2024 American Control Conference (ACC)
July 8-12, 2024. Toronto, Canada

A Set-Based Control Mode Selection Approach for Active Detection of False Data Injection Cyberattacks

Shilpa Narasimhan, Nael H. El-Farra, Matthew J. Ellis

Abstract— **In the last two decades, several highly sophisticated cyberattacks have targeted process control systems (PCSs) that operate chemical processes. To enhance PCS cybersecurity, cyberattack detection schemes utilizing operational data to reveal the presence of attacks on PCSs have received extensive attention. Stealthy attacks are designed to evade detection by an operational technology-based detection scheme. Their detection may require an active detection method, which perturbs the process by utilizing an external intervention for attack detection. In this work, two control modes that may be used to induce perturbations for active attack detection of steathly false-data injection cyberattacks are presented. A reachability analysis is used to develop a set-based condition indicating that if met by a specific stealthy attack, the attack will be detected and therefore, the control mode is considered to be "attack-revealing". Leveraging the condition, a screening algorithm that may be used to select an attack-revealing control mode is presented. Using an illustrative process, the application of the screening algorithm is demonstrated.**

I. Introduction

In the United States, the chemical manufacturing sector is one of 16 critical infrastructure sectors because of the sector's impact on security, national economics, and national public health and safety [1]. Process control systems (PCSs) are industrial control systems operating chemical manufacturing processes and, over the past two decades, have been subject to highly sophisticated false data injection (FDI) cyberattacks that aim to compromise the integrity of the data over the PCS communication channels [1]. To manage the risk of a cyberattack on a PCS, operational technology (OT)-based approaches for the detection, identification, and mitigation of the impact of a cyberattack are being developed [2], [3].

Many OT-based approaches for cyberattack detection are designed to detect an attack if the data over the PCS network deviate from their baseline values [4], [5]. OT-based attack detection approaches may be broadly classified as passive detection schemes and active detection methods. Passive attack detection schemes monitor a process for attacks without utilizing an external intervention. Approaches for passive attack detection in the literature include those using standard anomaly detection approaches such as the χ^2 detection scheme and the cumulative sum detection scheme [6]–[8]. Approaches utilizing machine learning and pattern mining to detect attacks have also been proposed [9]–[11].

Shilpa Narasimhan, Nael H. El-Farra, and Matthew J. Ellis are with the Department of Chemical Engineering, University of California, Davis, Davis, CA 95616, USA. Emails: shnarasimhan@ucdavis.edu, nhelfarra@ucdavis.edu, and mjellis@ucdavis.edu. Corresponding author: M. J. Ellis. Financial support from the National Science Foundation is gratefully acknowledged.

Stealthy FDI attacks are an important type of attacks, as they evade detection. Active detection methods using an external intervention to perturb the process may be used to detect stealthy attacks that cannot be detected by passive detection schemes. Active detection of stealthy attacks has received some attention [12]–[19]. For a process monitored by a residual-based detection scheme, an active detection method using injection of external signals has been explored in [12]. Some active detection methods have considered injection of randomized inputs [13] or injecting random inputs and designing a secure control architecture [14] to prevent an attack from being stealthy (indirectly enabling attack detection). Moving target defense is another active detection method, under which, an external auxiliary system with additional actuators and sensors and with time-varying dynamics is added [15]–[17]. In our prior work [18], [19], control parameter switching to operate the process under an "attack-sensitive" mode was proposed where the attack-sensitive mode is one where some attacks destabilize the process, thereby leading to attack detection. However, attack detection by destabilization may be undesirable.

In this work, alternative (non-destabilizing) active detection methods for detecting stealthy false-data injection cyberattacks that alter the data communicated over the PCS communication channels are considered. In particular, an active detection method employing one or both of two possible control modes, one involving changing set points and the other involving switching control parameters, is considered for the active detection of a class of stealthy false data injection attacks. Implementing either control mode induces perturbations in the closed-loop process. A reachability analysis is used to develop a set-based condition. If the condition is satisifed for a specific stealthy attack, the attack is guarantted to be detected under the control mode, forming the basis for "attack-revealing" control modes. Using the condition, a screening algorithm that may be used to choose a control mode to guarantee attack detection is presented. The application of the screening algorithm is demonstrated using an illustrative process example.

II. Preliminaries

A. Class of Processes

Processes modeled by discrete-time linear time-invariant dynamics are considered:

$$x_{t+1} = A^x x_t + B^u u_t + B^w w_t \tag{1a}$$

$$y_t = C^x x_t + v_t \tag{1b}$$

where $x_t \in \mathbb{R}^n$ is the state vector, $u_t \in \mathbb{R}^l$ is the manipulated input vector, $w_t \in \mathcal{W} \subset \mathbb{R}^p$ is the process disturbance vector, $y_t \in \mathbb{R}^m$ is the measured output vector, and $v_t \in \mathcal{V} \subset \mathbb{R}^m$ is the measurement noise vector. Without loss of generality, $t = 0$ is taken to be the initial time. The sets $\mathcal{W}$ and $\mathcal{V}$ are convex polytopes. A^x, B^u, and C^x are matrices of appropriate dimensions, and B^u has full column rank.

A Luenberger observer is used to estimate the states:

$$\hat{x}_{t+1} = A^x \hat{x}_t + B^u u_t + L(y_t - \hat{y}_t) \tag{2a}$$

$$\hat{y}_t = C^x \hat{x}_t \tag{2b}$$

where $\hat{x}_t \in \mathbb{R}^n$ is the estimated state vector, $\hat{y}_t \in \mathbb{R}^m$ is the estimated output vector, and $L \in \mathbb{R}^n \times \mathbb{R}^m$ is the observer gain. The control objective is to operate at a desired operating steady-state $x^s \in \mathbb{R}^n$. To achieve the control objective, a linear tracking controller is employed, given by:

$$u_t = -K(\hat{x} - x^s) + u_s \tag{3}$$

where $K \in \mathbb{R}^l \times \mathbb{R}^n$ is the controller gain and u_s is the controller bias used to achieve offset-free control. For simplicity, the expected value of the process disturbance is assumed to be zero. The bias may be computed from:

$$u_s = G^u(I - A^x)x^s \tag{4}$$

where $G^u = ((B^u)^T B^u)^{-1}(B^u)^T \in \mathbb{R}^l \times \mathbb{R}^n$ is the left pseudo-inverse of B^u. In this work, changing the operating steady-state is considered. The operating steady-state is referred to as the set point for simplicity. The set points are assumed to be selected such that they are reachable in the sense that there exists $u_s \in \mathbb{R}^l$ satisfying Eq. 4.

The state estimation error dynamics are given by: $e_{t+1} = (A^x - LC^x)e_t + B^w w_t - Lv_t$ where $e_t \coloneqq x_t - \hat{x}_t \in \mathbb{R}^n$ denotes the state estimation error. The collective dynamics of the closed-loop process encompass both the process states and the estimation error. To facilitate the analysis, an augmented state vector $\xi_t \coloneqq [x_t^T \ e_t^T]^T$ is defined, and its dynamics are given by:

$$\begin{aligned}\xi_{t+1} = &\underbrace{\begin{bmatrix} A^x - B^u K & B^u K \\ 0 & A^x - LC^x \end{bmatrix}}_{=:A^{\xi}(K,L)} \xi_t \\ &+ \underbrace{\begin{bmatrix} B^w & 0 \\ B^w & -L \end{bmatrix}}_{=:B^d(L)} d_t + \underbrace{\begin{bmatrix} B^u(G^u(I - A^x) + K) \\ 0 \end{bmatrix}}_{=:B^s(K)} x^s\end{aligned} \tag{5}$$

where $d_t \coloneqq [w_t^T \ v_t^T]^T \in \mathcal{D}$ and $\mathcal{D} \coloneqq \mathcal{W} \times \mathcal{V}$. For simplicity of presentation, the vector d_t is called the disturbance vector and the set $\mathcal{D}$ is called the disturbance set. Since $\mathcal{W}$ and $\mathcal{V}$ are assumed to be convex polytopes, $\mathcal{D}$ is a convex polytope.

To ensure stable closed-loop behavior, the controller and observer gains are selected such that all eigenvalues of the matrices $A^x - B^u K$ and $A^x - LC^x$ are strictly within the unit circle. Due to the presence of process disturbances and measurement noise, the closed-loop process is persistently perturbed. As a result, the augmented state of the process is ultimately bounded within the minimum invariant set, which is the limit set of all trajectories of the process [20]. The minimum invariant set of the process is given by [21]:

$$\mathcal{R}_\infty^\xi(x^s) = \bigoplus_{i=0}^{\infty} A^\xi(K,L)^i \mathcal{D}^e(x^s) \tag{6}$$

where $\oplus$ represents the Minkowski sum, $\bigoplus_{i=0}^{\infty} F^i\mathcal{X} = \mathcal{X} \oplus F\mathcal{X} \oplus F^2\mathcal{X} \oplus \ldots$ ($\mathcal{X}$ is a set $\mathcal{X} \subseteq \mathbb{R}^n$ and F is a square matrix), and $\mathcal{D}^e(x^s) = B^d(L)\mathcal{D} \oplus B^s(K)\{x^s\}$.

B. Class of False Data Injection Attacks

The process is vulnerable to false data injection (FDI) attacks which alter the output (y_t) transmitted via the sensor-controller link and the input (u_t) conveyed over the controller-actuator link so that the altered values are received by the controller and actuators. Additive and multiplicative FDI attacks are considered where the relationships between the unaltered and altered values may be described as:

$$y_t^a = \Lambda^y y_t + \delta_t^y \tag{7a}$$

$$u_t^a = \Lambda^u u_t + \delta_t^u \tag{7b}$$

If $\Lambda^\theta \neq I$ ($\theta \in \{y, u\}$), the attack alters the data over a communication link by multiplying it with the factor Λ^θ. If $\delta_t^\theta \neq 0$, the attack alters the data over a communication link by adding a bias δ_t^θ. $\theta = y$ represents the sensor-controller link, and $\theta = u$ represents the controller-actuator link. The variables δ_t^y and δ_t^u are assumed to be bounded within a convex polytope, i.e., $\delta_t \coloneqq \begin{bmatrix} \delta_t^u \\ \delta_t^y \end{bmatrix} \in \Delta$ for all $t \in \mathbb{Z}_+$

An attack on the closed-loop process alters the evolution of its augmented state as follows:

$$\begin{aligned}\xi_{t+1} = &\underbrace{\begin{bmatrix} A^x - B^u\Lambda^u K & B^u\Lambda^u K \\ L(I - \Lambda^y)C^x & A^x - LC^x \end{bmatrix}}_{=:A^{\xi_a}(K,L)} \xi_t + \underbrace{\begin{bmatrix} 0 & B^u \\ -L & 0 \end{bmatrix}}_{=:B^{\delta_a}(L)} \delta_t \\ &+ \underbrace{\begin{bmatrix} B^w & 0 \\ B^w & -L\Lambda^y \end{bmatrix}}_{=:B^{d_a}(L)} d_t + \underbrace{\begin{bmatrix} B^u\Lambda^u(G^u(I - A^x) + K) \\ 0 \end{bmatrix}}_{=:B^{s_a}(K)} x^s\end{aligned} \tag{8}$$

From Eq. 8, the process can be destabilized by a multiplicative attack, with $\rho(A^{\xi_a}(K,L)) \coloneqq \max_i |\lambda_i(A^{\xi_a}(K,L))| > 1$, where $\lambda_i(A^{\xi_a}(K,L)$ is ith eigenvalue of the matrix $A^{\xi_a}(K,L)$. However, an additive attack with $\Lambda^u = I$ and $\Lambda^y = I$ cannot destabilize the closed-loop process.

When the process is subjected to an FDI attack, the process is referred to as the attacked process. The term attack-free process is used to describe the closed-loop process without an attack. When the attacked process is stable, the augmented states are ultimately bounded within its minimum invariant set, which is a compact set. The minimum invariant set is:

$$\mathcal{R}_\infty^\xi(x^s) = \bigoplus_{i=0}^{\infty} A^{\xi_a}(K,L)^i \mathcal{D}^a(x^s) \tag{9}$$

where $\mathcal{D}^a(x^s) \coloneqq B^{d_a}\mathcal{D} \oplus B^{\delta_a}\Delta \oplus B^{s_a}(K)\{x^s\}$.

C. Monitoring Variable and Set-Based Detection Scheme

An attack alters the evolution of the augmented state from its expected attack-free evolution. However, since the augmented state cannot be measured directly, detection schemes monitor the evolution of a monitoring variable to detect anomalous behavior. A monitoring variable ($\eta := [y^T \ r^T]^T$) that is a concatenation of the measured output and the residual vector ($r := y - \hat{y}$) is used to monitor the process. For the attack-free process, its monitoring variable may be expressed as a linear combination of the augmented state and the disturbance vectors:

$$\eta_t = \underbrace{\begin{bmatrix} C^x & 0 \\ 0 & C^x \end{bmatrix}}_{=:C^\eta} \xi_t + \underbrace{\begin{bmatrix} 0 & I \\ 0 & I \end{bmatrix}}_{=:D^\eta} d_t \tag{10}$$

For the attacked process, its monitoring variable may be expressed as a linear combination of the augmented state, the disturbance vector, and the attack biases added to the sensor-controller link as:

$$\eta_t = \underbrace{\begin{bmatrix} \Lambda^y C^x & 0 \\ (\Lambda^y - I)C^x & C^x \end{bmatrix}}_{=:C^{\eta_a}} \xi_t + \underbrace{\begin{bmatrix} 0 & \Lambda^y \\ 0 & \Lambda^y \end{bmatrix}}_{=:d_t^a} d_t + \begin{bmatrix} \delta_t^y \\ \delta_t^y \end{bmatrix} \tag{11}$$

where $d_t^a \in \mathcal{D}^{\eta_a} := \begin{bmatrix} 0 & \Lambda^y \\ 0 & \Lambda^y \end{bmatrix} \mathcal{D} \oplus \begin{bmatrix} 0 & I \\ 0 & I \end{bmatrix} \Delta$ for $t \in \mathbb{Z}_+$.

Chemical processes typically operate at steady-state for extended duration, ensuring that the augmented state remains within its minimum invariant set. In the absence of attacks, the monitoring variable is contained within a well-defined set when $\xi_t \in \mathcal{R}_\infty^\xi(x^s)$. From Eq. 9 and Eq. 10, this set, called the terminal set, is represented as:

$$\mathcal{R}_\infty^\eta(\mathcal{R}_\infty^\xi(x^s)) = C^\eta \mathcal{R}_\infty^\xi(x^s) \oplus D^\eta \mathcal{D} \tag{12}$$

The terminal set for the attack-free process encompasses all conceivable values of the monitoring variable across all time steps ($t \in \mathbb{Z}_+$) and under all disturbances ($d_t \in \mathcal{D}$) when $\xi_t \in \mathcal{R}_\infty^\xi(x^s)$. Therefore, the terminal set can be used to verify the integrity of monitoring variable values. To monitor for attacks, a terminal set membership-based detection scheme can be used where the containment of the monitoring variable in the terminal set is verified. If the monitoring variable is outside the set, an alarm is raised and an attack is detected (refer to [18] for more details).

If the process operates for a sufficiently long period after an attack and the closed-loop process is stable, the augmented state will converge to the minimum invariant set under the attack ($\mathcal{R}_\infty^{\xi_a}(x^s)$). For analysis purposes, the corresponding terminal set of the monitoring variable can be computed from $\mathcal{R}_\infty^{\xi_a}(x^s)$, given by:

$$\mathcal{R}_\infty^{\eta_a}(\mathcal{R}_\infty^{\xi_a}(x^s)) = C^{\eta_a} \mathcal{R}_\infty^{\xi_a}(x^s) \oplus \mathcal{D}^{\eta_a} \tag{13}$$

III. Active Detection for Stealthy Attacks

The terminal set-based detection scheme described above is passive, as it monitors the process for attacks without utilizing any external interventions or perturbations. An attacker may be able to carry out stealthy attacks capable of evading detection. Stealthy attacks with respect to the passive terminal set-based detection scheme are attacks where the monitoring variable is maintained within its expected terminal set ($\eta_t \in \mathcal{R}_\infty^\eta(\mathcal{R}_\infty^\xi(x^s))$). As a result, the terminal set-based detection scheme will fail to detect the attack.

An active detection method involving operating under the attack-sensitive mode could be applied to detect stealthy attacks that destabilize the process [18], [19]. However, operating in a mode that allows some attacks to destabilize the process may be undesirable. In this work, two active detection methods are considered: changing the control parameters and the set point, which define alternative operating modes of the control system to enable the detection of stealthy attacks. A framework for evaluating if an attack is guaranteed to be detected under this active method is developed for these control modes.

A. Alternative Active Detection without Destabilization

The process after extended operation near the initial set point x_i^s under control parameters (K^i, L^i) is considered so that the augmented state has converged to either $\mathcal{R}_\infty^\xi(x_i^s)$ or $\mathcal{R}_\infty^{\xi_a}(x_i^s)$, depending on whether the process is attack-free or subjected to an attack. Under the active detection method, the set point changes from $x^s = x_i^s$ to $x^s = x_f^s$, and/or the control parameters switch from $(K, L) = (K^i, L^i)$ to $(K, L) = (K^f, L^f)$. The time step in which this change occurs is taken to be $t = 0$. The initial set point and control parameters (K^i, L^i) are selected based on the desired operating set point and standard controller design methods. However, the final set point and control parameters are selected to enable the detection of a given attack, defined by Λ^u, Λ^y, and Δ, that is stealthy with respect to the terminal set-based detection scheme. The attack-free and attacked process are assumed to be stable under both sets of control parameters in the sense that $\rho(A^{\xi_a}(K^i, L^i)) < 1$ and $\rho(A^{\xi_a}(K^f, L^f)) < 1$.

The change(s) perturb(s) the process by exciting the process dynamics, in the sense that after the change(s), the augmented state may be outside its corresponding minimum invariant set. From Eq. 10 and Eq. 11, the values of the monitoring variable depend upon the augmented state, the disturbances acting on the process, and the attack (if the process is attacked). During the transient period (when the state is outside the minimum invariant set), the monitoring variable may take values outside its terminal set. For such cases, the terminal set-based detection scheme will raise an alarm, even for the attack-free process, since the scheme does not account for such transient behavior. A reachable set-based detection scheme can monitor the process during the transient period without generating false alarms [22]. During the transient period, the possible states reached for the attack-free and attacked process are described by the reachable sets of the process. The reachable sets of the attack-free and the attacked process are:

$$\mathcal{R}_t^\xi(x_f^s) = A^\xi(K, L)\mathcal{R}_{t-1}^\xi(x_f^s) \oplus \mathcal{D}^e(x_f^s) \tag{14a}$$

$$\mathcal{R}_t^{\xi_a}(x_f^s) = A^{\xi_a}(K, L)\mathcal{R}_{t-1}^{\xi_a}(x_f^s) \oplus \mathcal{D}^a(x_f^s) \tag{14b}$$

for $t > 0$ where, with slight abuse of notation, the initial sets are $\mathcal{R}_0^{\xi} = \mathcal{R}_\infty^{\xi}(x_i^s)$ and $\mathcal{R}_0^{\xi_a} = \mathcal{R}_\infty^{\xi_a}(x_i^s)$. Eq. 14a describes the evolution of the reachable sets for the attack-free augmented states from an initial set of states that is the minimum invariant set of the attack-free process at the initial steady-state. Eq. 14b describes the evolution of the reachable sets for the augmented states of the attacked process from an initial set of states that is the minimum invariant set of the attacked process at the initial steady-state. From Eq. 10 and Eq. 11, the reachable sets of the monitoring variables describe their evolution for the attacked and the attack-free processes as (for $t > 0$):

$$\mathcal{R}_t^{\eta}(\mathcal{R}_t^{\xi}(x_f^s)) = C^{\eta}\mathcal{R}_t^{\xi}(x_f^s) \oplus D^{\eta}\mathcal{D} \tag{15a}$$

$$\mathcal{R}_t^{\eta_a}(\mathcal{R}_t^{\xi_a}(x_f^s)) = C^{\eta_a}\mathcal{R}_t^{\xi_a}(x_f^s) \oplus \mathcal{D}^{\eta_a} \tag{15b}$$

For the attack-free process, its monitoring variable values evolve within its reachable sets. A reachable set-based detection scheme designed to utilize the reachable sets for the attack-free process to monitor the process for attacks can be used [22], given by:

$$\phi_t(\eta_t) = \begin{cases} 0, \ \eta_t \in \mathcal{R}_t^{\eta}(\mathcal{R}_t^{\xi}(x_f^s)) \\ 1, \ \eta_t \notin \mathcal{R}_t^{\eta}(\mathcal{R}_t^{\xi}(x_f^s)) \end{cases} \tag{16}$$

where $\phi_t(\eta_t)$ is the output of the detection scheme at the time step $t > 0$. The detection scheme generates an output of 1 if the monitoring variable is not contained within its attack-free reachable set, meaning that an attack is detected. However, if the monitoring variable is contained within the attack-free reachable set, then the detection scheme generates an output of 0 indicating a lack of attack detection.

B. Selecting an Attack-Revealing Control Mode for Active Detection

From Eq. 15a and Eq. 15b, the monitoring variable values for the attack-free and the attacked processes are contained within their respective reachable sets. If at some time, the reachable sets of the monitoring variable for the attacked and the attack-free processes at that time step do not intersect, then there exist no values of monitoring variable values of the attacked process, that are contained within the reachable set of the attack-free process. It follows from this reasoning that the perturbation induced by switching the control mode is attack-revealing if, at some time step $t > 0$, the reachable sets of the monitoring variable for the attacked and the attack-free process satisfy:

$$\mathcal{R}_t^{\eta}(\mathcal{R}_t^{\xi}(x_f^s)) \cap \mathcal{R}_t^{\eta_a}(\mathcal{R}_t^{\xi_a}(x_f^s)) = \emptyset \tag{17}$$

The reachable set-based detection scheme in Eq. 16 monitors the process based on the reachable sets for the attack-free process, meaning that attack detection is guaranteed at the time step t if the perturbation induced is attack-revealing, i.e., if Eq. 17 is satisfied.

In the discussion that follows, a screening algorithm that leverages Eq. 17 to enable the selection of a control mode that guarantees attack detection is presented. The algorithm is implemented offline and requires that the reachable sets for the attacked and the attack-free processes operated under a given control mode be computed, and the satisfaction of Eq. 17 be checked. If at some time step $t_d \in \mathbb{Z}_+$, the reachable sets of the attack-free and the attacked processes satisfy Eq. 17, then the control mode chosen induces attack-revealing perturbations in that it guarantees the attack detection. However, if Eq. 17 is never satisfied, then the control mode induced does not guarantee attack detection.

First, a practical implementation challenge is discussed. Ensuring the satisfaction of Eq. 17 requires computing reachable sets for both the attack-free and attacked processes, potentially extending to an infinite number of time steps, which is infeasible. The algorithm must strike a balance between checking a finite number of reachable sets and the computational complexity. More specifically, a possibility exists that the condition in Eq. 17 is satisfied for some time step after the algorithm is terminated. To manage this tradeoff, a parameter $t_f > 0$ is introduced, which is the number of time steps to compute the reachable sets before terminating the algorithm. Opting for a large t_f may reduce the possibility that Eq. 17 is satisfied for some time after t_f but may increase computational demands; selecting a small t_f may heighten this possibility but may reduce computation. This is grounded in the understanding that, given an error threshold, there exists a time duration large enough for the reachable sets to converge to an invariant set containing the minimum invariant set. The discrepancy between the invariant set and the true minimum invariant set depends on the chosen error threshold [21, Theorem 1]. On a more practical level, choosing t_f could involve selecting the number of time steps at which it becomes essential to detect the attack, especially if operating under the alternative control mode for prolonged periods is undesirable. With these considerations, the algorithm is as follows:

Algorithm 1: Algorithm to screen an active detection method for its ability to guarantee attack detection

Inputs: Λ^y, Λ^u, Δ, (K^f, L^f), x_f^s, t_f, $\mathcal{R}_t^{\xi}(x_f^s)$ and $\mathcal{R}_t^{\xi_a}(x_f^s)$ for $t \in (0, t_f]$.

Initialization: $t = 0$, $\mathcal{R}_0^{\xi} = \mathcal{R}_\infty^{\xi}(x_i^s)$, $\mathcal{R}_0^{\xi_a} = \mathcal{R}_\infty^{\xi_a}(x_i^s)$, $t_d = \infty$, $(K, L) = (K^f, L^f)$

do
- Compute reachable sets per Eq. 15a and Eq. 15b.
- **if** *Eq. 17 is satisfied* **then**
  - The chosen control mode guarantees attack detection at $t_d = t$.
- **else if** $t = t_f$ **then**
  - The chosen control mode does not guarantee attack detection.
- **else**
  - Set $t \leftarrow t + 1$

while $t_d = \infty$ *or* $t < t_f$;

IV. Application to an Illustrative Process

A process under a simultaneous sensor-controller link and controller-actuator link FDI attack is considered:

$$\begin{aligned} x_{t+1} &= x_t + u_t^a + w_t \\ u_t^a &= -\Lambda^u K(\hat{x}_t - x^s) + \delta_t^u \\ y_t^a &= \Lambda^y(x_t + v_t) + \delta_t^y \end{aligned}$$

where $x_t \in \mathbb{R}$ is the state, $u_t^a \in \mathbb{R}$ is the control action received by the control actuators, $w_t \in \mathcal{W} := \{w' \mid |w'| \leq 1\}$ is the process disturbance, $y_t^a \in \mathbb{R}^m$ is the measured output received by the controller, and $v_t \in \mathcal{V} := \{v' \mid |v'| \leq 1\}$ is the measurement noise. For this integrating process, there is an equilibrium manifold corresponding to the steady-state input $u^s = 0$. The initial steady-state is the origin, i.e., $x_i^s = 0$. The control parameters chosen to operate the process at the initial steady-state are $(K^i, L^i) = (0.8541, 0.618)$. An attack with $\Lambda^y = 0.86$, $\Lambda^u = 1.1$, $\delta_t^y = 0.1$, and $\delta_t^u = -0.028$ is considered. The MPT 3.0 toolbox is used for polytope computations [23].

To verify the detectability characteristics of the attack across the detection methods and schemes considered, 1000 simulation scenarios, each scenario spanning 100 time steps, are considered. Within each scenario, an initial condition is randomly selected from the minimum invariant set of the attacked process ($\mathcal{R}_\infty^{\xi_a}(x_i^s)$). To simulate process disturbances and measurement noise, random sequences with each element drawn from $\mathcal{N}(0, 3.33 \times 10^{-2})$ are generated.

The detectability properties of this attack under the terminal set-based detection scheme are first investigated. To analyze the detectability of the attack under the detection scheme, the terminal sets of the monitoring variable for the attack-free and the attacked process are computed. The terminal set of the attacked process is contained entirely within the terminal set of the attack-free process, i.e., $\mathcal{R}_\infty^{\eta_a}(\mathcal{R}_\infty^{\xi_a}(x_i^s)) \subset \mathcal{R}_\infty^{\eta}(\mathcal{R}_\infty^{\xi}(x_i^s))$. This implies that the attack is undetectable, i.e., stealthy with respect to the detection scheme because for any $\xi_t \in \mathcal{R}_\infty^{\xi_a}(x_i^s)$, $\eta_t \in \mathcal{R}_\infty^{\eta_a}(\mathcal{R}_\infty^{\xi_a}(x_i^s)) \subset \mathcal{R}_\infty^{\eta}(\mathcal{R}_\infty^{\xi}(x_i^s))$. To verify the undetectability of the attack, 1000 closed-loop simulations of the attacked process are performed. The attack is not detected in any of these simulations. One active detection approach that can enable the detection of this attack is to change the control parameters to so-called attack-sensitive parameters, where the closed-loop process with these parameters is stable under attack-free operation and is destabilized under the attack [18], [19]. However, destabilization for attack detection may be undesirable, and alternative active detection methods are considered. To check if a particular active detection method guarantees attack detection, Algorithm 1 is applied and implemented.

The method described in [20] is used to compute an outer approximation of the minimum invariant sets of the attack-free and attacked processes for the initial steady-state and control parameters and the final steady-state and control parameters, i.e., the sets $\mathcal{R}_\infty^{\xi}(x_i^s)$, $\mathcal{R}_\infty^{\xi_a}(x_i^s)$, $\mathcal{R}_\infty^{\xi}(x_f^s)$, and $\mathcal{R}_\infty^{\xi_a}(x_f^s)$. The specified error bound on these calculations is 1×10^{-3}. To determine the termination time of the algorithm (t_f), the reachable sets of the attack and attack-free process with initial sets $\mathcal{R}_\infty^{\xi}(x_i^s)$ and $\mathcal{R}_\infty^{\xi_a}(x_i^s)$ are computed until the reachable sets are contained within the outer approximation of the attack-free and attacked minimum invariant sets for the final steady-state and control parameters. Defining t_1 and t_2 as the time steps at which the attack-free and attacked reachable sets are first contained with their corresponding minimum invariant sets, respectively, t_f is taken to be $\max(t_1, t_2)$. The satisfaction of Eq. 17 is verified by checking for the existence of a point satisfying both sets of inequalities describing the two reachable sets of the monitoring variable for the attack-free and the attacked processes. Specifically, a feasibility problem, cast as a linear program, is constructed and solved for all $t \in [0, t_f]$.

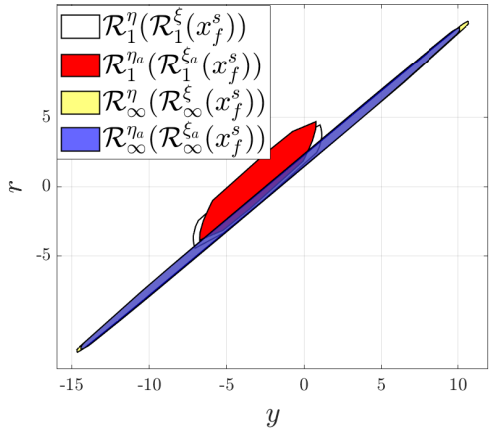

Fig. 1: Reachable sets and the terminal sets of the monitoring variable at $t = 1$ for the attacked and the attack-free process under a control mode with $x_f^s = -2$ and $(K^f, L^f) = (1.5, 0.1)$.

The first alternative active detection method considered utilizes a set point change to shift the operation of the process to a neighborhood of the steady-state $x_f^s = -2$ and with the control parameters $(K^f, L^f) = (1.5, 0.1)$. Under these control parameters, the eigenvalues of the closed-loop attacked process are -0.67 and 0.92, indicating that the closed-loop process is stable in the presence of attack. In this case, $t_1 = 53$ and $t_2 = 75$, so $t_f = 75$. Using Algorithm 1, Eq. 17 is not satisfied for any time step, and therefore, attack detection is not guaranteed. Fig. 1 illustrates the reachable sets of monitoring variable for the attacked and the attack-free processes at the time step $t = 1$, and the terminal sets of the monitoring variable values for the attacked and the attack-free processes under the chosen active detection method, showing that Eq. 17 is not satisfied because the sets always intersect. To verify the detection properties of the attack under the first alternative active detection method, 1000 closed-loop simulations of the attacked process monitored by the reachable set-based detection scheme in Eq. 16 are performed under the active detection method. The attack is detected in 114 simulations. For the simulations where the attack is detected, the attack is detected at either time step 1 or 2. The results demonstrate that while the attack may be

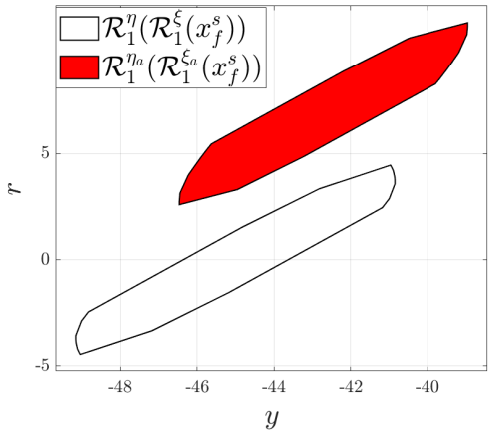

Fig. 2: Reachable sets of the monitoring variable for the attacked and the attack-free process at $t_d = 1$ under a control mode with $x_f^s = -30$ and $(K^f, L^f) = (1.5, 0.1)$.

detected with this active detection method, detection is not guaranteed.

A second active detection method using a set point change to $x_f^s = -30$, and a control parameter switch to $(K^f, L^f) = (1.5, 0.1)$ is considered. The reachable sets of the augmented state for the attacked and the attack-free processes are computed and the termination time step for Algorithm 1 is determined to be as $t_f = \max(t_1, t_2) = 119$, with $t_1 = 92$ and $t_2 = 119$. Algorithm 1 is applied, and the control mode chosen is determined to guarantee attack detection at the time step $t_d = 1$. Fig. 2 illustrates that the reachable sets for the attacked and the attack-free processes do not intersect at the time step $t_d = 1$, satisfying Eq. 17. One thousand simulations of the attacked process under this active detection method and monitored by the reachable set-based detection scheme are performed. The attack is detected in all simulations at the time step $t_d = 1$, demonstrating that the active detection method chosen guarantees attack detection.

V. Conclusions

Two control modes for the active detection of a class of stealthy false data injection cyberattacks were presented. Reachability analysis was used to present a screening algorithm that may be used to select an active detection method that guarantees attack detection. The application of the screening algorithm was demonstrated using an illustrative process example.

References

[1] Cybersecurity & Infrastructure Security Agency, "Critical infrastructure sectors," Tech. Rep., 2023, accessed August 2023. [Online]. Available: https://www.cisa.gov/topics/critical-infrastructure-security-and-resilience/critical-infrastructure-sectors

[2] T. Alladi, V. Chamola, and S. Zeadally, "Industrial control systems: Cyberattack trends and countermeasures," *Computer Communications*, vol. 155, pp. 1–8, 2020.

[3] S. Parker, Z. Wu, and P. D. Christofides, "Cybersecurity in process control, operations, and supply chain," *Computers & Chemical Engineering*, vol. 171, p. 108169, 2023.

[4] Y. Hu, A. Yang, H. Li, Y. Sun, and L. Sun, "A survey of intrusion detection on industrial control systems," *International Journal of Distributed Sensor Networks*, vol. 14, no. 8, p. 1550147718794615, 2018.

[5] J. Giraldo, D. Urbina, Á. A. Cárdenas, J. Valente, M. Faisal, J. Ruths, N. O. Tippenhauer, H. Sandberg, and R. Candell, "A survey of physics-based attack detection in cyber-physical systems," *ACM Computing Surveys*, vol. 51, no. 4, pp. 1–36, 2018.

[6] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "Revealing stealthy attacks in control systems," in *Proceedings of the 50th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, Illinois, USA, 01-05 October 2012, pp. 1806–1813.

[7] C. Murguia and J. Ruths, "Characterization of a CUSUM model-based sensor attack detector," in *Proceedings of the IEEE 55th Conference on Decision and Control*, Las Vegas, NV, 12-14 December 2016, pp. 1303–1309.

[8] ——, "CUSUM and chi-squared attack detection of compromised sensors," in *Proceedings of the IEEE Conference on Control Applications*, Buenos Aires, Argentina, 19-22 September 2016, pp. 474–480.

[9] S. Chen, Z. Wu, and P. D. Christofides, "A cyber-secure control-detector architecture for nonlinear processes," *AIChE Journal*, vol. 66, no. 5, p. e16907, 2020.

[10] M. Kravchik and A. Shabtai, "Detecting cyber attacks in industrial control systems using convolutional neural networks," in *Proceedings of the Workshop on Cyber-physical Systems Security and Privacy*, Toronto, Canada, 15-19 October 2018, pp. 72–83.

[11] K. Guibene, N. Messai, M. Ayaida, and L. Khoukhi, "A pattern mining-based false data injection attack detector for industrial cyber-physical systems," *IEEE Transactions on Industrial Informatics*, pp. 1–10, 2023.

[12] C. Trapiello and V. Puig, "Input design for active detection of integrity attacks using set-based approach," in *Proceedings of the 21st IFAC World Congress*, Berlin, Germany, 11-17 July 2020, pp. 11 094–11 099.

[13] H. Oyama, D. Messina, K. K. Rangan, F. L. Akkarakaran, K. Nieman, H. Durand, K. Tyrrell, K. Hinzman, and W. Williamson, "Development of directed randomization for discussing a minimal security architecture," *Digital Chemical Engineering*, vol. 6, p. 100065, 2023.

[14] M. Attar and W. Lucia, "An active detection strategy based on dimensionality reduction for false data injection attacks in cyber-physical systems," *IEEE Transactions on Control of Network Systems*, pp. 1–11, 2023.

[15] N. Babadi and A. Doustmohammadi, "A moving target defence approach for detecting deception attacks on cyber-physical systems," *Computers and Electrical Engineering*, vol. 100, p. 107931, 2022.

[16] Y. Hu, P. Xun, P. Zhu, Y. Xiong, Y. Zhu, W. Shi, and C. Hu, "Network-based multidimensional moving target defense against false data injection attack in power system," *Computers & Security*, vol. 107, p. 102283, 2021.

[17] M. Ghaderi, K. Gheitasi, and W. Lucia, "A blended active detection strategy for false data injection attacks in cyber-physical systems," *IEEE Transactions on Control of Network Systems*, vol. 8, no. 1, pp. 168–176, 2021.

[18] S. Narasimhan, N. H. El-Farra, and M. J. Ellis, "Active multiplicative cyberattack detection utilizing controller switching for process systems," *Journal of Process Control*, vol. 116, pp. 64–79, 2022.

[19] ——, "A control-switching approach for cyberattack detection in process systems with minimal false alarms," *AIChE Journal*, vol. 68, no. 12, p. e17875, 2022.

[20] S. V. Raković, E. C. Kerrigan, K. I. Kouramas, and D. Q. Mayne, "Invariant approximations of the minimal robust positively invariant set," *IEEE Transactions on Automatic Control*, vol. 50, no. 3, pp. 406–410, 2005.

[21] V. M. Kuntsevich and B. N. Pshenichnyi, "Minimal invariant sets of dynamic systems with bounded disturbances," *Cybernetics and Systems Analysis*, vol. 32, no. 1, pp. 58–64, 1996.

[22] S. Narasimhan, N. H. El-Farra, and M. J. Ellis, "A reachable set-based scheme for the detection of false data injection cyberattacks on dynamic processes," *Digital Chemical Engineering*, vol. 7, p. 100100, 2023.

[23] M. Herceg, M. Kvasnica, C. N. Jones, and M. Morari, "Multi-Parametric Toolbox 3.0," in *Proceedings of the European Control Conference*, Zürich, Switzerland, July 17–19 2013, pp. 502–510.