attack in terms of how much an attack can deviate the convergence of the trajectories with respect to the nominal conditions in expectation. To this end, for concreteness we assume that $\bar{\delta}^a(t)$ is constant, such that

$$\lim_{t\to\infty} \bar{z}(t) = -\bar{F}^{-1}\bar{G}\bar{\delta}^a, \tag{22}$$

leading to $\lim_{t\to\infty} \bar{x}(t) = (A + BK)^{-1}BK(A - L\boldsymbol{P}C)^{-1}L\boldsymbol{P}\bar{\delta}^a$. Therefore, we can quantify the impact of the attack as follows:

$$\mathcal{I}(\bar{\delta}^a) = \|M\bar{\delta}^a\|,$$

for $M = (A + BK)^{-1}BK(A - L\boldsymbol{P}C)^{-1}L\boldsymbol{P}$.

Performance under MTD: Notice from (12) and (21), that if A is Hurwitz, the trivial solution $\boldsymbol{P} = 0$, completely eliminates any effect of an attack in the system. However, this will eliminate the observer altogether and therefore, prevent us from building an estimate of the system state and generate adequate control actions. Therefore, we need to impose a performance criteria that the MTD must satisfy in the attack-free case (e.g., convergence speed).

In order to quantify the degradation caused by the MTD mechanism in the system, we use as a performance index the slowest eigenvalue of $\bar{F}_E = A - L\boldsymbol{P}C$, which is related to the convergence speed of the observer. Therefore, our goal is to design an MTD strategy that guarantees

$$\lambda_{max}(A - L\boldsymbol{P}C) \leq \gamma,$$

where $\gamma < 0$, and where $\lambda_{max}(X)$ is the maximum real part of the eigenvalues of X.

Finally, since we want to find $\boldsymbol{P}$ to reveal stealthy attacks according to Theorem 4.1, let $\Psi_j^+ = C_j\bar{F}_E^{-1}L(\boldsymbol{P} - I)\tau$ and $\Psi_j^- = C_j\bar{F}_E^{-1}L(\boldsymbol{P} - I)\tau + 2\tau_j$. Thus, we can define

$$\Psi = \sum_{j\in\mathcal{S}} \max\{\Psi_j^+, 0\} - \min\{\Psi_j^-, 0\},$$

such that *at least one* $h_j > \tau_j$ when $\Psi > 0$.

The optimization problem to find $\boldsymbol{P}$ that guarantees GAS a.s., ensures that the type of stealthy attacks are revealed, and that minimizes the impact of the attack is described as follows:

$$\begin{aligned} &\min_{\boldsymbol{P}} \|M\| \\ &s.t. \\ &0 < p_j \leq 1, \ \forall j \in \mathcal{S} \\ &\Psi > 0, \\ &(9), (13) \\ &\lambda_{max}^R(\bar{F}_E) \leq \gamma. \end{aligned} \tag{23}$$

Note that this is a nonlinear optimization problem that, at times, can be solved using interior-point or active-set algorithms.

Remark 5.1: The gain L can also be included as a design parameter in the proposed optimization problem, but it makes (23) non-convex, such that the optimal solution is not unique. In this case, it would be necessary to use different approaches to find a good combination of L, and $\boldsymbol{P}$. On the other hand, L can also be chosen to decrease the number of unstable subsystems such that it could be possible to find smaller probabilities p_j for the same performance degradation.

VI. Case Study

In order to verify the viability of our approach, we consider the LTI system with matrices

$$A = \begin{bmatrix} 1 & 0.5 & 0 \\ 0.3 & -2 & -0.5 \\ 0.1 & 1 & -2 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

The feedback control gain is $K = [-16.3262, -2.5487, 0.4054]$ and L corresponds to the steady state Kalman filter gain

$$L = \begin{bmatrix} 2.0726 & 0.1846 \\ 0.1846 & 0.0348 \\ 0.1312 & 0.0362 \end{bmatrix}.$$

Since the number of sensors is $q = 2$, there are four possible subsystems i.e., $s = 4$. Therefore, Assumption **A1** is satisfied for $\lambda = [2.09, -2.07, 2.05, -2.1]$, and the Lyapunov function for all $i \in \Sigma$ is $V_i(x) = V = x^\top Qx$, where

$$Q = \begin{bmatrix} 0.2098 & -0.0168 & -0.0294 \\ -0.0168 & 0.7087 & -0.0906 \\ -0.0294 & -0.0906 & 0.3963 \end{bmatrix},$$

such that $\mu = 1$. Let $\gamma = -1$, $T_{max} = 0.1$ and the detection thresholds $\tau = [0.01, 0.01]^\top$. The solution of the optimization problem in (23) is found using the interior-point algorithm and corresponds to $\boldsymbol{P}^* = diag([0.98, 0.503])$.

Suppose that an adversary injects the attack $\delta^a(t) = [-0.1, 1.7]^\top$ after 20 s. Figure 2 illustrates the Montecarlo simulation of the trajectories of the states and the norm $\|x(t)\|$. Clearly, the MTD approach is able to decrease the impact of this attack when compared to the case without MTD.

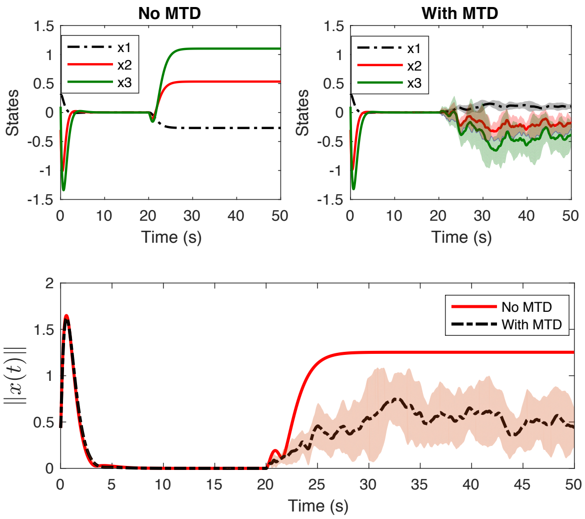

Fig. 2. System states without MTD (top left), and with the MTD strategy (top right). The shaded area indicate the maximum/minimum deviation of the Montecarlo simulation at each time instant. Notice that our approach decreases the deviation caused by the attack, and since $\gamma = -1$ and $p_1 = 0.98$, the performance degradation is small.

Now, suppose that an adversary launches a stealthy attack as described in (16) for the anomaly detection threshold τ. Figure 3 shows how without MTD, the attack remains completely stealthy. However, thanks to the random MTD mechanism, the attack is easily revealed.

VII. Conclusions

We have proposed and analyzed the security of an MTD strategy for improving the detectability of attacks, while at the same time minimizing the power that an adversary has when compromising a sensor signal. We showed that our strategy is effective against very powerful stealthy attacks even when the adversary knows the system dynamics, the detection strategy, and has access to all sensors and control inputs. We derived conditions for the MTD strategy to keep the system stable and defined an optimization problem that allows us to find the probability at which each sensor transmits its information that guarantee the detection of stealthy attacks and that minimize the impact caused by the attack. In practice the MTD strategy can be activated when we notice indicators of attacks, or if we notice that the system is deviating from the desired space without explanation; if the MTD is activated then, it will be able to mitigate the attack while at the same time revealing a previously undetected attack. Future work includes extending the results to the case of recurring attacks using the hybrid systems techniques used in [11]

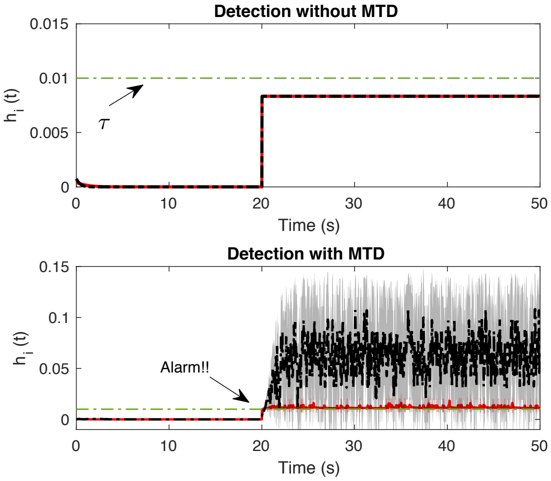

Fig. 3. Bad-data detection with thresholds $\tau_i = \tau = 0.01$ in the presence of a stealthy attack. The attack is never detected without MTD (top), but the addition of uncertainty makes possible to reveal the attack (bottom).

References

[1] S. Jajodia, A. K. Ghosh, V. Swarup, C. Wang, and X. S. Wang, *Moving target defense: creating asymmetric uncertainty for cyber threats*. Springer Science & Business Media, 2011, vol. 54.

[2] E. O. of the President, "Trustworthy cyberspace: Strategic plan for the federal cyber security research and development program," National Science and Technology Council, Tech. Rep., 2011.

[3] K. R. Davis, K. L. Morrow, R. Bobba, and E. Heine, "Power flow cyber attacks and perturbation-based defense," in *Proceedings of the IEEE Third International Conference on Smart Grid Communications (SmartGridComm), 2012*. IEEE, 2012, pp. 342–347.

[4] M. A. Rahman, E. Al-Shaer, and R. B. Bobba, "Moving target defense for hardening the security of the power system state estimation," in *Proceedings of the First ACM Workshop on Moving Target Defense*. ACM, 2014, pp. 59–68.

[5] J. Tian, R. Tan, X. Guan, and T. Liu, "Hidden moving target defense in smart grids," in *Proceedings of the 2nd Workshop on Cyber-Physical Security and Resilience in Smart Grids*. ACM, 2017, pp. 21–26.

[6] S. Weerakkody and B. Sinopoli, "Detecting integrity attacks on control systems using a moving target approach," in *Proceedings of the IEEE 54th Annual Conference on Decision and Control (CDC),*. IEEE, 2015, pp. 5820–5826.

[7] J. Giraldo, A. Cardenas, and M. Kantarcioglu, "Security and privacy trade-offs in cps by leveraging inherent differential privacy," in *2017 IEEE Conference on Control Technology and Applications (CCTA)*, Aug 2017, pp. 1313–1318.

[8] C. Murguia and J. Ruths, "Cusum and chi-squared attack detection of compromised sensors," in *2016 IEEE Conference on Control Applications (CCA)*, Sept 2016, pp. 474–480.

[9] D. Chatterjee and D. Liberzon, "Stabilizing randomly switched systems," *SIAM Journal on Control and Optimization*, vol. 49, no. 5, pp. 2008–2031, 2011.

[10] D. I. Urbina, J. A. Giraldo, A. A. Cardenas, N. O. Tippenhauer, J. Valente, M. Faisal, J. Ruths, R. Candell, and H. Sandberg, "Limiting the impact of stealthy attacks on industrial control systems," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16. New York, NY, USA: ACM, 2016, pp. 1092–1105.

[11] S. Phillips, A. Duz, F. Pasqualetti, and R. G. Sanfelice, "Hybrid attack monitor design to detect recurrent attacks in a class of cyber-physical systems," in *Proceedings of the IEEE Conference on Decision and Control*, 2017, pp. 1368–1373.