SCIENCE CHINA
Information Sciences

• **Supplementary File** •

Attack detectability and stealthiness in distributed optimal coordination of cyber-physical systems

Liwei An[1] & Guang-Hong Yang[1,2*]

[1]*College of Information Science and Engineering, Northeastern University, Shenyang, P.R.China;*
[2]*State Key Laboratory of Synthetical Automation for Process Industries (Northeastern University), Shenyang, P.R.China*

Appendix A Comparisons with the related works

• *Comparisons with results on fault detection and isolation (FDI):* In the existing literature, important works on distributed/decentralized sensor fault diagnosis for large-scale systems have been reported, such as [1–3]. In practice, the fault signal is usually independent from the system structure and system behavior such that the existing FDI methods can effectively detect and isolate the fault signal. In contrast to the system fault, the attack signal can exploit the vulnerability of the detection mechanism to damage or destroy the function of the system without being detected. As reported by [8], the classical residual-based fault detection filters, for instance those presented in [4], cannot be guaranteed to detect and isolate the signals that do excite exclusively zero dynamics (i.e., zero-dynamics attacks). Recently, various stealthy attack strategies against different types of detection mechanisms have been proposed, e.g., [5–7] and references therein. In our problem formulation we consider the malicious attack that aims to destroy the DOC of the system without being detected.
• *Comparisons with results on attack detection and identification (FDI):* Some effective distributed ADI methods for interconnected systems have also been proposed in the existing literature [10, 11]. However, the considered system models, attack models and control objectives considerably differ from the ones studied in this paper.

System models: In [10, 11], linear interconnected system models are considered. In this paper, the uncertain nonlinear strict-feedback system model is considered.

Attack models: In [10, 11], the attack models are assume to have no global knowledge of interconnected system or knowledge of the detectors. In this paper, we allow the adversarial attacker to know the overall system model, system state, control input and the possible detector.

Control objective: In [10, 11] the control objective is to stabilize the closed-loop system. In this paper, the control objective is to steer the physical system converge to the optimal solution.

Adapted techniques: In [10], the strong observability of local subsystem is used to detect and identify the attacks, and in [11] the stochastic hypothesis testing method is used to detect the attacks. In this paper, the double coupling residuals and "strongly-robust" thresholds based on the prescribed performance technique are proposed to detect and identify the attacks.

Appendix B DOC structure description

The DOC architecture is a double-layer structure illustrated in Figure B1. Similar hierarchical architectures can also be found in [3]. The cyber part $\mathcal{C}^{(j)}$ consists of a decision-making network. Each decision-making agent, denoted by $\mathcal{D}^{(j)}$, is responsible for sending the control command $y_r^{(j)}$ to $\Sigma^{(j)}$. Agent $\mathcal{D}^{(j)}$ contains an optimization module and a monitoring module, denoted by $\mathcal{O}^{(j)}$ and $\mathcal{M}^{(j)}$, respectively. Module $\mathcal{O}^{(j)}$ is used to optimize its local objective function $g^{(j)}$, while exchanging its output $y^{(j)}$ with its neighbors under an undirected network topology $\mathcal{G}$. Since the attackers can corrupt the transmitted output, the effect of the attack on $(\mathcal{P}^{(i)}, \mathcal{C}^{(i)})$, $i \in \mathbf{N}_j$ can also be propagated to $(\mathcal{P}^{(j)}, \mathcal{C}^{(j)})$ via information exchanges, where $\mathbf{N}_j$ denotes the set of neighbors of node j. Note that for the jth subsystem, $a^{(i)}$, $i \in \mathbf{N}_j$ is leveraged against the output of its neighbor i, rather than an attack on the information exchanged between these two subsystems. Each module $\mathcal{M}^{(j)}$ is used to detect and identify the local attack $a^{(j)}$ by monitoring module $\mathcal{O}^{(j)}$ and transmit the detection logic $\mathfrak{V}^{(j)}$ to $\mathcal{O}^{(j)}$. In $\mathcal{P}^{(j)}$, control module $\mathcal{K}^{(j)}$ drives dynamics $\Sigma^{(j)}$ in accordance with the control command $y_r^{(j)}$ coming from $\mathcal{D}^{(j)}$. Figure 1 illustrates *sufficient interactions* between cyber and physical parts in the CPS architecture.

* Corresponding author (email: yangguanghong@ise.neu.edu.cn)

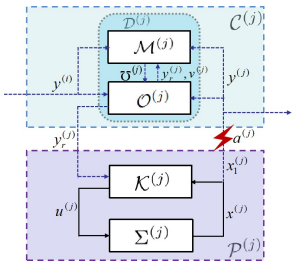

Figure B1 Secure DOC architecture under cyber attacks.

Appendix C Controller design and stability analysis

First, we present the design of control agent $\mathcal{K}^{(j)}$ (i.e., specify (4)) which consists of inner-loop module $\mathcal{K}^{(I,j)}$ and outer-loop module $\mathcal{K}^{(O,j)}$. We define the following changes of coordinates

$$z_1^{(j)} = x_1^{(j)} - y_r^{(j)}, \quad z_i = x_i^{(j)} - \alpha_{i-1}^{(j)}, \quad i = 2, \cdots, n \tag{C1}$$

where $\alpha_i^{(j)} = \alpha_{i,I}^{(j)} + \alpha_{i,O}^{(j)}$ is the virtual control function determined at the ith step and spitted into two parts: inner-loop control $\alpha_{i,I}^{(j)}$ and outer-loop control $\alpha_{i,O}^{(j)}$. Similarly, $u^{(j)} = u_I^{(j)} + u_O^{(j)}$ is the control input consisting of inner-loop control $u_I^{(j)}$ and outer-loop control $u_O^{(j)}$. They can be respectively expressed as follows:

• Inner-loop control $\mathcal{K}^{(I,j)}(x^{(j)}, \Im^{(j)})$

$$\alpha_{1,I}^{(j)} = -c_1^{(j)} z_1^{(j)} - \hat{\rho}^{(j)} z_1^{(j)} - \omega_1^{(j)} \hat{\lambda}^{(j)} + \hat{\pi}^{(j)} - S\left(\frac{z_1^{(j)}}{\delta^{(j)}}\right)$$

$$\alpha_{i,I}^{(j)} = -z_i^{(j)} - c_i^{(j)} z_i^{(j)} - \omega_i^{(j)} \hat{\lambda}^{(j)} + \Lambda_i^{(j)} \tag{C2}$$

$$u_I^{(j)} = \beta_j^{-1}\left[-z_n^{(j)} - c_n^{(j)} z_n^{(j)} - \omega_n^{(j)} \hat{\lambda}^{(j)} + \Lambda_n^{(j)}\right] \tag{C3}$$

The corresponding update laws are given as

$$\dot{\hat{\lambda}}^{(j)} = \Gamma^{(j)} \tau_n^{(j)} \tag{C4}$$

$$\dot{\rho}^{(j)} = \gamma_0^{(j)} \|z_1^{(j)}\|^2 \tag{C5}$$

$$\dot{\pi}^{(j)} = \gamma_1^{(j)} z_1^{(j)} \tag{C6}$$

where

$$\tau_1^{(j)} = \omega_1^{(j)} z_1^{(j)}, \ \tau_i^{(j)} = \tau_{i-1}^{(j)} + \omega_i^{(j)} z_i^{(j)}$$

$$\omega_i^{(j)} = \psi_i^{(j)} - \sum_{k=1}^{i-1} \frac{\partial \alpha_{i-2,I}^{(j)}}{\partial x_k^{(j)}} \psi_k^{(j)}$$

$$\Lambda_i^{(j)} = \sum_{k=1}^{i-1} \frac{\partial \alpha_{i-1,I}^{(j)}}{\partial x_k^{(j)}} x_{k+1}^{(j)} + \frac{\partial \alpha_{i-1,I}^{(j)}}{\partial \hat{\lambda}^{(j)}} \Gamma^{(j)} \tau_i^{(j)}$$
$$+ \sum_{k=2}^{i-1} \frac{\partial \alpha_{k-1,I}^{(j)}}{\partial \hat{\lambda}^{(j)}} \Gamma^{(j)} w_i^{(j)} z_k^{(j)} + \frac{\partial \alpha_{i-1,I}^{(j)}}{\partial \hat{\pi}^{(j)}} \dot{\hat{\pi}}^{(j)}$$

$$S\left(\frac{z_1^{(j)}}{\delta^{(j)}}\right) = \frac{1}{2} \ln\left(1 + \frac{z_1^{(j)}}{\delta^{(j)}}\right) - \frac{1}{2} \ln\left(1 - \frac{z_1^{(j)}}{\delta^{(j)}}\right)$$

with $\delta^{(j)}(t)$ being an exponentially decaying function with lower bound $k_b^{(j)}$ such that $|z_{1,s}^{(j)}(0)| < \delta^{(j)}(0)$, and $z_{1,s}^{(j)}$ denotes the sth $(s = 1, \cdots, m)$ element of $z_1^{(j)}$; $\psi_1^{(j)} = \text{diag}\{\varphi_1^{(j)}(x_1^{(j)}), 0\}$, $\psi_i^{(j)} = \text{diag}\{\varphi_i^{(j)}(\bar{x}_i^{(j)}), z_i^{(j)}\}$ for $i = 2, \cdots, N$; and $\hat{\lambda}^{(j)}$, $\hat{\rho}^{(j)}$ and $\hat{\pi}^{(j)}$ are the estimates of $\lambda^{(j)} =: [\theta_j^T, \mu]^T$ with $\mu =: ((1+\eta)^2\|L\| + \|L\|^3)\Pi^2/2$, $\rho^{(j)} := (2n-1+\eta)\|L\|$ and $\pi^{(j)} =: \sum_{i \in \mathbf{N}_j} (v_*^{(j)} - v_*^{(i)})$, respectively, where Π is defined in the appendix; $\Gamma^{(j)}$ is a positive definite matrix and $\gamma_0^{(j)}$, $\gamma_1^{(j)}$ and $c_i^{(j)}$ for $i = 1, \cdots, n$ are positive constants, all chosen by users.

• Outer-loop control $\mathcal{K}^{(O,j)}(\Im^{(j)})$

$$\alpha_{1,O}^{(j)} = -\nabla g^{(j)}(y_r^{(j)}) - 2\tilde{v}^{\mathbf{N}_j} \tag{C7}$$

$$\alpha_{i,O}^{(j)} = -\frac{\partial \alpha_{i-1,O}^{(j)}}{\partial y_r^{(j)}} \left[\nabla g^{(j)}(y_r^{(j)}) + \tilde{v}^{\mathbf{N}_j}\right] \tag{C8}$$

$$u_O^{(j)} = -\beta_j^{-1} \frac{\partial \alpha_{n-1,O}^{(j)}}{\partial y_r^{(j)}} \left[\nabla g^{(j)}(y_r^{(j)}) + \tilde{v}^{\mathbf{N}_j}\right] \tag{C9}$$

Summarizing the above procedure (C1)-(C9), we derive Algorithm 1 for the DOC of the overall CPS under healthy environment.

Algorithm 1: DOC under healthy environment

DO algorithm (Module $\mathcal{O}(y)$):

$$\begin{aligned}\dot{y}_r =& -\nabla g(y_r) - Lv - (1+\eta)Ly\\ \dot{v} =& Ly\\ \Im =& (y_r, \nabla g(y_r), \tilde{v})\end{aligned} \tag{C10}$$

where $\nabla g(y_r) = \mathrm{vec}(\nabla g^{(1)}(y_r^{(1)}), \cdots, \nabla g^{(N)}(y_r^{(N)}))$ and $\tilde{v} = \mathrm{vec}(\tilde{v}^{\mathbf{N}_1}, \cdots, \tilde{v}^{\mathbf{N}_N})$.

Adaptive tracking control (Module $\mathcal{K}(x, \Im)$):

Inner-loop control $\mathcal{K}^I(x, \Im)$:

$$\alpha_{1,I} = -C_1 z_1 - \hat{\rho} z_1 - \omega_1 \hat{\lambda} + \hat{\pi} - S\left(\frac{z_1}{\delta}\right) \tag{C11}$$

$$\begin{aligned}\alpha_{i,I} =& -z_i - C_i z_i - \omega_i \hat{\lambda} + \sum_{k=1}^{i-1} \frac{\partial \alpha_{i-1}}{\partial x_k} x_{k+1}\\ &+ \frac{\partial \alpha_{i-1}}{\partial \hat{\lambda}} \Gamma \tau_i + \sum_{k=2}^{n-1} \frac{\partial \alpha_{k-1}}{\partial \hat{\lambda}} \Gamma w_i z_k + \frac{\partial \alpha_{i-1}}{\partial \hat{\pi}} \dot{\hat{\pi}}\end{aligned} \tag{C12}$$

$$\begin{aligned}u_I =& B^{-1}\left[-z_n - C_n z_n - \omega_n \hat{\lambda} + \sum_{k=1}^{n-1} \frac{\partial \alpha_{n-1}}{\partial x_k} x_{k+1}\right.\\ &\left. + \frac{\partial \alpha_{n-1}}{\partial \hat{\lambda}} \Gamma \tau_n + \sum_{k=2}^{n-1} \frac{\partial \alpha_{k-1}}{\partial \hat{\lambda}} \Gamma w_i z_k + \frac{\partial \alpha_{n-1}}{\partial \hat{\pi}} \dot{\hat{\pi}}\right]\end{aligned} \tag{C13}$$

where $C_i = \mathrm{diag}\{c_i^{(1)}, \cdots, c_i^{(N)}\}$, $\hat{\rho} = \mathrm{diag}\{\hat{\rho}^{(1)}, \cdots, \hat{\rho}^{(N)}\}$, $\Gamma = \mathrm{diag}\{\Gamma^{(1)}, \cdots, \Gamma^{(N)}\}$, $B = \mathrm{diag}\{\beta_1, \cdots, \beta_N\}$, $S(z_1/\delta) = [S(z_1^{(1)}/\delta^{(1)}), \cdots, S(z_1^{(N)}/\delta^{(N)})]$, $\psi_i = \mathrm{diag}\{\psi_i^{(1)}, \cdots, \psi_i^{(N)}\}$, $\omega_i = \mathrm{diag}\{\omega_i^{(1)}, \cdots, \omega_i^{(N)}\}$ and

$$\begin{aligned}\tau_1 &= \omega_1 z_1\\ \tau_i &= \tau_{i-1} + \omega_i z_i\\ \omega_i &= \psi_i - \sum_{k=1}^{i-1} \frac{\partial \alpha_{i-2}}{\partial x_k} \psi_k\end{aligned}$$

Outer-loop control $\mathcal{K}^O(\Im)$:

$$\alpha_{1,O} = -\nabla g(y_r) - 2Lv \tag{C14}$$

$$\alpha_{i,O} = -\frac{\partial \alpha_{i-1,O}}{\partial y_r}\left[\nabla g(y_r) + Lv\right] \tag{C15}$$

$$u_O = -B^{-1}\frac{\partial \alpha_{n-1,O}}{\partial y_r}\left[\nabla g(y_r) + Lv\right] \tag{C16}$$

where $i = 2, \cdots, n-1$ and $\Gamma_0 = \mathrm{diag}\{\gamma_0^{(1)}, \cdots, \gamma_0^{(N)}\}$.

Update laws:

$$\dot{\lambda} = \Gamma \tau_n \tag{C17}$$

$$\dot{\rho} = \Gamma_0 z_1 \circ z_1 \tag{C18}$$

$$\dot{\pi} = \Gamma_1 z_1 \tag{C19}$$

where $\Gamma_0 = \mathrm{diag}\{\gamma_0^{(1)}, \cdots, \gamma_0^{(N)}\}$ and $\Gamma_1 = \mathrm{diag}\{\gamma_1^{(1)}, \cdots, \gamma_1^{(N)}\}$.

Now, we give the convergence analysis on the cyber dynamics (3) and physical dynamics (1) based on the Lyapunov method, respectively.

Cyber dynamics: Let $y^* = 1_N \otimes y^\star$ be a solution of (2). From [12, Proposition 3.2], there exists $v^* \in \mathbb{R}^{Nm}$ such that $\nabla g(y_r^*) + Lv^* + Ly_r^* = 0$ holds, and (y^*, v^*) is the saddle of $G(y, v) = g(y) + y^T Lv + \frac{1}{2} y^T Ly$. Consider the Lyapunov function of the cyber dynamics

$$V_c = \frac{1}{2}(\|y_r - y^*\|^2 + \|v - v^*\|^2)$$

Note that $z_1 = y - y_r$ under the healthy conditions. Then (3) becomes

$$\begin{aligned}\dot{y}_r =& -\nabla g(y_r) - Lv - (1+\eta)Ly_r - (1+\eta)Lz_1\\ \dot{v} =& Ly_r + Lz_1\end{aligned} \tag{C20}$$

The time derivative of V_c along with (C20) is

$$\begin{aligned}\dot{V}_c =&(y_r-y^*)^T[-\nabla g(y_r)-Lv-(1+\eta)Ly_r]+(v-v^*)^TLy_r-(1+\eta)z_1^TLy_r+(v-v^*)^TLz_1\\ \overset{(a)}{\leqslant}&G(y^*,v)-G(y^*,v^*)+G(y^*,v^*)-G(y,v^*)-\eta y_r^TLy_r-(1+\eta)z_1^TLy_r+(v-v^*)^TLz_1\\ \overset{(b)}{\leqslant}&-\eta y_r^TLy_r-(1+\eta)z_1^TLy_r+v^TLz_1-z_1^T\pi\end{aligned}\tag{C21}$$

where the equalities: (a) follows from the convexity of G in the first argument and the linearity of G in its second argument; (b) follows from the fact that (y^*,v^*) is the saddle point of G. Note that the mismatching terms v^TLz_1 and $-z_1^T\pi$ will be compensated by the following physical dynamics.

Physical system: The convergence analysis is discussed based on backstepping procedure. Rewrite (1) into a compact form

$$\mathcal{P}:\ \begin{cases}\dot{x}_i=x_{i+1}+\varphi_i(\bar{x}_i)\theta,\ i=1,\cdots,n-1\\ \dot{x}_n=Bu+\varphi_n(x)\theta,\\ y=x_1\end{cases}\tag{C22}$$

where $\varphi_i(\bar{x}_i)=\mathrm{diag}\{\varphi_i^{(1)}(\bar{x}_i^{(1)}),\cdots,\varphi_i^{(N)}(\bar{x}_i^{(N)})\}$ and $\theta=\mathrm{vec}(\theta_1,\cdots,\theta_N)$.

The error dynamics can be expressed as

$$\begin{cases}\dot{z}_1=\alpha_1+\varphi_1(\bar{x}_1)\theta-\dot{y}_r+z_2\\ \dot{z}_i=\alpha_i+\varphi_i(\bar{x}_i)\theta-\dot{\alpha}_{i-1}+z_{i+1}\\ \dot{z}_n=Bu+\varphi_n(x)\theta-\dot{\alpha}_{n-1}\end{cases}\tag{C23}$$

which can be spitted into inner-loop and outer-loop subsystems:

$$\begin{cases}\dot{z}_{1,I}=\alpha_{1,I}+\psi_1(\bar{x}_1)\lambda+z_2\\ \dot{z}_{i,I}=\alpha_{i,I}+\psi_i(\bar{x}_i)\lambda-\dot{\alpha}_{i-1,I}+z_{i+1}-\mu z_i\\ \dot{z}_{n,I}=Bu_I+\psi_n(x)\lambda-\dot{\alpha}_{n-1,I}-\mu z_n\end{cases}\tag{C24}$$

and

$$\begin{cases}\dot{z}_{1,O}=\alpha_{1,O}-\dot{y}_r\\ \dot{z}_{i,O}=\alpha_{i,O}-\dot{\alpha}_{i-1,O}\\ \dot{z}_{n,O}=Bu_O-\dot{\alpha}_{n-1,O}\end{cases}\tag{C25}$$

where $z_i=z_{i,I}+z_{i,O}$ for $i=1,\cdots,n$.

Next, we provide the Lyapunov analysis of the physical dynamics by considering

$$V_p=\frac{1}{2}\left(\sum_{i=1}^n\|z_i\|^2+\tilde{\lambda}^T\Gamma^{-1}\tilde{\lambda}+\tilde{\rho}^T\Gamma_0^{-1}\tilde{\rho}+\tilde{\pi}^T\Gamma_1^{-1}\tilde{\pi}\right)$$

where $\tilde{\lambda}=\lambda-\hat{\lambda}$, $\tilde{\rho}=\rho-\hat{\rho}$, $\tilde{\pi}=\pi-\hat{\pi}$.

The derivative of V_p can be computed as

$$\begin{aligned}\dot{V}_p=&\sum_{i=1}^n z_i\dot{z}_i-\tilde{\lambda}^T\Gamma^{-1}\dot{\hat{\lambda}}-\tilde{\rho}^T\Gamma_0^{-1}\dot{\hat{\rho}}-\tilde{\pi}^T\Gamma_1^{-1}\dot{\hat{\pi}}\\ =&\dot{V}_I+\dot{V}_O\end{aligned}$$

where $\dot{V}_I=\sum_{i=1}^n z_i\dot{z}_{i,I}-\tilde{\lambda}^T\Gamma^{-1}\dot{\hat{\lambda}}-\tilde{\rho}^T\Gamma_0^{-1}\dot{\hat{\rho}}-\tilde{\pi}^T\Gamma_1^{-1}\dot{\hat{\pi}}$ and $\dot{V}_O=\sum_{i=1}^n z_i\dot{z}_{i,O}$ represent the inner-loop and outer-loop Lyapunov derivatives, respectively.

Consider the inner-loop error dynamics (C24) with controls (C11)–(C13) and adaptive laws (C17)–(C19). Following the traditional backstepping procedure [13], along with (C24), we can obtain

$$\dot{V}_I\leqslant-\sum_{i=1}^n z_i^TC_iz_i-\mu\sum_{i=2}^n\|z_i\|^2-\rho\|z_1\|^2+z_1^T\pi-z_1^TS\left(\frac{z_1}{\delta}\right).\tag{C26}$$

Now we consider the outer-loop error dynamics (C25) with controls (C14)-(C16).

Step 1. In view of (C10) and (C25), we have

$$\dot{z}_{1,O}=\alpha_{1,O}+\nabla g(y_r)+Lv+(1+\eta)Ly\tag{C27}$$

To stabilize (C27), consider the Lyapunov derivative $\dot{V}_{1,O}=z_1\dot{z}_{1,O}$. Then using the virtual control (C14), we have

$$\begin{aligned}\dot{V}_{1,O}=&z_1^T[\alpha_{1,O}+\nabla g(y_r)+Lv+(1+\eta)Ly]\\ =&-z_1^TLv+(1+\eta)z_1^TLy\\ =&-z_1^TLv+(1+\eta)z_1^TL(y_r+z_1)\\ \leqslant&-z_1^TLv+(1+\eta)z_1^TLy_r+(1+\eta)\|L\|\|z_1\|^2\end{aligned}\tag{C28}$$

Step $i(2 \leqslant i \leqslant n)$. Note that the arguments of the function $\alpha_{i-1,O}$ involve y_r and v. From (C10) and (C25), we have

$$\begin{aligned}\dot{z}_{i,O}=&\alpha_{i,O}+\frac{\partial\alpha_{i-1,O}}{\partial y_r}\left[\nabla g(y_r)+Lv+(1+\eta)Ly\right]-\frac{\partial\alpha_{i-2,O}}{\partial y_r}L^2y\\ =&\left[(1+\eta)\frac{\partial\alpha_{i-1,O}}{\partial y_r}L-\frac{\partial\alpha_{i-2,O}}{\partial y_r}L^2\right](y_r+z_1)\end{aligned}\tag{C29}$$

On the compact set $\{V(t) \leqslant V(0)\}$, there exists a positive constant such that $\|\partial\alpha_{i,O}/\partial y_r\| \leqslant \Pi$ for all $i = 1, \cdots, n-1$. Based on the fact and using the triangular inequality, the Lyapunov derivative $\dot{V}_{i,O} = z_i\dot{z}_{i,O}$ along with (C29) can be expressed as

$$\begin{aligned}\dot{V}_{i,O}=&z_i^T\left[(1+\eta)\frac{\partial\alpha_{i-1,O}}{\partial y_r}L-\frac{\partial\alpha_{i-2,O}}{\partial y_r}L^2\right](y_r+z_1)\\ \leqslant&\mu\|z_i\|^2+2\|L\|\|z_1\|^2+2y_r^TLy_r\end{aligned}\tag{C30}$$

Combining (C28) and (C30), the outer-loop Lyapunov derivative satisfies

$$\dot{V}_O \leqslant -z_1^TLv+(1+\eta)z_1^TLy_r+\rho\|z_1\|^2+2(n-1)y_r^TLy_r+\mu\sum_{i=2}^{n}\|z_i\|^2\tag{C31}$$

Finally, construct the Lyapunov function $V = V_c + V_p$ for the overall CPS. Taking (C21), (C26) and (C31) into account, its time derivative satisfies

$$\begin{aligned}\dot{V}\leqslant&-(\eta-2(n-1))y_r^TLy_r-\sum_{i=1}^{n}z_i^TC_iz_i-\sum_{j=1}^{N}z_1^{(j)T}S\left(\frac{z_1^{(j)}}{\delta^{(j)}}\right)\\ \leqslant&-(\eta-2(n-1))y_r^TLy_r-\sum_{i=1}^{n}z_i^TC_iz_i\end{aligned}\tag{C32}$$

where the fact $z_1^{(j)T}S(z_1^{(j)}/\delta^{(j)}) \geqslant 0$ is used.

Choose $\eta > 2(n-1)$. Then $\dot{V} \leqslant 0$. Thus, $\{V(t) \leqslant V(0)\}$ is an invariant set. It implies that $z(t)$, $y_r(t)$, $v(t)$, $\hat{\lambda}(t)$, $\hat{\rho}(t)$, $\hat{\pi}(t)$ and $z_1^{(j)T}S(z_1^{(j)}/\delta^{(j)})$ are bounded. Then $y(t) = z_1(t)+y_r(t)$ is bounded. Along with the backstepping procedure, $\alpha_i(t)$, $u_i(t)$ and $x_i(t)$ are also bounded. Noting $\dot{y}_r(t), \dot{v}(t) \in L_\infty$ and $y_r^T(t)Ly_r(t), z_i(t) \in L_2$. According to Barbalat's Lemma, $\lim_{t\to\infty} y_r^T(t)Ly_r(t) = 0$ and $\lim_{t\to\infty} z_i(t) = 0$. Finally, following the proof of [12, Theorem 4.1], one obtains that $\lim_{t\to\infty} y_r(t) = y^*$. Thus, we can conclude that $\lim_{t\to\infty}[y(t) - y^*] = \lim_{t\to\infty}[y(t) - y_r(t) + y_r(t) - y^*] = 0$.

Appendix D ADI design procedure

Taking (3) and (5) into account, we can express the error dynamics as the following form

$$\dot{e}_r^{(j)} = -\eta^{(j)}(e_r^{(j)} + z_1^{(j)}) - \eta^{(j)}a^{(j)}\tag{D1}$$

$$\dot{e}_v^{(j)} = -w_{\mathbf{N}_j}(e_v^{(j)} - e_r^{(j)} - z_1^{(j)}) + w_{\mathbf{N}_j}a^{(j)}\tag{D2}$$

where $w_{\mathbf{N}_j} = \sum_{i\in\mathbf{N}_j} w_{ji}$ and $\eta^{(j)} = (1+\eta)w_{\mathbf{N}_j}$.

In the error dynamics (D1)–(D2), the tracking error $z_1^{(j)}$ is simultaneously affected by the local attack and multiple attacks propagated from its neighbors, which hinders the identification of local attacks. Later, we will propose a robust method to address this problem.

The jth detection thresholds, denoted by $\bar{e}_{r,H}^{(j)}(t)$ and $\bar{e}_{v,H}^{(j)}(t)$, are designed based on the bounds of residuals $e_r^{(j)}(t)$ and $e_v^{(j)}(t)$ under healthy conditions, respectively. The error dynamics under healthy conditions, denoted by $(e_{r,H}^{(j)}, e_{v,H}^{(j)})$, can be expressed by

$$\dot{e}_{r,H}^{(j)} = -\eta^{(j)}(e_{r,H}^{(j)} + z_1^{(j)})\tag{D3}$$

$$\dot{e}_{v,H}^{(j)} = -w_{\mathbf{N}_j}(e_{v,H}^{(j)} - e_{r,H}^{(j)} - z_1^{(j)})\tag{D4}$$

Along with (D3) and (D4), the residuals under healthy conditions $e_{r,H}^{(j)}(t)$ and $e_{v,H}^{(j)}(t)$ can be bounded by

$$\|e_{r,H}^{(j)}(t)\| \leqslant e^{-\eta^{(j)}t}e_{r,H}^{(j)}(0) + \Psi(\eta^{(j)}, z_1^{(j)}(t), 0, t)\tag{D5}$$

$$\|e_{v,H}^{(j)}(t)\| \leqslant e^{-w_{\mathbf{N}_j}t}e_{v,H}^{(j)}(0) + \Psi(w_{\mathbf{N}_j}, e_{r,H}^{(j)}(t) + z_1^{(j)}(t), 0, t)\tag{D6}$$

where $\Psi(\alpha, h(t), t_0, t) := \alpha\int_{\tau=t_0}^{t} e^{\alpha(\tau-t)}\|h(\tau)\|d\tau$.

Note that $z_1^{(j)}(t) \in \Delta_z^{\delta^{(j)}}$. Substituting it into (D5) and (D6), we can obtain the two "strongly-robust" thresholds

$$\bar{e}_{r,H}^{(j)}(t) = e^{-\eta^{(j)}t}e_{r,H}^{(j)}(0) + \bar{\Psi}^{(j)}_{\Delta_z^{\delta^{(j)}}}(\eta^{(j)}, 0, t)$$

$$\bar{e}_{v,H}^{(j)}(t) = e^{-w_{\mathbf{N}_j}t}e_{v,H}^{(j)}(0) + \bar{\Psi}^{(j)}_{\Delta_{ez}^{\delta^{(j)}}}(w_{\mathbf{N}_j}, 0, t)$$

where $\Delta_{ez}^{\delta^{(j)}} := \{e + z : \|e\| \leqslant \bar{e}_{r,H}^{(j)},\ z \in \Delta_z^{\delta^{(j)}}\}$ and $\bar{\Psi}^{(j)}_{\Delta^{\delta^{(j)}}}(\alpha, t_0, t) := \sup_{h(t)\in\Delta^{\delta^{(j)}}} \alpha\int_{\tau=t_0}^{t} e^{\alpha(\tau-t)}\|h(\tau)\|d\tau$.

Thus, the decision logic implemented in each module $\mathcal{M}^{(j)}$, denoted by $\mho^{(j)}(t)$, can be defined as

$$\mho^{(j)}(t) = \mho^{(j,r)}(t) \cup \mho^{(j,v)}(t)\tag{D7}$$

where $\mho^{(j,r)}(t)$: $\|e_r^{(j)}(t)\| \leqslant \bar{e}_{r,H}^{(j)}(t)$ and $\mho^{(j,v)}(t)$: $\|e_v^{(j)}(t)\| \leqslant \bar{e}_{v,H}^{(j)}(t)$. If $\mho^{(j)}(t)$ is violated, then $\mathcal{M}^{(j)}$ will generate an alarm.

Appendix E Proof of Theorem 1 and Remark

Proof. For sake of contradiction, we suppose that no local attack $a^{(j)}$ has occurred, then the error dynamics are expressed as (10) and (11). According to the PPT, one has $\|z_1^{(j)}(t)\| < \sqrt{m}\delta^{(j)}(t)$ for all $t \in [0, T_d^{(j)}]$. Further, if $\int_{t=0}^{T_d^{(j)}} \delta^{(j)2}(t)dt \leqslant \Omega/m$, then $\int_{t=0}^{T_d^{(j)}} \left\|z_1^{(j)}(t)\right\|^2 dt < m\int_{t=0}^{T_d^{(j)}} \delta^{(j)2}(t)dt \leqslant \Omega$. It means that there exists an instant $t_*^{(j)} > T_d^{(j)}$ such that $z_1^{(j)}(t) \in \Delta_z^{\delta^{(j)}}$ can still be ensured for any $t \in [0, t_*^{(j)}]$ even in the presence of the impacts of the attacks $a^{(i)}$ propagated from its neighbors $i \in \mathbf{N}_j$. Hence, $\mho^{(j)}(T_d^{(j)})$ is satisfied, a contradiction, which in turn implies that the occurrence of local attack $a^{(j)}$ is guaranteed. ■

Remark E1. From (D1) and (D2), the variable $z_1^{(j)}$ suffers the coupling effects of attack signals $a^{(i)}$ from the neighbors $i \in \mathbf{N}_j$ such that the local attack $a^{(j)}$ cannot be isolated from $a^{(i)}, i \in \mathbf{N}_j$ (i.e., the attack occurring at the neighbor may also cause the false alarm of the local monitoring module). To address it, in the control module $\mathcal{K}^{(j)}$ the prescribed performance technique (PPT) [14, 15] is used to restrict the bound of tracking error $z_1^{(j)}$ (the nonlinear function S is introduced into $\mathcal{K}^{(j)}$). As a result, the detection thresholds, or further the proposed ADI method, are strongly robust against propagated attacks. In other words, the prescribed performance bound $\|z_1^{(j)}(t)\| < \sqrt{m}\delta^{(j)}(t)$ restrains the propagating effects of the attacks $a^{(i)}, i \in \mathbf{N}_j$ on $e_r^{(j)}$ and $e_v^{(j)}$ such that the sensitivity to the local attack is improved. In fact, according to the special form of

$$S\left(\frac{z_1^{(j)}}{\delta^{(j)}}\right) = \frac{1}{2}\ln\left(1 + \frac{z_1^{(j)}}{\delta^{(j)}}\right) - \frac{1}{2}\ln\left(1 - \frac{z_1^{(j)}}{\delta^{(j)}}\right)$$

it makes the PPT to have strong robustness, i.e., the prescribed performance constraint $\|z_1^{(j)}(t)\| < \sqrt{m}\delta^{(j)}(t)$ can be guaranteed until the direct attack-associated terms $-\eta^{(j)}a^{(j)}$ and $w_{\mathbf{N}_j}a^{(j)}$ in (D1)–(D2) cause the alarm (if these two terms exist). Hence, the attack on a subsystem does not cause false alarm of the monitoring module of its neighbors and the local attack can be identified.

Appendix F Proof of Theorem 2 and Remarks

Following the stability analysis given in Appendix B, the proof of Theorem 2 can be formalized by proving

$$\begin{cases} \|e_r^{(j)}(t)\| \leqslant \bar{e}_{r,H}^{(j)}(t) \\ \|e_v^{(j)}(t)\| \leqslant \bar{e}_{v,H}^{(j)}(t) \end{cases} \Rightarrow a^{(j)}(t) \in L_2[0, +\infty)$$

A brief sketch of proof procedure is shown in Figure F1.

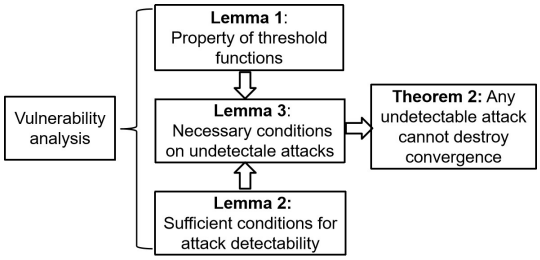

Figure F1 A brief sketch of vulnerability analysis

We first give some properties of functions $\bar{\Psi}^{(j)}_{\Delta_z^{\delta^{(j)}}}(\alpha, t_0, t)$ and $\bar{\Psi}^{(j)}_{\Delta_{ez}^{\delta^{(j)}}}(\alpha, t_0, t)$ in the "strongly-robust" thresholds, which are important for analyzing the detectability performance of the ADI mechanism.

Lemma F1. Let $\delta^{(j)}(t) = (k_0^{(j)}e^{-c^{(j)}t} + k_b^{(j)})/\sqrt{m}$, where $k_0^{(j)}$, $k_b^{(j)}$ and $c^{(j)}(\neq \alpha)$ are positive design parameters such that $|z_{1,s}^{(j)}(0)| < \delta^{(j)}(0)$, $s = 1, \cdots, m$. Then

(a) $\bar{\Psi}^{(j)}_{\Delta_z^{\delta^{(j)}}}(\alpha, 0, t) \leqslant k_b^{(j)}(1 - e^{-\alpha t}) + \frac{\alpha k_0^{(j)}}{\alpha - c^{(j)}}(e^{-c^{(j)}t} - e^{-\alpha t})$;

(b) $\bar{\Psi}^{(j)}_{\Delta_{ez}^{\delta^{(j)}}}(\alpha, 0, t) \leqslant 2k_b^{(j)}(1 - e^{-\alpha t}) + \frac{(2\alpha - c^{(j)})\alpha k_0^{(j)}}{(\alpha - c^{(j)})^2}(e^{-c^{(j)}t} - e^{-\alpha t}) + \alpha\left[k_b^{(j)} + \frac{\alpha k_0^{(j)}}{\alpha - c^{(j)}} + e_{r,H}^{(j)}(0)\right]te^{-\alpha t}$;

(c) $\int_{t=0}^{\infty} \bar{\Psi}^{(j)2}_{\Delta_z^{\delta^{(j)}}}(\alpha, 0, t)dt \leqslant \Omega$, $\int_{t=0}^{\infty} \bar{\Psi}^{(j)2}_{\Delta_{ez}^{\delta^{(j)}}}(\alpha, 0, t)dt \leqslant 2\Omega$.

Proof. (a) Note that $\Psi^{(j)}(\alpha, h(t), 0, t)$ increases as $\|h(t)\|$ increases. Based on the constraint $\|z_1^{(j)}(t)\| \leqslant k_0^{(j)}e^{-c^{(j)}t} + k_b^{(j)}$, we have

$$\begin{aligned} \bar{\Psi}^{(j)}_{\Delta_z^{\delta^{(j)}}}(\alpha, 0, t) \leqslant & \alpha\int_{\tau=0}^{t} e^{\alpha(\tau - t)}(k_0^{(j)}e^{-c^{(j)}t} + k_b^{(j)})d\tau \\ \leqslant & k_b^{(j)}(1 - e^{-\alpha t}) + \frac{\alpha k_0^{(j)}}{\alpha - c^{(j)}}(e^{-c^{(j)}t} - e^{-\alpha t}) \end{aligned}$$

(b) Based on (a) and using similar analysis, the proof can be completed.

(c) Let $h^*(t) := \arg\sup_{h(t) \in \Delta_z^{\delta^{(j)}}} \alpha\int_{\tau=0}^{t} e^{\alpha(\tau-t)}\|h(\tau)\|d\tau$, i.e., $\bar{\Psi}^{(j)}_{\Delta_z^{\delta^{(j)}}}(\alpha, t_0, t) = \alpha\int_{\tau=0}^{t} e^{\alpha(\tau-t)}\|h^*(\tau)\|d\tau$. Since $\Delta_z^{\delta^{(j)}}$ is a compact set, $h^*(t)$ satisfies $\int_{\tau=0}^{\infty}\|h^*(\tau)\|^2 d\tau \leqslant \Omega$. To show (c), we construct the auxiliary dynamics

$$\dot{\chi}(t) = -\alpha\chi(t) + \alpha\|h^*(t)\|, \ \chi(0) = 0 \tag{F1}$$

By integrating (F1) we can find $\chi(t) = \bar{\Psi}^{(j)}_{\Delta^{\delta^{(j)}}_{z}}(\alpha, t_0, t)$. On the other hand, considering the Lyapunov function $V = \chi^2/2$, its derivative along with (F1) satisfies

$$\dot{V} = \chi(-\alpha\chi + \alpha\|h^*\|) \leqslant -\alpha V + \frac{\alpha}{2}\|h^*\|^2$$

integrating two sides of which yields $\int_{t=0}^{\infty}\chi^2(t)dt \leqslant \Omega$. Using similar procedure to $\bar{\Psi}^{(j)}_{\Delta^{\delta^{(j)}}_{ez}}(\alpha, 0, t)$, it is easily obtained that $\int_{t=0}^{\infty}\bar{\Psi}^{(j)2}_{\Delta^{\delta^{(j)}}_{ez}}(\alpha, 0, t)dt \leqslant 2\Omega$. ■

Remark F1. From Lemma F1-(c), one has $\lim_{t\to\infty}\bar{\Psi}^{(j)}_{\Delta^{\delta^{(j)}}_{z}}(\alpha, 0, t) = 0$ and $\lim_{t\to\infty}\bar{\Psi}^{(j)}_{\Delta^{\delta^{(j)}}_{ez}}(\alpha, 0, t) = 0$ following Barbalat's Lemma. It means that only if $\mho^{(j)}(t)$ is satisfied, the bound functions $\bar{e}^{(j)}_{r,H}(t)$ and $\bar{e}^{(j)}_{v,H}(t)$ will converge to zero, which in turn implies that $e^{(j)}_r(t)$ and $e^{(j)}_v(t)$ converge to zero. Lemma F1-(a) and -(b) give prescribed performance bounds of $\bar{\Psi}^{(j)}_{\Delta^{\delta^{(j)}}_{z}}$ and $\bar{\Psi}^{(j)}_{\Delta^{\delta^{(j)}}_{ez}}$. By replacing $\bar{\Psi}^{(j)}_{\Delta^{\delta^{(j)}}_{z}}$ and $\bar{\Psi}^{(j)}_{\Delta^{\delta^{(j)}}_{ez}}$ with the prescribed performance bounds, we can obtain low-complexity thresholds. However, such relaxations will weaken the detectability and extend the detection time. Also, the two prescribed performance bounds converge to $k^{(j)}_b$ and $2k^{(j)}_b$ instead of zero, respectively, which may results in that the detection mechanism is not able to detect some small but persistent attacks. However, we can choose $k^{(j)}_b$ to be sufficiently small to reduce the effects of such attacks.

To examine the sensitivity of the proposed attack detection scheme, the following attack detectability is analyzed.

Lemma F2 (Detectable attacks). Consider cyber attack $a^{(j)}$ occurring at the subsystem $(\mathcal{P}^{(j)}, \mathcal{C}^{(j)})$. If there exists some time instant $T^{(j)}_d > T^{(j)}_a$ such that the attack satisfies

$$\begin{aligned}\eta^{(j)}\left\|\int_{t=T^{(j)}_a}^{T^{(j)}_d} e^{\eta^{(j)}(t-T^{(j)}_d)}a^{(j)}(t)dt\right\| >& 2e^{\eta^{(j)}(T^{(j)}_a-T^{(j)}_d)}\|e^{(j)}_r(T^{(j)}_a)\| + \bar{\Psi}^{(j)}_{\Delta^{\delta^{(j)}}_{z}}(\eta^{(j)}, T^{(j)}_a, T^{(j)}_d)\\ &+ \eta^{(j)}\int_{t=T^{(j)}_a}^{T^{(j)}_d} e^{\eta^{(j)}(t-T^{(j)}_d)}\left\|z^{(j)}_1(t)\right\|dt\end{aligned} \tag{F2}$$

or

$$\begin{aligned}w_{\mathbf{N}_j}\left\|\int_{t=T^{(j)}_a}^{T^{(j)}_d} e^{w_{\mathbf{N}_j}(t-T^{(j)}_d)}a^{(j)}(t)dt\right\| >& 2e^{w_{\mathbf{N}_j}(T^{(j)}_a-T^{(j)}_d)}\|e^{(j)}_v(T^{(j)}_a)\| + \bar{\Psi}^{(j)}_{\Delta^{\delta^{(j)}}_{ez}}(w_{\mathbf{N}_j}, T^{(j)}_a, T^{(j)}_d)\\ &+ w_{\mathbf{N}_j}\int_{t=T^{(j)}_a}^{T^{(j)}_d} e^{w_{\mathbf{N}_j}(t-T^{(j)}_d)}\left\|e^{(j)}_r(t) + z^{(j)}_1(t)\right\|dt\end{aligned} \tag{F3}$$

then the attack $a^{(j)}(t)$ is detected by $\mho^{(j)}$ at $t = T^{(j)}_d$.

Proof. After the first occurrence of the attack $a^{(j)}$, i.e., $t > T^{(j)}_a$, the time derivative of $e^{(j)}_r(t)$ becomes $\dot{e}^{(j)}_r = -\eta^{(j)}(e^{(j)}_r + z^{(j)}_1) + \eta^{(j)}a^{(j)}$. Integrating both sides and applying the triangular inequality yield

$$\begin{aligned}\|e^{(j)}_r(T^{(j)}_d)\| \geqslant& \eta^{(j)}\left\|\int_{t=T_f}^{T^{(j)}_d} e^{\eta^{(j)}(t-T^{(j)}_d)}a^{(j)}(t)dt\right\| - e^{\eta^{(j)}(T^{(j)}_a-T^{(j)}_d)}\|e^{(j)}_r(T^{(j)}_a)\|\\ &- \eta^{(j)}\int_{t=T^{(j)}_a}^{T^{(j)}_d} e^{\eta^{(j)}(t-T^{(j)}_d)}\left\|z^{(j)}_1(t)\right\|dt,\end{aligned}$$

substituting (F2) into which yields that $\|e^{(j)}_r(T^{(j)}_d)\| > e^{\eta^{(j)}(T^{(j)}_a-T^{(j)}_d)}e^{(j)}_r(T^{(j)}_a) + \bar{\Psi}^{(j)}_{\Delta^{\delta^{(j)}}_{z}}(\eta^{(j)}, T^{(j)}_a, t)$.

Following similar analysis, (F3) ensures

$$\|e^{(j)}_v(T^{(j)}_d)\| > e^{w_{\mathbf{N}_j}(T^{(j)}_a-T^{(j)}_d)}e^{(j)}_v(T^{(j)}_a) + \bar{\Psi}^{(j)}_{\Delta^{\delta^{(j)}}_{ez}}(w_{\mathbf{N}_j}, T^{(j)}_a, t)$$

From the definition of $\mho^{(j)}(t)$, the attack $a^{(j)}(t)$ satisfying (F2) or (F3) provokes the violation of decision logic $\mho^{(j)}(t)$ and resultantly $a^{(j)}(t)$ is detected when $t = T^{(j)}_d$. ■

Remark F2. The inequalities (F2)–(F3) characterize a class of detectable attacks under the worst-case detectability. The computation of detection time $T^{(j)}_d$ may be somewhat conservative. However, differing from the fault, the attacker may strategically design the (worst-case) attack to extend the detection time as much as possible. Thus, the real-time detection time may sufficiently approach to $T^{(j)}_d$ but not exceed than it. In general, from (F2)–(F3), if the cyber attack on the time interval $[T^{(j)}_a, T^{(j)}_d]$ is sufficiently large, then the attack can be detected. However, a malicious attacker may strategically inject the attack signals which are not detected by the proposed distributed ADI scheme, yet degrade the system performance. The following lemma gives necessary conditions on undetectable attacks.

Lemma F3 (Undetectable attacks). If a cyber attack $a^{(j)}(t)$ occurring at subsystem $(\mathcal{P}^{(j)}, \mathcal{C}^{(j)})$ is undetectable by $\mho^{(j)}$, the following inequality holds

$$\int_{t=T^{(j)}_a}^{\infty}\left(\int_{\tau=T^{(j)}_a}^{t} e^{\eta^{(j)}(\tau-t)}\|a^{(j)}(\tau)\|d\tau\right)^2 dt \leqslant M_1 \tag{F4}$$

$$\int_{t=T^{(j)}_a}^{\infty}\left(\int_{\tau=T^{(j)}_a}^{t} e^{w_{\mathbf{N}_j}(\tau-t)}\|a^{(j)}(\tau)\|d\tau\right)^2 dt \leqslant M_2 \tag{F5}$$

where $M_1 = \frac{8\Omega}{\eta^{(j)2}} + \frac{4\|e_r^{(j)}(T_a^{(j)})\|^2}{\eta^{(j)4}}$ and $M_2 = \frac{16\Omega}{w_{\mathbf{N}_j}^2} + \frac{4\|e_v^{(j)}(T_a^{(j)})\|^2}{w_{\mathbf{N}_j}^4}$. Further, $\int_{t=T_a^{(j)}}^{\infty} \|a^{(j)}(t)\|^2 dt < +\infty$.

Proof. Here we directly show (F5) from (F3), and (F4) is easily obtained by applying similar procedure to (F2). If the attack $a^{(j)}(t)$ occurring at time $T_a^{(j)}$ is not detectable, from (F3) in Lemma F2, then for any $t \geqslant T_a^{(j)}$,

$$w_{\mathbf{N}_j} \left\| \int_{\tau=T_a^{(j)}}^{t} e^{w_{\mathbf{N}_j}(\tau-t)} a^{(j)}(\tau) d\tau \right\| \leqslant 2e^{w_{\mathbf{N}_j}(T_a^{(j)}-t)} \|e_v^{(j)}(T_a^{(j)})\| + 2\bar{\Psi}^{(j)}_{\Delta_{ez}^{\delta(j)}}(w_{\mathbf{N}_j}, T_a^{(j)}, t) \tag{F6}$$

Consider the right-hand side of (F6). Taking square and integral consecutively to each term yields

$$4\|e_v^{(j)}(T_a^{(j)})\|^2 \int_{t=T_a^{(j)}}^{\infty} e^{2w_{\mathbf{N}_j}(T_a^{(j)}-t)} dt \leqslant \frac{2\|e_v^{(j)}(T_a^{(j)})\|^2}{w_{\mathbf{N}_j}},$$
$$4\int_{t=T_a^{(j)}}^{\infty} \bar{\Psi}^{(j)2}_{\Delta_{ez}^{\delta(j)}}(w_{\mathbf{N}_j}, T_a^{(j)}, t) dt \leqslant 8\Omega.$$

where the second inequality follows from Lemma F1-(c).

Then using the Cauchy-Buniakowsky-Schwarz inequality, one has

$$4\int_{t=T_a^{(j)}}^{\infty} \left(e^{w_{\mathbf{N}_j}(T_a^{(j)}-t)} \|e_v^{(j)}(T_a^{(j)})\| + \bar{\Psi}^{(j)}_{\Delta_{ez}^{\delta(j)}}(w_{\mathbf{N}_j}, T_a^{(j)}, t) \right)^2 dt \leqslant \frac{4\|e_v^{(j)}(T_a^{(j)})\|^2}{w_{\mathbf{N}_j}^2} + 16\Omega \tag{F7}$$

Combining (F6) and (F7), Eq. (F5) follows at once.

Next, to prove $\int_{t=T_a^{(j)}}^{\infty} \|a^{(j)}(t)\|^2 dt < +\infty$, we consider the error dynamics

$$\dot{e}_v^{(j)} = -w_{\mathbf{N}_j}(e_v^{(j)} - e_r^{(j)} - z_1^{(j)}) + w_{\mathbf{N}_j} a^{(j)}.$$

Noting that $\mho^{(j)}$ is always satisfied, then $e_r^{(j)}, e_v^{(j)}, z_1^{(j)} \in L_2[0, +\infty)$ from Lemma F1-(c). Therefore, there exist a sufficiently big $T \geqslant T_a^{(j)}$ and a time interval $\Xi_{v,s}$ with $\nu(\Xi_{v,s}) = 0$ such that

$$\frac{a_s^{(j)}(t)}{e_{v,s}^{(j)}(t)} < 1, \ \forall t \in [T, \infty) \backslash \Xi_{v,s}$$

which means that there exists a function $\bar{\phi}_{v,s}(t) < -\|e_{v,s}^{(j)}(t)\|$ such that

$$|a_s^{(j)}(t)| < |e_{v,s}^{(j)}(t)| \text{ or } a^{(j)}(t) = \bar{\phi}_{v,s}(t)\text{sgn}(e_{v,s}^{(j)}(t)) \tag{F8}$$

for any $t \in [T, \infty) \backslash \Xi_{v,s}$, where $a_s^{(j)}$ and $e_{v,s}^{(j)}$ represent the sth element of vectors $a^{(j)}$ and $e_v^{(j)}$, respectively. Applying similar procedure to $\dot{e}_r^{(j)} = -\eta^{(j)}(e_r^{(j)} + z_1^{(j)}) + \eta^{(j)} a^{(j)}$, there exist a function $\bar{\phi}_{r,s}(t) \leqslant -\|e_{r,s}^{(j)}(t)\|$ and a time interval $\Xi_{r,s}$ with $\nu(\Xi_{r,s}) = 0$ such that

$$|a_s^{(j)}(t)| < |e_{r,s}^{(j)}(t)| \text{ or } a_s^{(j)}(t) = \bar{\phi}_{r,s}(t)\text{sgn}(e_{r,s}^{(j)}(t)) \tag{F9}$$

for any $t \in [T, +\infty) \backslash \Xi_{r,s}$.

Compared (F8) with (F9), and noting that the equality

$$\text{sgn}(e_{v,s}^{(j)}(t)) = \text{sgn}(e_{r,s}^{(j)}(t)), \ \forall t \in [T, +\infty) \backslash (\Xi_{r,s} \cup \Xi_{v,s})$$

does not hold, it yields that $\|a^{(j)}(t)\| < \max\{\|e_r^{(j)}(t)\|, \|e_v^{(j)}(t)\|\}$ for any $t \in [T, +\infty) \backslash \bigcup_{s=1}^{m} (\Xi_{r,s} \cup \Xi_{v,s})$, which guarantees $\int_{t=T_a^{(j)}}^{\infty} \|a^{(j)}(t)\|^2 dt < +\infty$. ■

Remark F3. From its proof, we can see the design of double coupling residuals plays a key role in removing the existence of the attacks $a_s^{(j)}(t) = \bar{\phi}_{r,s}(t)\text{sgn}(e_{r,s}^{(j)}(t))$ and $a_s^{(j)}(t) = \bar{\phi}_{v,s}(t)\text{sgn}(e_{v,s}^{(j)}(t))$ which are stealthy against two single decision logics $\mho^{(j,r)}$ and $\mho^{(j,v)}$, respectively. Lemma F3 implies that any undetectable attack must belong to $L_2[0, +\infty)$.

Appendix G Simulation illustration

As a practical application of the studied problem framework, we apply our algorithms to the problem of motion coordination of multiple Remotely Operated Vehicles (ROVs): rendezvous at a location which is optimal for the formation. The dynamics equation of each ROV can be expressed in two coordinate frames [16]:

$$\begin{aligned} \dot{\boldsymbol{\eta}} &= J(\boldsymbol{\eta})\boldsymbol{\nu} \\ M\dot{\boldsymbol{\nu}} + C(\boldsymbol{\nu}) + D(\boldsymbol{\nu})\boldsymbol{\nu} + g(\boldsymbol{\eta}) &= \boldsymbol{\tau} + \Delta \boldsymbol{f} \end{aligned} \tag{G1}$$

where $\boldsymbol{\eta} = [x, y, z, \phi, \theta, \psi]^T$ is the position and orientation described in the earth-fixed frame ($|\theta| < \pi/2$ and $|\phi| < \pi/2$), $\boldsymbol{\nu} = [u, v, w, p, q, r]^T$ is the linear and angular velocity in the body-fixed frame, $M = M_{RB} + M_A$ and M is positive definite, $C(\boldsymbol{\nu}) = C_{RB}(\boldsymbol{\nu}) + C_A(\boldsymbol{\nu})$ satisfying $C(\boldsymbol{\nu}) = -C^T(\boldsymbol{\nu})$, M_{RB} is the rigid-body inertia matrix, M_A is the added inertia matrix; $C_{RB}(\boldsymbol{\nu})$ is the rigid-body Coriolis and centripetal matrix, $C_A(\boldsymbol{\nu})$ is the hydrodynamic Coriolis and centripetal matrix in cluding added mass, $D(\boldsymbol{\nu})$ is hydrodynamic damping and lift matrix, $g(\boldsymbol{\eta})$ is a vector of gravitational forces and moment, $\boldsymbol{\tau}$ is the control force and torque vector, $\Delta \boldsymbol{f}$ is the bounded disturbance vector.

According to [16], the velocity dynamics can be expressed as linear-parametric form

$$M\dot{\boldsymbol{\nu}}_v + C(\boldsymbol{\nu}) + D(\boldsymbol{\nu})\boldsymbol{\nu} + g(\boldsymbol{\eta}) = \Phi^T(\boldsymbol{\nu}, \dot{\boldsymbol{\nu}}_v, \boldsymbol{\eta})\sigma$$

where $\sigma = [m_\nu - X_{\dot{u}}, m_\nu - Y_{\dot{v}}, X_u, X_{|u|u}, Y_v, Y_{|v|v}, m_\nu - Z_{\dot{w}}, Z_w, Z_{w|w|}, W - B, I_z - N_{\dot{r}}, N_r, N_{r|r|}]$ is unknown system parameter vector, $\boldsymbol{\nu}_v$ is the virtual control and $\Phi(\boldsymbol{\nu}, \dot{\boldsymbol{\nu}}_v, \boldsymbol{\eta})$ is a known reduced regressor matrix function. The specific forms of M, $J(\boldsymbol{\eta})$, $C(\boldsymbol{\nu})$, $D(\boldsymbol{\nu})$, $g(\boldsymbol{\eta})$, $\Phi(\boldsymbol{\nu}, \dot{\boldsymbol{\nu}}_v, \boldsymbol{\eta})$ and related system parameters can be found in [16] and are omitted here for saving space.

Consider a ROV formation which consists of 4 ROVs. The communication topology $\mathcal{G}$ is given by a 2-regular graph and the connection weight $w_{ji} = 1$ if v_i and v_j are connected. The objective of multi-agent coordination is to find a distributed control strategy that is able to drive each ROV from its initial position to rendezvous at the target position which minimizes the square sum of distances from these initial positions. The objective can be formulated as the problem: $\min_{\boldsymbol{\eta}^{(j)}} \sum_{j=1}^{4} \|\boldsymbol{\eta}^{(j)} - \boldsymbol{\eta}_0^{(j)}\|^2$, s.t. $\boldsymbol{\eta}^{(1)} = \cdots = \boldsymbol{\eta}^{(4)}$ where $\boldsymbol{\eta}_0^{(j)}$ represents the initial state of the jth ROV.

Consider the cyber attacks (also including the sensor faults or some extraneous factors such as ocean currents) occurring in the complex underwater environment. When the cyber core detects the existence of the cyber attacks, it will drive the attacked ROV to the secure state $\boldsymbol{\eta}_s = 0$. In the simulation, the initial state conditions of these four ROVs are set as $\boldsymbol{\eta}^{(1)}(0) = [0.3\ 0.4\ 1\ 0]^T$, $\boldsymbol{\eta}^{(2)}(0) = [0.1\ 0.1\ 0.5\ -\pi/6]^T$, $\boldsymbol{\eta}^{(3)}(0) = [0\ 0\ 0\ -\pi/8]^T$ and $\boldsymbol{\eta}^{(4)}(0) = [0.2\ 0.5\ 1\ 0]^T$. Assume that the 4th ROV suffers the cyber attack at $t = 30$s, and $\phi^{(4)}(t) = e^{0.5(t-30)-1}[\sin(t)\ \cos(t)\ -\sin(t)\ -\cos(t)]^T$. For simplifying calculation, only the decision logic $\mho^{(j,r)}(t)$ rather than $\mho^{(j)}(t)$ is used in the proposed ADI approach.

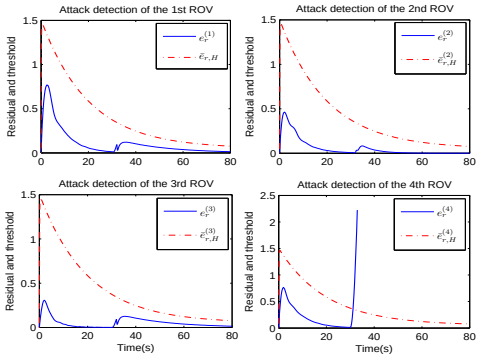

Figure G1 ADI by the decision logic $\mho^{(j,r)}(t)$.

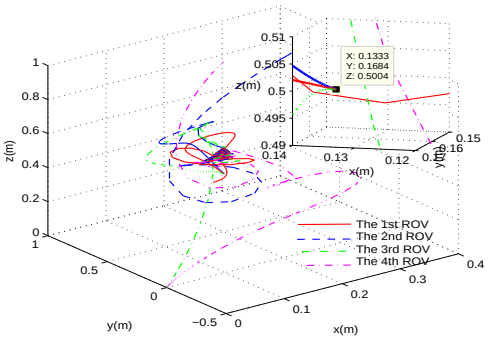

Figure G2 Routes of four ROVs.

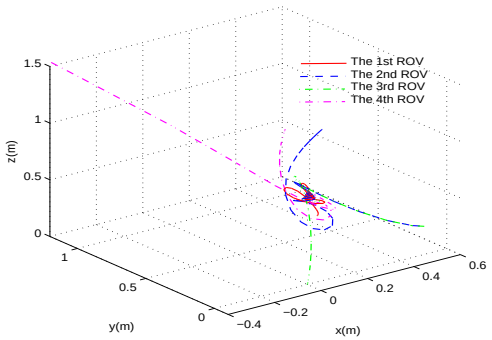

Figure G3 Routes of four ROVs on [0s,34.45s] with the basic DOC.

The ADI mechanism based on the decision logic $\mho^{(j,r)}(t)$ formulated by $e_r^{(j)}$ and $e_{r,H}^{(j)}$ is shown in Figure G1. It can be observed that although the residuals $e_r^{(j)}$, $j = 1, 2, 3$, fluctuate a little after the cyber attack occurs, the decision logics $\mho^{(j,r)}(t)$ generated by $\mathcal{M}^{(j)}$ are still satisfied. Meanwhile, $\mho^{(4,y)}(t)$ is immediately violated (about at $t = 31.5$s), thus indicating the cyber attack occurs in the 4th ROV. After detecting the cyber attack, the module will send the reference $y_r^{(4)} = 0$ to the 4th ROV. Then the ROV converges to the secure position $\boldsymbol{\eta}_s = 0$, while the remaining three ROVs achieve the consensus at the optimal solution of $\min_{\boldsymbol{\eta}^{(j)}} \sum_{j=1}^{3} \|\boldsymbol{\eta}^{(j)} - \boldsymbol{\eta}_0^{(j)}\|^2$, s.t. $\boldsymbol{\eta}^{(1)} = \boldsymbol{\eta}^{(2)} = \boldsymbol{\eta}^{(3)}$. The trajectory curves of these four ROVs in the whole process are illustrated in Figure G2. For comparison, we also apply the existing basic version of DOC (without the ADI mechanism) [17] under adversarial environment. Figure G3 shows that under the case all the ROVs follow the wrong control commands and move along with wrong (insecure) routes due to the attack propagation through the exchange of information between neighboring subsystems (The Simulink reports "ERROR" at $t = 34.45$s and terminates).

References

1 Shames I, Teixeira A, Sandberg H, Johansson K H. Distributed fault detection for interconnected second-order systems. Automatica, 2021, 47: 2757–2764.

2 Zhang Q, Zhang X, Distributed sensor fault diagnosis in a class of interconnected nonlinear uncertain systems. In: Proc 8th IFAC SAFEPROCESS, Mexico City, Mexico, 2012, 1101–1106.

3 Reppa V, Polycarpou M M, Panayiotou C G, Distributed sensor fault diagnosis for a network of interconnected cyber-physical systems. IEEE Trans Control Netw Syst, 2015, 2: 11–23.

4 M. Massoumnia, G. Verghese, and A. Willsky, "Failure detection and identification," IEEE Trans. Autom. Control, vol. 34, no. 3, pp. 316–321, 1989.

5 Y. Liu, M.K. Reiter, and P. Ning, "False data injection attacks against state estimation in electric power grids," in Proc. the 16th ACM conf. Computer communications security, 2009, pp. 21–32.

6 Y. Chen, S. Kar, and J.M.F. Moura, "Optimal attack strategies subject to detection constraints against cyber-physical systems," IEEE Trans. Control Netw. Syst., vol. 5, no. 3, pp. 1157–1168, 2019.

7 A. Teixeira, I. Shamesb, H. Sandberg, K. Henrik Johansson, "Secure control framework for resource-limited adversaries," Automatica, vol. 51, pp. 135–148, 2015.

8 Pasqualetti F, Dorfler F, Bullo F, Attack detection and identification in cyber-physical systems. IEEE Trans Autom Control, 2013, 58: 2715–2729

9 Teixeira A, Shamesb I, Sandberg H, Johansson K H, Secure control framework for resource-limited adversaries. Automatica, 2015, 51: 135–148.

10 A. Barboni, H. Rezaee, F. Boem, and T. Parisini, "Detection of covert cyber-attacks in interconnected systems: a distributed model-based approach," IEEE Trans. Autom. Control, vol. 65, no. 9, pp. 3728–3741, 2020.

11 R. Anguluri, V. Katewa, and F. Pasqualetti, "Centralized versus decentralized detection of attacks in stochastic interconnected systems," IEEE Trans. Autom. Control, vol. 65, no. 9, pp. 3903–3910, 2020.

12 B. Gharesifard and J. Cortes, "Distributed continuous-time convex optimization on weight-balanced digraphs," IEEE Trans. Autom. Control, vol. 59, no. 3, pp. 781–786, Mar. 2014.

13 W. Wang and C. Wen, "Adaptive actuator failure compensation control of uncertain nonlinear systems with guaranteed transient performance," Automatica, vol. 46, pp. 2082–2091, 2010.

14 C. P. Bechlioulis and G. A. Rovithakis, "A low-complexity global approximation-free control scheme with prescribed performance for unknown pure feedback systems," Automatica, vol. 50, no. 4, pp. 1217–1226, 2014.

15 A. Theodorakopoulos and G. A. Rovithakis, "Guaranteeing preselected tracking quality for uncertain strict-feedback systems with deadzone input nonlinearity and disturbances via low-complexity control," Automatica, vol. 54, pp. 135–145, 2015.

16 K. Zhu, L. Gu, "A MIMO nonlinear robust controller for work-class ROVs positioning and trajectory tracking control," in Proc. Annu. Conf. Control Decision, Hangzhou, China, 2011, pp. 2565–2570.

17 Y. Zhang, Z. Deng, and Y. Hong, "Distributed optimal coordination for multiple heterogeneous Euler-Lagrangian systems," *Automatica*, vol. 79, pp. 207–213, 2017.