2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Passivity-Based Attack Detection and Mitigation with Switching Adaptive Controller and Quadratic Storage Function

Pushkal Purohit and Anoop Jain

Abstract— **This paper studies the consensus problem in a networked multi-agent system subject to actuator attacks. The concept of passivity is leveraged for the detection of destabilizing attacks, which relies on satisfying the dissipation-inequality with the quadratic storage functions. On identification of an attack, the controller switches to the defense mode, where the attack signal is mitigated via an estimator based on available measurements. We show that the state error, incurred due to an attack, remains bounded with bounded attack signals, and the system achieves consensus once the attack is mitigated. Simulations are provided to illustrate the theoretical findings.**

I. Introduction

Communication technology and power-efficient microcontrollers have improved exponentially, making cyber-physical systems (CPSs) easy to implement. Distributed control of these large-scale systems uses wireless channels and onboard sensing to share information. This, however, introduces vulnerabilities to malicious attacks through the communication network [1], [2]. In such scenarios, the main challenge is to detect and mitigate attacks early to prevent severe damage to the system. Moreover, it is essential to minimize resource usage, prompting the use of switching controllers that can function in normal or defense modes.

From a control system perspective, some challenges regarding attack detection and mitigation for secure CPSs are discussed in [3] and the references therein. These works primarily characterize different types of possible attacks on a CPS, classified based on the target point of attack, viz. plant, network, sensor, actuator, as summarized in [4]. Though there exist several articles [5]–[11] on this topic, however, there are certain challenges in implementing these works. For instance, [7], [8] require beforehand calculation of gains and matrices to be used in control law for attack detection and mitigation strategies. Specifically, [7], [8] consider that the system is always under attack, resulting in unnecessary consumption of the (already) limited resources in the mobile systems. A switching controller, on the other hand, saves resources when there is no attack and goes defensive under attack to maintain system performance. We adapt this approach in this work by exploiting the ideas from the passivity theory. The proposed scheme in this paper only needs system's current measurement, and we mitigate the attack rather than isolating the compromised agents, unlike [5] and [6].

The authors are with the Department of Electrical Engineering, Indian Institute of Technology Jodhpur, India 342030 (e-mail: purohit.1@iitj.ac.in, anoopj@iitj.ac.in). This work is supported in parts by DST-SERB (Project No. SRG/2020/001112) and Meity projects (Registration No: S/MeitY/SKS/20210139), GoI.

Among the various existing approaches for attack identification and mitigation, observer-based methods have received significant attention in the literature. For instance, the seminal work [12] presented an observer-based attack identification scheme for LTI systems. An attack detector, requiring side initial state information, is proposed in [6]. In [7], an adaptive controller-cum-observer guarantees the ultimate boundedness in a leader-follower framework against data injection attacks. While most of the existing works focus on residual-based and estimation-based attack detectors, these might be vulnerable to stealthy attacks [13], [14]. Note that it is computationally challenging to identify an attack signal using an observer alone, as it has to monitor numerous signals. As a remedy, in this paper, we propose a relatively simple approach relying on passivity theory for attack detection, followed by its mitigation using an observer under the limited availability of output measurements.

Significant research has been devoted to use of passivity theory in designing resilient control for a CPS [15]. Besides providing tools for investigating system stability, passivity theory facilitates an easy attack detection method. Based on the energy balance of the passive systems, an attack detection mechanism using the local and networked monitor is proposed in [10]. In contrast, our approach relies on checking an inequality condition and does not require a separate local monitor. Passivity-based defense mechanism with a switching controller is proposed in [11]. Whereas our method applies to multi-agent systems and does not rely on an observer for attack detection. Our paper considers completely localized independent actuator attack detection and estimation for the networked agents. Further, a switching-based control strategy is proposed to switch controller modes between the *normal* and the *defensive* modes.

Contributions: First, we obtain the condition for passivity for a network of linear systems, cooperating towards consensus, using passivity-inequality with the quadratic storage function. For clarity, we provide analysis for both scenarios when the complete system measurements are available and when limited measurements are available. Second, we give the condition on the attack signal which can be detected and may destabilize the system preventing consensus. Third, we show that, with the proposed estimator, the attack is mitigated and the (relative) state errors remain bounded. We consider no upper limit on the number of agents that can be attacked simultaneously. It is shown that the estimation of the attack signal converges with the actual attack signal.

Preliminaries: The symbols $\mathbb{R}$ and $\mathbb{R}_+$ represent the set of real and non-negative real numbers, respectively. We

979-8-3503-0123-6/23/$31.00 ©2023 IEEE

denote by $\text{diag}\{k_1, \ldots, k_n\} \in \mathbb{R}^{n\times n}$ a diagonal matrix with diagonal entries $k_i, i = 1, \ldots, n$. I_n is an identity matrix of size $n \times n$, and $\mathbf{0}_n \in \mathbb{R}^n$ represents a column vector having all entries 0. $\| \bullet \| \in \mathbb{R}_+$ represents induced 2-norm (resp., Euclidean norm) for any matrix $\bullet \in \mathbb{R}^{m\times n}$ (resp., vector $\bullet \in \mathbb{R}^n$). For any complex number λ, $\Re(\lambda)$ denotes its real part. The Kronecker product of two matrices is denoted by $\otimes$. The Moore–Penrose pseudo inverse of a matrix $M \in \mathbb{R}^{m\times n}$ is $M^\dagger \in \mathbb{R}^{n\times m}$, having the property $MM^\dagger M = M$ [16]. The Laplacian matrix of a graph $\mathcal{G}$ is represented by $L \in \mathbb{R}^{N\times N}$. One may refer to [17] for the properties of Laplacian L.

Lemma 1 ([18]). *Let M be an $n \times n$ matrix with negative real parts of all eigenvalues λ_i, $i = 1, \ldots, n$, then there exists $\gamma > 0$ and $\alpha > 0$ such that $\|e^{Mt}\| \le \gamma e^{-\alpha t}$, $\forall t \ge 0$. In fact, $-\alpha = \max_i \Re(\lambda_i)$.*

Definition 1 (Passive System [19]). *Consider the system*

$$\dot{x} = f(x, u); \quad y = h(x, u), \tag{1}$$

with state $x \in \mathbb{R}^m$, control input $u \in \mathbb{R}^p$, and output $y \in \mathbb{R}^p$. The function f is locally Lipschitz, h is continuous, $f(\mathbf{0}_m, \mathbf{0}_p) = \mathbf{0}_m$ and $h(\mathbf{0}_m, \mathbf{0}_p) = \mathbf{0}_p$. The system (1) is said to be passive if there exists a differentiable storage function $S(x) : \mathbb{R}^m \to \mathbb{R}$, $S(x) \ge 0, S(\mathbf{0}_m) = 0$, such that

$$u^T y \ge \dot{S} = \frac{\partial S}{\partial x} f(x, u), \quad \forall (x, u) \in \mathbb{R}^m \times \mathbb{R}^p. \tag{2}$$

Here, (2) is referred to as the dissipation-inequality, which interprets that the rate of change of stored energy $\dot{S}$ is less than the supplied power $u^T y$. We characterize (2) using the quadratic positive-definite storage function

$$S(x) = (1/2)x^T x, \tag{3}$$

which offers advantages over other storage functions for LTI systems. It is computationally moderate, easy to implement, and resembles energy in physical systems − proportional to the square of the system parameter. Notably, any system with linear dynamics can be represented by a quadratic function as a storage function [20].

II. System and Attack Models, and Problem Description

A. System Model

Consider a multi-agent system comprising N agents

$$\dot{x}_i = A_i x_i + B_i u_i \tag{4a}$$
$$y_i = C_i x_i, \quad i = 1, \ldots, N, \tag{4b}$$

where $x_i \in \mathbb{R}^m, u_i \in \mathbb{R}^p, y_i \in \mathbb{R}^p$ are the state, input, and output vectors, respectively. Further, $A_i \in \mathbb{R}^{m\times m}, B_i \in \mathbb{R}^{m\times p}, C_i \in \mathbb{R}^{p\times m}$ are the system, input, and output matrices, respectively. It is assumed that u_i and y_i are vectors of the same dimension, and (4) satisfies the below properties:

Assumption 1. *The matrix pair (A_i, B_i) is controllable and (A_i, C_i) is observable for all $i = 1, \ldots, N$.*

We consider that the agents are diffusively-coupled according to a fixed and strongly connected directed communication topology having Laplacian $L \in \mathbb{R}^{N\times N}$. A consensus-based control law for such systems is given by [21]

$$u_i = K \sum_{j\in\mathcal{N}_i} (y_j - y_i), \quad \forall i = 1, \ldots, N, \tag{5}$$

where $K > 0$ is a positive gain term and $\mathcal{N}_i$ is set of neighbors of i. For all $i = 1, \ldots, N$, (5) can be represented in the vector-matrix form as:

$$u = -K\bar{L}y, \tag{6}$$

where $\bar{L} \coloneqq (L \otimes I_p) \in \mathbb{R}^{Np\times Np}$ is the extended Laplacian, and $u = [u_1^T, \ldots, u_N^T]^T \in \mathbb{R}^{Np}, y = [y_1^T, \ldots, y_N^T]^T \in \mathbb{R}^{Np}$ are the stacked input and output vectors, respectively. Relying on the passivity assumption of system (4), we have the following result, which will be revisited later:

Lemma 2 (Convergence of Passive systems [21]). *Consider the systems (4) with control (5) and assume that Assumption 1 holds. Suppose the agents are input-output passive with a radially unbounded positive-definite storage function, and the communication graph is strongly connected. In that case, the coupled system (4), (5) is globally stable and the agents' output synchronize, i.e., $\lim_{t\to\infty} \|y_j - y_i\| = 0$, $\forall i, j$.*

B. Attack Model

Lemma 2 relies on the fact that the control signal is directly applied to the controlled agents. However, this may not be true in practice due to any malicious attack on the actuator, causing the applied control signal to be different from the signal generated by the controller, see Fig. 1, where the controller output u_i may be corrupted by an unknown attack $u_i^a \in \mathbb{R}^p$. Consequently, the compromised input can be written as

$$u_i^c = u_i + u_i^a, \tag{7}$$

for any i. Clearly, if there is no attack, i.e., $u_i^a = \mathbf{0}_p$ for every i, $u_i^c = u_i, \forall i$, implying that the overall system achieve state consensus according to Lemma 2. We further consider the following reasonable limitations on the attack signal:

Assumption 2. *The attack signal u_i^a is Lipschitz, that is, $\|u_i^a\| \le \bar{u}^a$ and $\|\dot{u}_i^a\| \le \tilde{u}^a$, for some $\bar{u}^a, \tilde{u}^a \in \mathbb{R}_+$ for all i.*

Such assumption are generally considered in literature, see [7], [22], as the attacker also has limited resources practically.

To facilitate the further analysis, we represent the compromised system (4) for all i, in matrix notations as:

$$\dot{x} = \bar{A}x + \bar{B}u^c = \bar{A}x + \bar{B}(u + u^a), \quad y = \bar{C}x, \tag{8}$$

where $\bar{A} = \text{diag}\{A_i\} \in \mathbb{R}^{mN\times mN}, \bar{B} = \text{diag}\{B_i\} \in \mathbb{R}^{mN\times pN}, \bar{C} = \text{diag}\{C_i\} \in \mathbb{R}^{pN\times mN}$ are the block di-

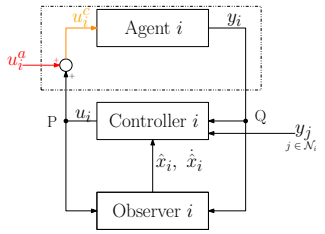


Fig. 1: The i^{th} agent under actuator attack.

agonal matrices, and $x = [x_1^T, \ldots, x_N^T]^T \in \mathbb{R}^{Nm}, u^c = [(u_1^c)^T, \ldots, (u_N^c)^T]^T \in \mathbb{R}^{Np}$ are the stacked state and applied input vectors, respectively. Similarly, $u = [u_1^T, \ldots, u_N^T]^T \in \mathbb{R}^{Np}$ and $u^a = [(u_1^a)^T, \ldots, (u_N^a)^T]^T \in R^{Np}$.

C. Problem Formulation

Our main aim is to design controller u in (8) for achieving consensus in the output, even under an external attack u^a. The approach relies on a passivity-based mechanism to detect attack signals. The controller switches between *normal mode* and *defense mode*, governed by the logic parameter $\delta = 0, 1$, where $\delta = 0$ means "no attack detection" and $\delta = 1$ means "attack detection." This leads to our proposed switching-based control strategy:

$$u_i = u_\delta \coloneqq (1 - \delta)u_i^n + \delta u_i^d, \tag{9}$$

where u_i^n, u_i^d are the control actions in normal (no attack detection) and defense (attack detection) modes. In case of no attack detection (i.e., $\delta = 0$), the control vector is simply chosen as $u = u^n = -K(L \otimes I_p)y$ (see (6)), where $u^n = [(u_1^n)^T, \ldots, (u_N^n)^T]^T \in \mathbb{R}^{Np}$. This results in consensus, as per Lemma 2, for the strongly connected and passive agents with no attack. Thus, it remains to design $u^d = [(u_1^d)^T, \ldots, (u_N^d)^T]^T \in \mathbb{R}^{Np}$ when an attack is detected in the network.

Let $x^n(t)$ and $x^a(t)$ be the solutions of system dynamics (8) under no-attack and attack conditions, respectively. These can be obtained as

$$x^n(t) = \mathrm{e}^{\bar{A}t}x_0 + \int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}u^n(\tau)d\tau \tag{10}$$

$$x^a(t) = \mathrm{e}^{\bar{A}t}x_0 + \int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}(u^d(\tau) + u^a(\tau))d\tau, \tag{11}$$

where $x_0 = x(0)$. We further define

$$e(t) = x^a(t) - x^n(t), \tag{12}$$

as the error between the two state vectors.

Problem. *Consider the network* (8) *under the influence of the attack signal* (7)*, as shown in Fig. 1. Suppose that the agents are strongly connected, and Assumptions 1 and 2 hold. The following problems are addressed:*

(P1) *Devise a passivity-based mechanism to detect an attack modeled by* (7) *on the network* (8)*.*

(P2) *Design the switching control law* u *in* (9) *(i.e.,* u^n *and* u^d*) such that the error* $e(t)$ *in* (12) *remains bounded, that is,* $\|e(t)\| \leq \Omega,\ \forall t \geq 0$*, where* Ω *is an arbitrarily small positive constant.*

It is often the case that all the measurements are not available from the system due to a lack of desired sensors or a non-measurable state. In this work, we consider two cases depending on the availability and non-availability of the state information and its derivative. The former case is discussed for better motivation and easy following of the latter case.

III. Attack Detection: A Passivity-Based Approach with Quadratic Storage Function

This section proposes a passivity-based attack detection approach for attacks that might prevent outputs from reaching a consensus. We motivate the approach by analyzing the case where the complete state information is available. We then extend these ideas for the case when complete state information is not available by incorporating a state observer.

As seen by the controller, the system, enclosed by the dashed rectangle in Fig. 1, should behave as a passive system to assure consensus as per Lemma 2. Alternatively, the inequality (2) must be satisfied concerning points P and Q in Fig. 1. However, the passivity concerning points P and Q might not hold in the presence of an injected attack signal u^a. We leverage this fact to detect the presence of an attack. In this direction, we first discuss the following lemma, which provides a sufficient condition such that the considered network system (8) remains passive under no attack.

Lemma 3. *Under no attack condition, the network* (8) *remains passive with the control law* (6) *and the quadratic storage function* (3) *if and only if matrix* $\mathcal{M} \coloneqq K(\bar{B}\bar{L} - \bar{C}^T\bar{L}^T)\bar{C} - \bar{A} \in \mathbb{R}^{Nm\times Nm}$ *is positive semi-definite, that is, for any* $\zeta \in \mathbb{R}^{Nm}$*,*

$$\zeta^T\mathcal{M}\zeta \geq 0. \tag{13}$$

Proof. For (8) to be passive, it holds from (2) that $u^T y \geq \dot{S}(x)$, where $S(x)$ is the positive definite quadratic storage function (3) with time-derivative $\dot{S}(x) = x^T\dot{x}$. This implies

$$u^T y \geq x^T\dot{x}. \tag{14}$$

Substituting for u, $\dot{x}$, y from (6) and (8), respectively, under no attack condition (i.e., $u^a = \mathbf{0}_{Np}$), we obtain $-Kx^T\bar{C}^T\bar{L}^T\bar{C}x \geq x^T\bar{A}x - Kx^T\bar{B}\bar{L}^T\bar{C}x \implies x^T\left[K\left((\bar{B}\bar{L} - \bar{C}^T\bar{L}^T)\bar{C}\right) - \bar{A}\right]x \geq 0 \implies x^T\mathcal{M}x \geq 0$, for any $x \in \mathbb{R}^{Nm}$. This proves the necessity. If $\mathcal{M}$ is not a positive semi-definite matrix, then it can be trivially shown that (13) does not hold, proving the sufficiency. □

We exploit the passivity property (14) (equivalently (13)) to detect the presence of any malicious attack on the system by considering that the controller is equipped with a device that can verify the dissipation inequality (14) in real-time.

Definition 2. *The attack signal* u^a *that might cause the network* (8) *to satisfy (resp., not to satisfy) inequality* (14) *are referred to as undetectable (resp., detectable) attacks.*

Definition 2 states that undetectable attack signals, which do not violate condition (14), have no impact on the network's consensus properties near the convergence point. These signals remain passive, as Lemma 2 requires. However, detectable attack signals that might violate condition (14) can affect consensus. To identify conditions preventing consensus, we present the following result divided into subsections based on the availability of measurements x and $\dot{x}$ at point Q in Fig. 1.

A. Availability of Measurements x and $\dot{x}$

In this subsection, we present conditions for a destabilizing attack when all measurements x and $\dot{x}$ are readily accessible to the controller at point Q. Also, the control signal sent by the controller at point P is known as it is generated at the designer's end. Following theorem summarizes the results.

Theorem 1. *Let $\mathcal{M}$ be a positive semi-definite matrix defined in Lemma 3. If there exists an attack signal u^a such that*

$$\zeta^T \mathcal{M} \zeta < \zeta^T \bar{B} u^a, \tag{15}$$

for any $\zeta \in \mathbb{R}^{Nm}$, the system (8) *does not satisfy the inequality* (13)*. Consequently, an attack on the network* (8) *is detected as per Definition 2.*

Proof. We prove this by contradiction. Assume that the dissipation-inequality (14) is satisfied across the points P and Q in Fig. 1. Substituting for u, y and $\dot{x}$ from (6) and (8) into (14) in the presence of attack signal u^a, we obtain $x^T[K(\bar{B}\bar{L} - \bar{C}^T\bar{L}^T)\bar{C} - \bar{A}]x \geq x^T\bar{B}u^a \implies x^T\mathcal{M}x \geq x^T\bar{B}u^a$, after substituting for $\mathcal{M}$ from Lemma 3. Now, it can be concluded that any value of u^a, which violates the preceding condition, causes the system (8) to lose passivity and hence prevent consensus of the agents under control law (6), as in Lemma 2. Alternatively, if u^a satisfies the condition (15), the same can be concluded, proving our claim. □

B. Unavailability of Measurements

When the complete state information is not available, the unknown states can be constructed using a state observer, as shown in Figure 1. We consider the following observer dynamics for the complete system:

$$\dot{\hat{x}} = \bar{A}\hat{x} + \bar{B}u - \bar{H}(\hat{y} - y) \tag{16a}$$

$$\hat{y} = \bar{C}\hat{x}, \quad i = 1, \ldots, N, \tag{16b}$$

where $\bar{H} = \text{diag}\{H_i\} \in \mathbb{R}^{mN \times pN}$ with $H_i \in \mathbb{R}^{m \times p}$ being a corrective feedback gain matrix, and $\hat{x} = [\hat{x}_1^T, \ldots, \hat{x}_N^T]^T \in \mathbb{R}^{Nm}$ is the stacked observer state vector with $\hat{x}_i$ being the state estimation for the agent i. Further, the observer gain matrix $\bar{H}$ is selected such that the error $\hat{x} - x$ converges to $\mathbf{0}_{Nm}$ while satisfying the positive semi-definiteness of the matrix $\tilde{\mathcal{M}}$ defined by

$$\tilde{\mathcal{M}} = K(\bar{B}\bar{L} - \bar{C}^T\bar{L}^T + \bar{H})\bar{C} - \bar{A}, \tag{17}$$

which is analogous to $\mathcal{M}$ in Theorem 1. Since the accurate values of the states are necessary for precisely detecting the presence of an attack, we consider the following mild assumption on the initial states of the observer:

Assumption 3. *The initial observer's state is synchronized with the initial system's state, i.e., $\hat{x}(t) \approx x(t), 0 \leq t < t_a$ where t_a is the time of start of an attack on the system* (8)*.*

It is worth noticing that, in terms of the estimated measurements $\hat{y}$, the control law (9) for normal operation, i.e., for $\delta = 0$, is given according to (6) as

$$u = u^n = -K\bar{L}\hat{y}, \tag{18}$$

which will be used in the subsequent analysis. Analogous to Theorem 1, we state the following theorem in case of unavailability of the state measurements.

Theorem 2. *Let $\tilde{\mathcal{M}}$ in* (17) *be a positive semi-definite matrix for appropriately chosen $\bar{H}$. Network* (8) *does not satisfy the inequality* (14) *with respect to observer dynamics* (8) *if the attacking signal u^a fails to satisfy the following inequality:*

$$\hat{x}^T\tilde{\mathcal{M}}\hat{x} \geq \hat{x}^T\bar{H}\bar{C}\left(\int_0^t \bar{A}x d\tau - \int_0^t \bar{B}\bar{L}\bar{C}\hat{x}d\tau + \int_0^t \bar{B}u^a d\tau\right). \tag{19}$$

Consequently, an attack on the network (8) *is detected.*

Proof. For the estimated state $\hat{x}$ and output $\hat{y}$, (14) becomes $u^T\hat{y} \geq \hat{x}^T\dot{\hat{x}}$. Substituting for $\dot{\hat{x}}$, $\hat{y}$ and u from (16) and (18), respectively, and following the steps similar to the proof in Theorem 1, it can be concluded that $\hat{x}^T[K(\bar{B}\bar{L} - \bar{C}^T\bar{L}^T + \bar{H})\bar{C} - \bar{A}]\hat{x} \geq \hat{x}^T\bar{H}\bar{C}x \implies \hat{x}^T\tilde{\mathcal{M}}\hat{x} \geq \hat{x}^T\bar{H}\bar{C}x$. Now, substituting $x(t) = \int_0^t \dot{x}dt$, where $\dot{x}$ is given by (8), we get $\hat{x}^T\tilde{\mathcal{M}}\hat{x} \geq \hat{x}^T\bar{H}\bar{C}[\int_0^t \bar{A}x d\tau - \int_0^t \bar{B}\bar{L}\bar{C}\hat{x}d\tau + \int_0^t \bar{B}u^a d\tau]$. So any value of u^a not satisfying the condition (19) may render the system non-passive and hence, the consensus in states cannot be guaranteed. □

Remark 1. *Though the conditions* (13) *and* (19) *are analytically obtained (through* (14)*), these are not required in implementing the switching control law* (9)*. It is sufficient to verify only* (14) *to know whether the system is passive or not as seen from points P and Q in Fig. 1.*

IV. Attack Mitigation

After detection of an attack, $\delta = 1$ in (9), and hence, $u = u^d$, which needs to be designed for attack mitigation such that error (12) remains bounded, where the switching between normal and defense modes occurs with δ. We again split the discussion into two parts, without a state observer and with a state observer, for clarity and ease of the analysis.

A. Without Observer: Availability of Measurements

Let the control u^d in (9) be proposed as:

$$u^d = -K\bar{L}y - \hat{u}^a, \tag{20}$$

where $\hat{u}^a = [(\hat{u}_1^a)^T, \ldots, (\hat{u}_N^a)^T]^T$ is an estimation of the attack signal u^a, and is given by

$$\hat{u}^a(t) = \bar{B}^\dagger(\dot{x}(t) - \bar{A}x(t) - \bar{B}(-\bar{L}y(t) - \hat{u}^a(t - t_d))), \tag{21}$$

where $\bar{B}^\dagger = (\text{diag}\{B_i\})^\dagger = \text{diag}\{B_i^\dagger\}$, and $t_d > 0$ is a small constant reaction time, representing the time required in observing the effect of control $u(t)$ on the output of the plant. This indicates that an output measurement received by the controller at time t was caused by an input signal sent at time $t - t_d$, which is usually a case in practice. Clearly, $\hat{u}^a$ is Lipschitz, implying that

$$\|\hat{u}^a(t) - \hat{u}^a(t - t_d)\| \leq \eta|t_d|, \tag{22}$$

for some constant $\eta \in \mathbb{R}_+$. Ideally, when $t_d \to 0$, we get an accurate estimation of the attack signal. However, it is not practically feasible due to the controller's and other components' processing time. Note that (21) is a simple algebraic equation motivated by Luenberger observer and does not require the implementation of any dynamical system for attack estimation. We have the following result:

Lemma 4. *As $t_d \to 0$, the difference $\hat{u}^a(t) - u^a(t)$ belongs to the null space of matrix $\bar{B}$, that is, $\lim_{t_d \to 0} \bar{B}\,(\hat{u}^a(t) - u^a(t)) = 0$.*

Proof. Substituting $\dot{x}$ from (8) into (21), yields $\hat{u}^a(t) = \bar{B}^\dagger(\bar{A}x(t)+\bar{B}u^n(t)-\bar{B}\hat{u}^a(t)+\bar{B}u^a(t)-\bar{A}x(t)-\bar{B}u^n(t)+\bar{B}\hat{u}^a(t-t_d)) = \bar{B}^\dagger\bar{B}\hat{u}^a(t) + \bar{B}^\dagger\bar{B}u^a(t) - \bar{B}^\dagger\bar{B}\hat{u}^a(t-t_d)$. Rearranging this, we get $\hat{u}^a(t) - \bar{B}^\dagger\bar{B}u^a(t) = \bar{B}^\dagger\bar{B}(\hat{u}^a(t) - \hat{u}^a(t-t_d))$. By left multiplying by $\bar{B}$ and using the property of generalized matrix inverse, as defined in subsection I, we have $\bar{B}(\hat{u}^a(t) - u^a(t)) = \bar{B}(\hat{u}^a(t) - \hat{u}^a(t-t_d))$. It can be seen that as $t_d \to 0$, the effect of attack will be removed from the system since $\lim_{t_d\to 0} \bar{B}\,(\hat{u}^a(t) - u^a(t)) = 0$, implying that $\hat{u}^a(t) - u^a(t)$ belongs to the null space of $\bar{B}$. □

It is clear that if $\bar{B}$ is full rank matrix, $\hat{u}^a(t) \to u^a(t)$ as $t_d \to 0$. This also indicates that for bounded attack signal $u^a(t)$ and small values of t_d, the attack estimate $\hat{u}^a(t)$ also remains bounded. We now describe the following theorem.

Theorem 3. *Consider network* (8)*, operating under control law* (9) *where u^n and u^d for all i are defined in* (6) *and* (20)*, respectively. If Assumptions 1 and 2 hold and measurements x and $\dot{x}$ are available. Then, the error* (12) *remains bounded.*

Proof. Substituting for u^d and $\hat{u}^a$ from (20) and (21), respectively, into (11), yields

$$x^a(t) = \underbrace{\mathrm{e}^{\bar{A}t}x_0 + \int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}u^n d\tau}_{T_1} + \underbrace{\int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}u^a(\tau)d\tau}_{T_2} - \underbrace{\int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}\bar{B}^\dagger(\dot{x}(\tau)-\bar{A}x(\tau)-\bar{B}(u^n(\tau)-\hat{u}^a(\tau-t_d)))d\tau}_{T_3}. \quad (23)$$

Simplifying the term T_3 in (23) separately, we observe that $T_3 = \int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}\bar{B}^\dagger\bar{B}u^n(\tau)d\tau - \int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}\bar{B}^\dagger\dot{x}(\tau)d\tau + \int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}\bar{B}^\dagger\bar{A}x(\tau)d\tau - \int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}\bar{B}^\dagger\bar{B}\hat{u}^a(\tau-t_d)d\tau$. Using the property $\bar{B}\bar{B}^\dagger\bar{B} = \bar{B}$ of the generalized matrix inverse, we obtain $T_3 = \int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}u^n(\tau)d\tau - \int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}\bar{B}^\dagger\dot{x}(\tau)d\tau + \int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}\bar{B}^\dagger\bar{A}x(\tau)d\tau - \int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}\hat{u}^a(\tau-t_d)d\tau$. Substituting $\dot{x}$ from (8), we have $T_3 = \int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}u^n(\tau)d\tau + \int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}\bar{B}^\dagger\bar{A}x(\tau)d\tau - \int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}\bar{B}^\dagger(\bar{A}x(\tau)+\bar{B}(u^n(\tau)-\hat{u}^a(\tau)+u^a(\tau))d\tau - \int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}\hat{u}^a(\tau-t_d)d\tau$, which on substitution into (23), results in

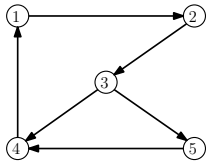
$$x^a(t) = \mathrm{e}^{\bar{A}t}x_0 + \int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}u^n d\tau + \int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}(\hat{u}^a(\tau) - \hat{u}^a(\tau-t_d))d\tau. \quad (24)$$

Further, substituting (10) and (24) into (12), it is clear that $\|e(t)\| = \|\int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}(\hat{u}^a(\tau) - \hat{u}^a(\tau - t_d))d\tau\| \le \int_0^t \|\mathrm{e}^{\bar{A}(t-\tau)}\bar{B}(\hat{u}^a(\tau) - \hat{u}^a(\tau - t_d))\|d\tau \le \int_0^t \|\mathrm{e}^{\bar{A}(t-\tau)}\|\|\bar{B}\|\|(\hat{u}^a(\tau)-\hat{u}^a(\tau-t_d))\|d\tau$, which follows by exploiting the triangle inequality. Using (22), it can be written that $\|e(t)\| \le \eta\|\bar{B}\||t_d|\int_0^t \|\mathrm{e}^{\bar{A}(t-\tau)}\|d\tau$. Now, applying Lemma 1, it follows that $\|e(t)\| \le \alpha\eta\gamma\|\bar{B}\||t_d|(\mathrm{e}^{-\alpha t}+1)$, where $\gamma > 0$ is a constant and $-\alpha = \max\Re(\lambda(\bar{A})) < 0$. □

B. With Observer: Unavailability of Measurements

Analogously to the previous case, the control u^d in (9) is now proposed in terms of the estimated measurements as

$$u^d = -K\bar{L}\hat{y} - \hat{u}^a, \quad (25)$$



(a) Network Topology

$$L = \begin{bmatrix} 1 & 0 & 0 & -1 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}$$

(b) Laplacian

Fig. 2: Communication Topology

where $\hat{y}$ is the estimated output, and $\hat{u}^a$ is given by

$$\hat{u}^a(t) = \bar{B}^\dagger(\dot{\hat{x}}(t) - \bar{A}\hat{x}(t) - \bar{B}(-\bar{L}\hat{y}(t) - \hat{u}^a(t-t_d))). \quad (26)$$

Theorem 4. *Consider network* (8)*, operating under control* (9) *and observer* (16) *where u^n and u^d are defined in* (25) *and* (26)*, respectively. Suppose that Assumptions 1 to 3 hold and measurements x and $\dot{x}$ are not available. Then, the error* (12) *remains bounded.*

Proof. Following an approach similar to the previous case, it can be written using (25) and (26) that

$$\begin{aligned} x^a(t) &= \mathrm{e}^{\bar{A}t}x_0 + \int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}(u^n(\tau) - \hat{u}^a(\tau) + u^a(\tau))d\tau \\ &= \mathrm{e}^{\bar{A}t}x_0 - K\int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}\bar{L}\bar{C}\hat{x}(\tau)d\tau \\ &\quad + \int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}(\hat{u}^a(\tau) - \hat{u}^a(\tau - t_d))d\tau \\ &\quad + \int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}\bar{B}^\dagger\bar{H}\bar{C}(\hat{x}(\tau) - x(\tau))d\tau. \end{aligned} \quad (27)$$

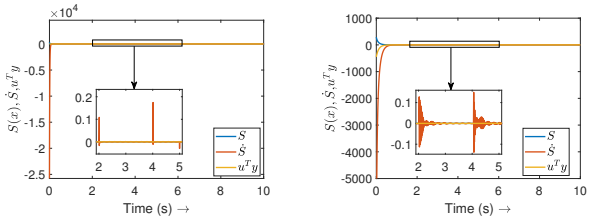
From (10) and (27), it is clear that the error $e(t)$ in (12) satisfies $\|e(t)\| \le \|\int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}(\hat{u}^a(\tau) - \hat{u}^a(\tau - t_d))d\tau + \int_0^t \mathrm{e}^{\bar{A}(t-\tau)}\bar{B}\bar{B}^\dagger\bar{H}(\hat{x} - x)d\tau\| \le \int_0^t \|\mathrm{e}^{\bar{A}(t-\tau)}\bar{B}(\hat{u}^a(\tau) - \hat{u}^a(\tau - t_d))\|d\tau + \int_0^t \|\mathrm{e}^{\bar{A}(t-\tau)}\bar{B}\bar{B}^\dagger\bar{H}(\hat{x} - x)\|d\tau \le \int_0^t \|\mathrm{e}^{\bar{A}(t-\tau)}\|\|\bar{B}\|\|(\hat{u}^a(\tau) - \hat{u}^a(\tau - t_d))\|d\tau + \int_0^t \|\mathrm{e}^{\bar{A}(t-\tau)}\|\|\bar{B}\bar{B}^\dagger\bar{H}\|\|(\hat{x} - x)\|d\tau$. Using (22), it can be concluded that $\|e(t)\| \le \eta\|\bar{B}\||t_d|\int_0^t \|\mathrm{e}^{\bar{A}(t-\tau)}\|d\tau + \|\bar{B}\bar{B}^\dagger\bar{H}\bar{C}\|\int_0^t \|\mathrm{e}^{\bar{A}(t-\tau)}\|\|(\hat{x} - x)\|d\tau$. Now, applying Lemma 1, it yields that $\|e(t)\| \le \alpha\eta\gamma\|\bar{B}\||t_d|(\mathrm{e}^{-\alpha t} + 1) + \gamma\|\bar{B}\bar{B}^\dagger\bar{H}\bar{C}\|\int_0^t \mathrm{e}^{-\alpha(t-\tau)}\|(\hat{x} - x)\|d\tau$, where $\gamma > 0$ is constant and $-\alpha = \max\Re(\lambda(\bar{A}))$. Further, since $\|(\hat{x} - x)\| \le \rho$ for some $\rho > 0$ for an appropriately chosen $\bar{H}$, in conclusion, $\|e(t)\|$ remains bounded for all $t \ge 0$. □

V. Simulation Example

Consider $N = 5$ agents having dynamic matrices as:

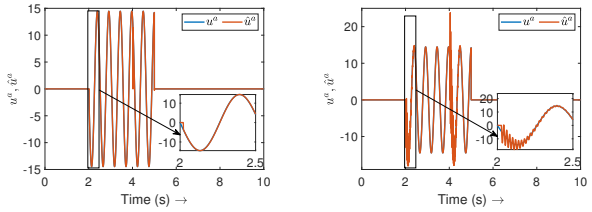
$$A_i = i\begin{bmatrix} -6 & i & 0.5i \\ i & -9 & 0.4i \\ 0.5i & 0.4i & -9 \end{bmatrix},\ B_i = i\begin{bmatrix} 0 \\ 0 \\ 0.5 \end{bmatrix},\ C_i = \frac{1}{i}\begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}^T,$$

where $i = 1, \ldots, 5$, and interacting according to a directed strongly connected topology as shown in Fig. 2. One can easily verify that Assumption 1 holds and the agents are passive. The initial states of the agents are taken randomly in the interval $[-20, 25]$. Further, the initial states of the observer are set the same as the actual system to satisfy Assumption 3. The attack considered on agents is of the form $u_i^a = a_i \sin(\omega_i t)$, where a_i, ω_i are chosen randomly in the



(a) Agent 2, Case I (b) Agent 2, Case II

Fig. 3: Illustration of passivity inequality (2).



(a) u_2^a, $\hat{u}_2^a$, Case I (b) u_2^a, $\hat{u}_2^a$, Case II

Fig. 4: Behavior of attack and estimated attack signals.

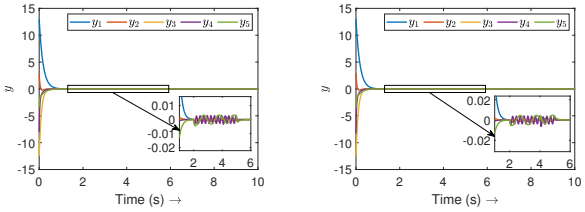
interval $[10, 20]$ and $[0, 10\pi]$, for $i = 1, 2, 4, 5$, respectively. Note that the attack signal is unknown to the controller. We consider agent 3 is not under attack and set $a_3 = 0$. The attack satisfies Assumption 2 and is active for $t \in (2, 5)s$.

Figure 3(a) shows the graph of storage function S, change in stored energy $\dot{S}$, and input energy $(u^T y)$ for Agent 2, in case of availability of measurements (**Case I**). Other attacked agents also show similar behavior and these plots are omitted for brevity. It is clear that the agent loses passivity at $t = 2s$, where the attack starts. Consequently, the attack is detected almost immediately, and the controller switches to the defense mode, according to (9). Figure 4(a) shows the estimated attack signal, which converges to the actual attack signal with almost negligible error for Agent 2. Other agents show similar behavior and their plots are omitted. Figure 5(a) shows the evolution of output of the networked system. It can be seen that the system output remain bounded near the consensus value, as derived in Theorem 3.

Figure 3(b) shows that the agent loses passivity at the beginning of the attack, and the controller switches to the defense mode, in case of non-availability of measurements (**Case II**). The behavior of attack estimation and the output with observer are shown in Figs 4(b) and 5(b), respectively. Similar conclusions can be drawn in this case too.

VI. Conclusions

Leveraging ideas from the passivity theory, we proposed a switching-based control scheme for attack detection and mitigation on a networked linear multi-agent system. Two scenarios were considered, with and without the availability of state measurements. It was shown that the state error remains bounded in both cases. The proposed method is computationally moderate as it does not involve the calculation of any gains, parameters, or coefficients.



(a) Output, Case I (b) Output, Case II

Fig. 5: Output of system (8).

References

[1] J. P. Hespanha, P. Naghshtabrizi, and Y. Xu, "A survey of recent results in networked control systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 138–162, 2007.

[2] M. Wolf and D. Serpanos, "Safety and security in cyber-physical systems and internet-of-things systems," *Proceedings of the IEEE*, vol. 106, no. 1, pp. 9–20, 2017.

[3] S. Tan, J. M. Guerrero, P. Xie, R. Han, and J. C. Vasquez, "Brief survey on attack detection methods for cyber-physical systems," *IEEE Systems Journal*, vol. 14, no. 4, pp. 5329–5339, 2020.

[4] S. M. Dibaji, M. Pirani, D. B. Flamholz, A. M. Annaswamy, K. H. Johansson, and A. Chakrabortty, "A systems and control perspective of CPS security," *Annual reviews in control*, vol. 47, 2019.

[5] D. Zhao, Y. Lv, X. Yu, G. Wen, and G. Chen, "Resilient consensus of higher-order multi-agent networks: An attack-isolation-based approach," *IEEE Transactions on Automatic Control*, 2021.

[6] Y. Chen, S. Kar, and J. M. Moura, "Dynamic attack detection in cyber-physical systems with side initial state information," *IEEE Transactions on Automatic Control*, vol. 62, no. 9, 2016.

[7] M. Meng, G. Xiao, and B. Li, "Adaptive consensus for heterogeneous multi-agent systems under sensor and actuator attacks," *Automatica*, vol. 122, p. 109242, 2020.

[8] X. Jin, W. M. Haddad, and T. Yucelen, "An adaptive control architecture for mitigating sensor and actuator attacks in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 62, no. 11, pp. 6058–6064, 2017.

[9] Y. Joo, Z. Qu, and T. Namerikawa, "Resilient control of cyber-physical system using nonlinear encoding signal against system integrity attacks," *IEEE Transactions on Automatic Control*, vol. 66, no. 9, pp. 4334–4341, 2020.

[10] E. Eyisi and X. Koutsoukos, "Energy-based attack detection in networked control systems," in *Proceedings of the 3rd international conference on High confidence networked systems*, 2014, pp. 115–124.

[11] Y. Yan, P. Antsaklis, and V. Gupta, "A resilient design for cyber physical systems under attack," in *2017 American Control Conference (ACC)*. IEEE, 2017, pp. 4418–4423.

[12] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE transactions on automatic control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[13] A. Khazraei and M. Pajic, "Attack-resilient state estimation with intermittent data authentication," *Automatica*, vol. 138, 2022.

[14] M. Choraria, A. Chattopadhyay, U. Mitra, and E. G. Ström, "Design of false data injection attack on distributed process estimation," *IEEE Transactions on Information Forensics and Security*, vol. 17, 2022.

[15] H. Zakeri and P. J. Antsaklis, "Recent advances in analysis and design of cyber-physical systems using passivity indices," in *2019 27th Mediterranean Conference on Control and Automation (MED)*. IEEE, 2019, pp. 31–36.

[16] C. Rao, S. K. Mitra, and J. Mitra, *Generalized Inverse of Matrices and Its Applications*, ser. Probability and Statistics Series. Wiley, 1971.

[17] C. Godsil and G. F. Royle, *Algebraic graph theory*. Springer Science & Business Media, 2001, vol. 207.

[18] L. Perko, *Differential Equations and Dynamical Systems*, ser. Texts in Applied Mathematics. Springer New York, 2013.

[19] H. Khalil, *Nonlinear Systems*. Prentice Hall, 2002.

[20] H. L. Trentelman and J. C. Willems, "Storage functions for dissipative linear systems are quadratic state functions," in *Proceedings of the 36th IEEE Conference on Decision and Control*, vol. 1. IEEE, 1997.

[21] N. Chopra, "Output synchronization on strongly connected graphs," *IEEE Transactions on Automatic Control*, vol. 57, no. 11, pp. 2896–2901, 2012.

[22] C. Peng and H. Sun, "Switching-like event-triggered control for networked control systems under malicious denial of service attacks," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3943–3949, 2020.