Perfect Attackability of Linear Dynamical Systems with Bounded Noise

Amir Khazraei and Miroslav Pajic

***Abstract*— This paper addresses the problem of secure state estimation in the presence of attacks on sensor measurements of a linear time invariant (LTI) system. We assume that the system is equipped with a common l_0-based attack-resilient state estimator and a sound anomaly detector. We introduce the notion of *perfect attackability* (PA) for LTI systems with bounded noise, when the attacker may introduce an unbounded estimation error while remaining undetected by the anomaly detector. Finally, necessary and sufficient conditions for perfectly attackable systems are provided, and illustrated on examples.**

I. Introduction

The widespread use of dedicated communication networks in modern control systems has increased the number of security related incidents against control of physical processes [1]. With such attacks, a control system's sensory and control unit data may be altered by the attacker, potentially damaging physical components of the system. As a result, the use of control theory techniques to increase resiliency of control systems against attacks has attracted significant attention in recent years.

The main idea has been to utilize knowledge of the system dynamics for attack detection and attack-resilient control (e.g., [2], [3], [4], [5], [6], [7], [8], [9]). For example, consider the problem of attack-resilient control in the presence of false-data injection attacks on system sensors. One line of work focuses on widely-used legacy control architectures that employ standard Kalman filter-based observers for state estimation; in addition, they employ standard residual probability-based detectors, such as $\mathcal{X}^2$ detectors, to detect system anomalies including the presence of attack [3], [10], [11]. For such Kalman filter-based controllers of linear time-invariant (LTI) systems, there are some standard types of attacks that can impose unbounded state estimation errors, while remaining stealthy (e.g., as shown in [3], [10], [4]).

The attack impact depends on the subset (including the number) of compromised sensors. For systems with a Gaussian noise model, where a Kalman filter-based controller and a standard residual probability-based detector are employed, the notion of perfect attackability is introduced in [3], [10]. It is shown in [3], [10] for χ^2 detectors, and in [11] for a more general type of statistical detectors, that such an LTI system is perfectly attackable if and only if its dynamics is unstable and the corresponding unstable eigenvectors satisfy certain properties associated with the set of under attacked sensors. However, it is unclear what are the capabilities (and limitations) of stealthy attackers in systems with bounded noise where statistical detectors cannot be used.

For such systems, the problem of attack-resilient control is commonly mapped into the problem of attack-resilient state estimation [5], [12]. The idea is that the use of resilient state estimation to correctly estimate system states and attack vectors from corrupted sensor measurements, will enable the use of standard feedback controllers even under attack. For noiseless LTI systems for which the knowledge of the model is exact, a resilient state estimator can be formulated as a l_0 optimization problem [5]. SMT-based state estimation technique is also introduced in [6] for a noiseless LTI system. In [7], l_0 and l_1 optimization based estimator is studied for systems with bounded noise, and showed that the worst-case error is linear with the size of the noise. However, the common assumption among all existing works on this topic (e.g., [5], [6], [7], [13]) is that the number of compromised sensors is bounded – at best, only less than a half of sensors should be under attack. Furthermore, the impact of stealthy attacks has never been considered for such systems.

Consequently, in this paper, we consider the problem of resilient state estimation for LTI systems with bounded-size noise, for an arbitrary number of corrupted senors and in the presence of stealthy attacks. We assume that the system is equipped with a l_0-based resilient state estimator (RSE) and an intrusion detector (ID). The main contributions of the paper are as follows. First, we introduce two different notions of perfectly attackability (PA) – PA at a single time point and PA over time for such systems; in PA systems, the attacker is capable of introducing an unbounded estimation error while remaining stealthy. Second, we provide a necessary and sufficient condition for these two notions of PA. We show that unlike PA in the Kalman filter-based estimators, with l_0-based RSEs, a system can be perfectly attackable over time even if the physical plant is not unstable.

The paper is organized as follows. In Section II, we present the system model and formalize our problem. Section III introduces the concept of perfectly attackable systems and find the necessary and sufficient conditions for PA. Finally, in Section IV, we provide a numerical example to illustrate these conditions, before concluding remarks in Section V.

Notation: We use $\mathbb{B}$ and $\mathbb{R}$ to denote the set of Boolean and real numbers, while $\mathbb{I}(.)$ denotes the indicator function. For a matrix A, $\mathcal{N}(A)$ denotes the null space of the matrix, while A^T denotes its transpose, $A^\dagger$ its Moore-Penrose pseudoinverse, and $||A||$ the l_2 norm. For a vector $x \in \mathbb{R}^n$, we denote by $||x||$ its 2-norm. We use x_i to denote the i^{th} element of x, while $\text{supp}(x) = \{i \mid i \in \{1, ..., n\}, x_i \neq 0\}$. Projection

The authors are with the Department of Electrical & Computer Engineering, Duke University, Durham, NC 27708, USA. Email: {amir.khazraei, miroslav.pajic}@duke.edu.

This work is sponsored in part by the ONR under agreements N00014-17-1-2504 and N00014-20-1-2745, AFOSR under the award number FA9550-19-1-0169, as well as the NSF CNS-1652544 grant.

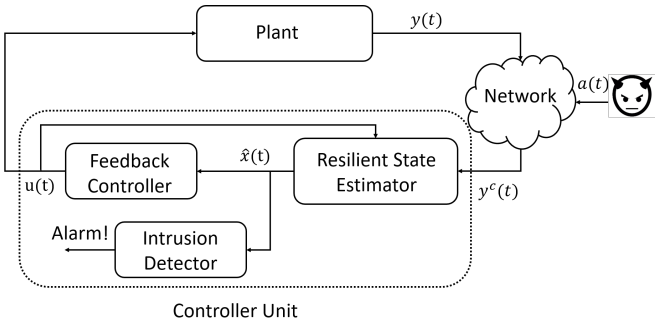

Fig. 1: Control system architecture, considered in this work, in the presence of network-based attacks.

vector e_i is the unit vector (of the appropriate size) with a 1 in its i^{th} position being the only nonzero element of the vector. For set $\mathcal{S}$, $|\mathcal{S}|$ is used to denote the cardinality of the set and $\mathcal{S}^{\complement}$ its complement. We use $\mathcal{P}_{\mathcal{K}}x$ to denote the projection from the set $\mathcal{S}$ to set $\mathcal{K}$ ($\mathcal{K} \subseteq \mathcal{S}$) by keeping only elements of x with indices from $\mathcal{K}$; formally, $\mathcal{P}_{\mathcal{K}} = \left[e_{j_1} | ... | e_{j_{|\mathcal{K}|}} \right]^T$, where $\mathcal{K} = \{s_{j_1}, ..., s_{j_{|\mathcal{K}|}}\} \subseteq S$ and $j_1 < j_2 < ... < j_{|\mathcal{K}|}$.

II. Problem Description

In this section, we introduce the considered system and attack model, as well as formally capture the problem addressed in this work.

A. System and Attack Model

We consider the setup from Figure 1 where each of the components is modeled as follows.

1) Plant Model: We assume that the plant is an observable LTI system modeled in the standard state-space form as

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) + v_p(t), \\ y(t) &= Cx(t) + v_m(t). \end{aligned} \tag{1}$$

Here, $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^p$ are the state, input and output vectors, respectively. The plant output vector captures measurements from the set of plant sensors $\mathcal{S} = \{s_1, s_2, ..., s_p\}$.[1] In addition, $v_p \in \mathbb{R}^n$ and $v_m \in \mathbb{R}^p$ are the process and measurement noises that are assumed to be bounded – i.e., there exist $\delta_{v_p}, \delta_{v_m} > 0$ such that

$$\|v_p(t)\|_2 \leq \delta_{v_p}, \quad \|v_m(t)\|_2 \leq \delta_{v_m}, \qquad \text{for all } t \geq 0; \tag{2}$$

note that we make no assumption regarding the distribution of the sensor and measurement noises.

2) Attack Model: We assume that the attacker has compromised information-flow from a subset of sensors $\mathcal{K} \subseteq \mathcal{S}$;[2] note that we make no assumption about the set $\mathcal{K}$ (e.g., its size or elements). Thus, the sensor measurements delivered to the controller can be modeled as

$$y^c(t) = y(t) + a(t), \tag{3}$$

where $a(t) \in \mathbb{R}^p$ denotes the sparse attack signal injected by the attacker at time t via the compromised information flows (i.e., sensors) from $\mathcal{K}$. Therefore, $\mathcal{K} = \text{supp}(a(t))$.

In this work, we consider the commonly adopted threat model as in e.g., [15], where the attacker has full knowledge of the system, its dynamics and employed architecture. In addition, the attacker has the required computation power to calculate suitable attack signals to inject via the set $\mathcal{K}$, while planning ahead as needed. Finally, the attacker's goal is to design attack signal $a(t)$ such that it always remain *stealthy* – i.e., undetected by the intrusion detection system – while *maximizing control performance degradation*. The notions of *stealthiness* and *control performance degradation* depend on the employed control architecture, and thus will be formally defined after the controller design has been introduced.

3) Controller Design: The controller, illustrated in Figure 1, is equipped with a resilient state estimator (RSE), whose output is used for standard feedback control, as well as an intrusion detector (ID). In what follows, we provide more details on RSE and ID.

To simplify our notation while describing the RSE functionality, the model from (1) can be considered in the form

$$\begin{aligned} x(t+1) &= Ax(t), \\ y(t) = y^c(t) &= Cx(t) + w(t) + a(t), \end{aligned} \tag{4}$$

as we can ignore the contribution of $u(t)$ because it is a known signal (no attacks on actuator are considered in this work), and thus has no effect on the problem of resilient state estimation. As shown in [13], [7], the bounds on the size of measurement noise w in (4) can be related to the bounds on the size of process and measurement noise vectors v_p and v_m – i.e., there exists $\delta_w > 0$ such that

$$||w(t)|| \leq \delta_w, \qquad \text{for all } \ t \geq 0. \tag{5}$$

Resilient State Estimator: The goal of an RSE is to reconstruct the state of the system $x(t)$ from a set of measurements $\{y(t), ..., y(t+N-1)\}$, where N denotes the number of previous sensor measurements that are used for estimation. Throughout the paper, we assume that $N = n$; yet, the results can be extended to the case $N < n$ or $N > n$.

To formally capture requirements for RSE, we rewrite the model from (4) in the form

$$\mathbf{y}(t) = \mathbf{O}x(t) + \mathbf{a}(t) + \mathbf{w}(t), \tag{6}$$

where $\mathbf{O} = [\mathbf{O}_1^T \mid ... \mid \mathbf{O}_p^T]^T$. Here, for each sensor i and a subset of sensors $\mathcal{K}$, we define the matrices $\mathbf{O}_i$ and $\mathbf{O}_{\mathcal{K}}$ as

$$\mathbf{O}_i = \begin{bmatrix} e_i^T C \\ e_i^T CA \\ \vdots \\ e_i^T CA^{N-1} \end{bmatrix}, \quad \mathbf{O}_{\mathcal{K}} = \begin{bmatrix} \mathcal{P}_{\mathcal{K}} C \\ \mathcal{P}_{\mathcal{K}} CA \\ \vdots \\ \mathcal{P}_{\mathcal{K}} CA^{N-1} \end{bmatrix}. \tag{7}$$

Note that each of the block vectors $\mathbf{a}$, $\mathbf{y}$, $\mathbf{w} \in \mathbb{R}^{pN}$, satisfies $\mathbf{a}(t) = [\mathbf{a}_1^T(t) \mid ... \mid \mathbf{a}_p^T(t)]^T$, $\mathbf{y}(t) = [\mathbf{y}_1^T(t) \mid ... \mid \mathbf{y}_p^T(t)]^T$ and $\mathbf{w}(t) = [\mathbf{w}_1^T(t) \mid ... \mid \mathbf{w}_p^T(t)]^T$. Furthermore, for each sensor $i \in \mathcal{S}$, it holds that

$$\mathbf{y}_i(t) = \mathbf{O}_i x(t) + \mathbf{a}_i(t) + \mathbf{w}_i(t), \tag{8}$$

[1] To simplify our notation, unless otherwise stated, we will use i instead of s_i to denote the i-th sensor.

[2] To simplify our presentation, we will refer to these sensors as compromised sensors since the effects of network-based attack are mathematically equivalent to having specific sensors compromised, as described in [14].

with $\mathbf{a}_i(t) = [a_i(t) \mid a_i(t+1) \mid ... \mid a_i(t+N-1)]^T \in \mathbb{R}^N$ denoting the values injected via i^{th} sensor's information flow at time steps $t, ..., t+N-1$, with $\mathbf{a}_i(t) = 0$ if $i \notin \mathcal{K}$. Finally, $\mathbf{y}_i(t) = [y_i(t) \mid y_i(t+1) \mid ... \mid y_i(t+N-1)]^T \in \mathbb{R}^N$ and $\mathbf{w}_i(t) = [w_i(t) \mid w_i(t+1) \mid ... \mid w_i(t+N-1)]^T \in \mathbb{R}^N$ are the values of measurement and measurement noise of sensor i.

In the most general form, the RSE functionality can be captured as [5]

$$\mathcal{E} : \mathbb{R}^{Np} \mapsto \mathbb{R}^n \times \mathbb{R}^{Np} \quad s.t. \quad \mathcal{E}(\mathbf{y}(t)) = \Big(\hat{x}(t), \hat{\mathbf{a}}(t)\Big). \tag{9}$$

Here, $\hat{x}(t)$ and $\hat{\mathbf{a}}(t)$ denote the estimation of the state and attack vectors obtained by the RSE from the delivered sensor measurements. To evaluate performance of an RSE, we define the estimation error as

$$\Delta x(t) = \hat{x}(t) - x(t). \tag{10}$$

A commonly used RSE is the l_0 decoder [7], or its equivalent forms (e.g., [5], [6]). Such RSE may be defined by the following optimization problem

$$\begin{aligned} \min_{\hat{x}(t), \hat{\mathbf{a}}(t)} & \sum_{i=1}^{p} \mathbb{I}\big(\|\hat{\mathbf{a}}_i(t)\| > 0 \big) \\ s.\ t. \quad & \mathbf{y}(t) = \mathbf{O}\hat{x}(t) + \hat{\mathbf{w}}(t) + \hat{\mathbf{a}}(t) \\ & \hat{\mathbf{w}}(t) \in \Omega. \end{aligned} \tag{11}$$

Here, Ω denotes the feasible set of noise vectors, determined by the noise bounds from (5). The vectors $\hat{\mathbf{w}}(t)$ and $\hat{\mathbf{a}}(t)$ are estimated at time t independently from the estimated vectors at time $t-1$. Hence, we may denote $\hat{\mathbf{w}}(t) = [\hat{\mathbf{w}}_1^T(t) \mid ... \mid \mathbf{w}_p^T(t)]^T$, $\hat{\mathbf{a}}(t) = [\hat{\mathbf{a}}_1^T(t) \mid ... \mid \hat{\mathbf{a}}_p^T(t)]^T$, with $\hat{\mathbf{w}}_i(t) = [\hat{w}_i^{(t)}(t)|...|\hat{w}_i^{(t)}(t+N-1)]^T$ and $\hat{\mathbf{a}}_i(t) = [\hat{a}_i^{(t)}(t)|...|\hat{a}_i^{(t)}(t+N-1)]^T$, where $\hat{a}_i^{(t)}(k)$ and $\hat{w}_i^{(t)}(k)$ are the estimated noise and attack vectors at time k, as computed at time t (i.e., for $k = t, ..., t+N-1$).

The RSE from (11) ensures that if less than s number of sensors are compromised in a system that is $2s$-sparse observable [16], then the state will be estimated with a bounded error [7]; whether the system is $2s$sparse observable depends on the observability matrix of (A, C).

Intrusion Detector (ID): The system is equipped with an ID that should detect the presence of system anomalies. As we consider systems with bounded-size noise (and not stochastic noise model), we capture the ID functionality in the general form as mapping $\mathcal{D} : \mathbb{R}^{Np} \mapsto \mathbb{B}$ defined as

$$\mathcal{D}(\hat{\mathbf{a}}(t)) = \mathbb{I}(\|\hat{\mathbf{a}}(t)\| > 0); \tag{12}$$

i.e., if the estimated attack vector is non-zero, the ID will sound an alarm. The threshold in (12) is set to zero due to the fact that for $\hat{\mathbf{a}}(t) \neq 0$ it is impossible to have $\mathbf{a}(t) = 0$, because if $\mathbf{a}(t) = 0$ then $\hat{x}(t) = x(t)$, $\hat{\mathbf{w}}(t) = \mathbf{w}(t)$ and $\hat{\mathbf{a}}(t) = 0$ are feasible solutions that minimize the objective function of optimization problem (11). Note that in this work, we only focus on determining whether the whole system is under attack or not, rather than identifying the exact set of attacked sensors. Designing sound attack identification for the latter problem would require the use of nonzero comparison thresholds in (12), as shown in [7].

Finally, in the rest of this work, we denote the described control system from (4) as $\Sigma(A, C, \delta_w, \mathcal{K})$.

B. Problem Formulation

Our goal is to capture conditions under which a stealthy attacker could introduce unbounded estimation error $\Delta x(t)$, as defined in (10). Here, by an attack being stealthy we assume that under the attack $\mathbf{a}(t)$ the stealthiness condition for the ID

$$\mathcal{D}(\hat{\mathbf{a}}(t)) = 0 \tag{13}$$

holds. Due to the batch-processing nature of the employed RSE, the approach and the derived conditions from [3], [10] cannot be used. Consequently, we start by introducing an equivalent notion of perfectly-attackable systems for LTI dynamical systems with bounded-size noise.

III. Perfect Attackability of Systems with Bounded Noise

To capture the attacker's impact on a system $\Sigma(A, C, \delta_w, \mathcal{K})$, we start with the definition of perfect attackability (PA). In [3], [10], where the notion of PA in the presence of stealthy attacks is first introduced, the considered controller employs a statistical ID (χ^2) and a Kalman-filter that implements continuous (i.e., streamed) processing of sensor measurements. On the other hand, existing RSEs for systems with bounded noise (e.g., [7], [6], [17], [18]) are based on batch-processing of sensor data – i.e., they process a window of sensor measurements at each time step (mostly even without taking previous computations into account).

Consequently, the notion of PA needs to differentiate between PA at a single time point vs. PA over any time interval. In this section, we first define both of these PA notions for any system $\Sigma(A, C, \delta_w, \mathcal{K})$, before providing the necessary and sufficient condition for both types of PA.

Definition 1. *System $\Sigma(A, C, \delta_w, \mathcal{K})$ is perfectly attackable at a single time step if for all $M > 0$, there exists a sequence of attack signals $\mathbf{a}(t)$ over N time steps, such that the RSE's estimation error satisfies $||\Delta x(t)|| > M$, while the attack is stealthy from the ID (i.e.,* (13) *holds). Such an attack vector $\mathbf{a}(t)$ is called a perfect attack for the system $\Sigma(A, C, \delta_w, \mathcal{K})$.*

Definition 1 only describes a perfect situation for the attacker at a single time step t when the system $\Sigma(A, C, \delta_w, \mathcal{K})$ is under attack. This definition means there is no guarantee that the attack remains stealthy at time steps before and/or after time t. In the following definitions, we characterize a more realistic requirements for a stealthy attacker, by considering PA over an interval of time, as well as introduce a notion of a perfect attack vector.

Definition 2. *System $\Sigma(A, C, \delta_w, \mathcal{K})$ is perfectly attackable over time if for all $M > 0$ there exists a sequence of attack signals $\mathbf{a}(t), \mathbf{a}(t+1), ...$ and a time point $t' \geq t$ such that for all k, where $k \geq t'$, it holds that $||\Delta x(k)|| > M$, and*

for all time steps, the estimated attack vectors $\hat{\mathbf{a}}$ satisfies the stealthiness requirements ($\mathcal{D}(\hat{\mathbf{a}}(k)) = 0$) during the attack.

Remark 1. *Note that from the system model, $\Delta x(t)$ effectively depends on a free input argument $\mathbf{a}(t)$ and a bounded input $\mathbf{w}(t)$. Therefore, to simplify our presentation we sometimes state the conditions of Definition 1 as $\Delta x(t)$ being unbounded, meaning that the mapping $\Delta x(\mathbf{a}(t))$ is unbounded for stealthy attacks $\mathbf{a}(t)$.*

From the above definitions, PA over time is a stronger notion than PA at a single time step, because $\mathcal{D}\big(\hat{\mathbf{a}}(t)\big)$ should be equal to zero for all time steps. Therefore, the following directly holds.

Proposition 1. *If system $\Sigma(A, C, \delta_w, \mathcal{K})$ is perfectly attackable over time, then it is also perfectly attackable at a single time step.*

Example 1. *To illustrate PA, consider system $\Sigma(A, C, \delta_w, \mathcal{K})$ with $\delta_w = 0$, $\mathcal{K} = \mathcal{S} = \{s_1, s_2\}$, $N = 2$,*

$$A = \begin{bmatrix} 1 & 1 \\ 1 & .5 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}.$$

It is straightforward to check that attack vector $\mathbf{a}(t) = \begin{bmatrix} a^T(t) & a^T(t+1) \end{bmatrix}^T = \mathbf{O}z$ results in the estimation error $\Delta x(t) = z$ and $\hat{\mathbf{a}}(t) = 0$, for z being any arbitrary nonzero vector; thus may be used as a perfect attack vector.

A. Conditions for Perfect Attackability at a Time Point

The following result captures necessary and sufficient conditions for PA at a single time.

Theorem 1. *The system $\Sigma(A, C, \delta_w, \mathcal{K})$ is perfectly attackable at a single time step if and only if the pair $(A, \mathcal{P}_{\mathcal{K}^\complement} C)$ is not observable.*

Proof. (*Only if:*) For the sake of contradiction, let us assume that the pair $(A, \mathcal{P}_{\mathcal{K}^\complement} C)$ is observable, while the system $\Sigma(A, C, \delta_w, \mathcal{K})$ is perfectly attackable at a single time step, which we denote as t. Then, there exists a stealthy attack sequence $\mathbf{a}(t)$ for which the RSE estimated attack vector $\hat{\mathbf{a}}(t) = 0$ and $\|\Delta x(t)\|$ is unbounded. Consider the impact of attack vectors on measurements of noncompromised sensors from $\mathcal{K}^\complement$. We have that

$$\begin{aligned} \mathcal{P}_{\mathcal{K}^\complement} \mathbf{y}(t) &\overset{\text{(i)}}{=} \mathbf{O}_{\mathcal{K}^\complement} x(t) + \mathcal{P}_{\mathcal{K}^\complement} \mathbf{w}(t) = \\ &\overset{\text{(ii)}}{=} \mathbf{O}_{\mathcal{K}^\complement} \hat{x}(t) + \mathcal{P}_{\mathcal{K}^\complement} \hat{\mathbf{w}}(t), \end{aligned} \tag{14}$$

where (i) holds because the attacker does not inject false data over noncompromised sensors, while (ii) holds since the attack is stealthy (i.e., $\hat{\mathbf{a}}(t) = 0$). Hence, it follows that

$$\mathbf{O}_{\mathcal{K}^\complement} \Delta x(t) = \mathcal{P}_{\mathcal{K}^\complement} \Delta \mathbf{w}(t) \tag{15}$$

where $\Delta \mathbf{w}(t) = \mathbf{w}(t) - \hat{\mathbf{w}}(t)$. Since the the matrix $\mathbf{O}_{\mathcal{K}^\complement}$ is full rank (from $(A, \mathcal{P}_{\mathcal{K}^\complement} C)$ being observable), we have that

$$\Delta x(t) = \big(\mathbf{O}_{\mathcal{K}^\complement}\big)^\dagger \Big(\mathcal{P}_{\mathcal{K}^\complement} \Delta \mathbf{w}(t)\Big), \tag{16}$$

and thus

$$\|\Delta x(t)\| \leq \Big\| \big(\mathbf{O}_{\mathcal{K}^\complement}\big)^\dagger \Big\| \Big(\|\mathcal{P}_{\mathcal{K}^\complement} \Delta \mathbf{w}(t)\| \Big). \tag{17}$$

The matrix $\big(\mathbf{O}_{\mathcal{K}^\complement}\big)^\dagger$ has a bounded norm, and $\mathbf{w}(t)$ and $\hat{\mathbf{w}}(t)$ are also bounded. Thus, the right side of (17) is bounded, which implies that $\Delta x(t)$ is also bounded, contradicting our initial assumption.

(If): Suppose that the pair $(A, \mathcal{P}_{\mathcal{K}^\complement} C)$ is not observable. This means that there exists a nonzero vector z such that $\mathbf{O}_{\mathcal{K}^\complement} z = 0$. Now, let us assume that the system is in state $x(t)$ when attack vector $\mathbf{a}(t) = \begin{bmatrix} \mathcal{P}_{\mathcal{K}} \mathbf{a}(t) \\ \mathcal{P}_{\mathcal{K}^\complement} \mathbf{a}(t) \end{bmatrix} = \mathbf{O}z = \begin{bmatrix} \mathbf{O}_{\mathcal{K}} z \\ 0 \end{bmatrix}$ is applied. For such signal, the information $\mathbf{y}(t)$ delivered to the RSE is captured in (6), for some noise realization $\mathbf{w}(t)$. On the other hand, the output of RSE $(\hat{x}(t), \hat{\mathbf{a}}(t))$ satisfies

$$\mathbf{y}(t) = \mathbf{O}x(t) + \mathbf{w}(t) + \mathbf{a}(t) = \mathbf{O}\hat{x}(t) + \hat{\mathbf{w}}(t) + \hat{\mathbf{a}}(t) \tag{18}$$

Consider $\hat{\mathbf{w}}'(t) = \hat{\mathbf{w}}(t)$, $\hat{x}'(t) = \hat{x}(t) + z$ and $\hat{\mathbf{a}}'(t) = 0$. Now, $(\hat{x}'(t), \hat{\mathbf{w}}'(t), \hat{\mathbf{a}}'(t))$ is a feasible point for the RSE optimization problem from (11) that also minimizes the objective function to zero. Thus, the output of RSE $(\hat{x}(t), \hat{\mathbf{a}}(t))$ also has to have the same value for the objective function, meaning that $\hat{\mathbf{a}} = \mathbf{0}$, and thus the attack will not be detected.

On the other hand, since (A, C) is observable, from (18) and (10) it follows that

$$\Delta x(t) = \mathbf{O}^\dagger \Delta \mathbf{w}(t) + \mathbf{O}^\dagger \mathbf{a}(t) = \mathbf{O}^\dagger \Delta \mathbf{w}(t) + z$$

Since $\Delta \mathbf{w}(t)$ is bounded, and z is any nonzero vector in the null-space of $\mathbf{O}_{\mathcal{K}^\complement}$, it can be chosen to have an arbitrarily large norm. Therefore, the system $\Sigma(A, C, \delta_w, \mathcal{K})$ is perfectly attackable at a single time step. □

In the system model, we assumed that the plant (A, C) is observable. Hence, from Theorem 1 and its proof, the next results directly follow.

Corollary 1. *System $\Sigma(A, C, \delta_w, \mathcal{S})$, i.e., if all sensors are compromised, is perfectly attackable at a single time step.*

Corollary 2. *If an attack vector injected to the system $\Sigma(A, C, \delta_w, \mathcal{K})$ has the form $\mathbf{a}(t) = \mathbf{O}z$, for some $z \in \mathbb{R}^n$, then $\hat{\mathbf{a}}(t) = 0$ and the RSE error satisfies*

$$\Delta x(t) = \mathbf{O}^\dagger \Delta \mathbf{w}(t) + z.$$

These results also provide a starting point for analysis of the notation of PA over time.

B. Conditions for Perfect Attackability over Time

Theorem 1 describes a necessary and sufficient condition for a system to be perfectly attackable at a single time step. The following theorem provides the condition under which system $\Sigma(A, C, \delta_w, \mathcal{K})$ is perfectly attackable over time.

Theorem 2. *Consider the system $\Sigma(A, C, \delta_w, \mathcal{K})$ and let us define the matrix $F(\mathcal{K}, N)$ as*

$$F(\mathcal{K}, N) = \begin{bmatrix} \mathbf{O}_{\mathcal{K}^\complement}^T & (\mathcal{P}_{\mathcal{K}} C)^T & \dots & (\mathcal{P}_{\mathcal{K}} C A^{N-2})^T \end{bmatrix}^T. \tag{19}$$

a) Suppose that the matrix $F(\mathcal{K}, N)$ is not full rank. Then, the system $\Sigma(A, C, \delta_w, \mathcal{K})$ is perfectly attackable over time if and only if the system $\Sigma(A, C, \delta_w, \mathcal{K})$ is perfectly attackable at a single time step.

b) Suppose that the matrix $F(\mathcal{K}, N)$ is full rank. Then the system $\Sigma(A, C, \delta_w, \mathcal{K})$ is perfectly attackable over time if and only if the system $\Sigma(A, C, \delta_w, \mathcal{K})$ is perfectly attackable at a single time step and the matrix A is unstable.

From Theorem 2 it holds that unlike the notion of PA in systems with probabilistic noise and statistical IDs [3], [10], [11], for systems with bounded noise and l_0-based RSEs considered in this work, a system can be perfectly attackable over time even if the matrix A is not unstable. Before proving Theorem 2, we introduce the following lemmas used in the proof – due to space constraint, the proofs of all lemmas can be found in [19]).

Lemma 1. *Consider an attack vector for the system $\Sigma(A, C, \delta_w, \mathcal{K})$ in the form $\mathbf{a}(t) = \mathbf{O}z(t)$, where if $\mathcal{N}(F(\mathcal{K}, N)) = 0$ then $z(t) \notin \mathcal{N}(A)$. If $z(t+1) = Az(t) + \alpha(t)$, where $\alpha(t) \in \mathcal{N}(F(\mathcal{K}, N))$, then $\mathbf{a}(t+1) = \mathbf{O}z(t+1)$ is also a stealthy attack vector for the system $\Sigma(A, C, \delta_w, \mathcal{K})$.*

Lemma 2. *Let the system $\Sigma(A, C, \delta_w, \mathcal{K})$, at two consecutive time steps t and $t+1$ have estimation error $\Delta x(t)$ and $\Delta x(t+1)$, while $\mathcal{D}(\hat{\mathbf{a}}(t)) = \mathcal{D}(\hat{\mathbf{a}}(t+1)) = 0$. Then $\Delta x(t+1) = A\Delta x(t) + \alpha(t) + p(t)$ where $\alpha(t) \in \mathcal{N}(F(\mathcal{K}, N))$ and $p(t)$ is a bounded vector.*

Lemma 3. *Suppose that the system $\Sigma(A, C, \delta_w, \mathcal{K})$ is perfectly attackable at a single time step and $\Delta x(t)$ is bounded while $\hat{\mathbf{a}}(t) = 0$. If the matrix $F(\mathcal{K}, N)$ is full rank, then there exists no attack vector $\mathbf{a}(t+1)$ such that $\Delta x(t+1)$ becomes unbounded while $\hat{\mathbf{a}}(t+1) = 0$.*

Corollary 3. *For the system $\Sigma(A, C, \delta_w, \mathcal{K})$, if the matrix $F(\mathcal{K}, N)$ is full rank, it is impossible to initiate attack with unbounded estimation error while remaining stealthy.*

Proof. Before starting the attack at time t_0, the estimation error is obviously bounded. Now, based on Lemma 3, if the matrix $F(\mathcal{K}, N)$ is full rank, it will be impossible to have unbounded estimation error $\Delta x(t_0)$ while $\hat{\mathbf{a}}(t_0) = 0$ □

Lemma 4. *There exists a nonzero attack vector $\mathbf{a}(t)$ ($\epsilon < \|\mathbf{a}(t)\|$ with $\epsilon > 0$) such that $\mathcal{D}(\hat{\mathbf{a}}(i)) = 0$ for any $t - (N-1) \leq i \leq t + N - 1$.*

We now prove Theorem 2.

Proof of Theorem 2. a) First, assume that the system is perfectly attackable over time. Based on Proposition 1, the system is also perfectly attackable at a single time step.

Inversely, assume that the system is perfectly attackable at a single time step. Suppose that the attack starts at time t_0. Therefore, $\mathcal{D}(\hat{\mathbf{a}}(t)) = 0$ for any $t < t_0 - (N-1)$. The augmented attack vector $\mathbf{a}(t_0 - (N-1))$ can be captured as

$$\mathbf{a}(t_0 - (N-1)) = \begin{bmatrix} 0 \\ \mathcal{P}_{\mathcal{K}}\mathbf{a}(t_0 - (N-1)) \end{bmatrix} \tag{20}$$

where

$$\mathcal{P}_{\mathcal{K}}\mathbf{a}(t_0 - (N-1)) = \begin{bmatrix} 0 & \dots & 0 & (\mathcal{P}_{\mathcal{K}} a(t_0))^T \end{bmatrix}^T. \tag{21}$$

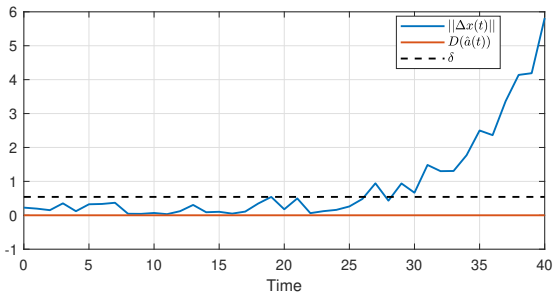

Fig. 2: The evolution of $\Delta x(t)$ and $\mathcal{D}(\hat{\mathbf{a}}(t))$ when the matrix $F(\mathcal{K}, N)$ is full rank. It is assumed that the attack starts at $t = 20$.

Since the matrix $F(\mathcal{K}, N)$ is not full rank, there exists a nonzero vector (referred to as $z(t_0 - (N-1))$) that satisfies $F(\mathcal{K}, N)z(t_0 - (N-1)) = 0$, as well as

$$\begin{aligned} \mathbf{a}(t_0 - (N-1)) &= \begin{bmatrix} 0 & \dots & 0 & (\mathcal{P}_{\mathcal{K}} a(t_0))^T \end{bmatrix}^T \\ &= \begin{bmatrix} \mathbf{O}_{\mathcal{K}^{\complement}} \\ \mathbf{O}_{\mathcal{K}} \end{bmatrix} z(t_0 - (N-1)) = \mathbf{O}z(t_0 - (N-1)). \end{aligned} \tag{22}$$

Here, $z(t_0 - (N-1))$ can be chosen arbitrarily large and this means that $\mathbf{a}(t_0 - (N-1))$ is a perfect attack vector for the system. Now, based on the Lemma 1, the consecutive perfect attack vectors can also be constructed using $\mathbf{a}(t) = \mathbf{O}z(t)$ with $z(t) = A^{t-t_0+(N-1)}z(t_0 - (N-1)) + \sum_{i=t_0}^{t-1} A^{t-i-1}\alpha(i)$ for any $t > t_0 - (N-1)$, where $\alpha(i) \in \mathcal{N}(F(\mathcal{K}, N))$. Since $\alpha(i)$ can be unbounded, the system will have unbounded estimation error for $t \geq t_0 - (N-1)$ while remaining stealthy from ID. Therefore, the system is perfectly attackable over time.

b) (*If*) Let us assume that the matrix A has at least one eigenvalue outside the unit circle. From Lemma 4, we know there exists a nonzero attack vector $\mathbf{a}(t_0)$ such that for any $t_0 - (N-1) \leq i \leq t_0 + N - 1$, $\mathcal{D}(\hat{\mathbf{a}}(i)) = 0$. This means there exists $\epsilon > 0$ such that $\|\mathbf{a}(t_0)\| = \epsilon$. Since $\mathbf{a}(t_0)$ can have any structure, we can assume $\mathbf{a}(t_0) = \mathbf{O}z(t_0)$. Since $\mathbf{O}$ is full rank, $z(t_0)$ can be any nonzero vector that satisfies $\|\mathbf{O}z(t_0)\| \leq \epsilon$; any such vector $z(t_0)$ may be chosen arbitrarily by the attacker.

Based on Lemma 1 if $z(t_0+1) = Az(t_0) + \mathcal{N}(F(\mathcal{K}, N))$, then it is possible to have $\mathbf{a}(t_0 + 1) = \mathbf{O}z(t_0 + 1)$. Since $F(\mathcal{K}, N)$ is full rank, $\mathcal{N}(F(\mathcal{K}, N)) = 0$. By continuing injecting attack vector in the form of $\mathbf{a}(t) = \mathbf{O}z(t)$ for a period of time $[t_0, t]$, we can get $z(t) = A^{t-t_0}z(t_0)$. Let us assume the unstable eigenvalues of the matrix A are diagonizable (the results can be extended to non-diagonizable case), and denote the corresponding unstable eigenvectors of A by $v_1, ..., v_q$. Moreover, let us assume that $z(t_0) = cv_i \neq 0$, such that $\|\mathbf{O}z(t_0)\| \leq \epsilon$. Therefore, we get that $z(t) = A^{t-t_0}z(t_0) = c\lambda_i^{t-t_0}v_i$. Since $|\lambda_i| > 1$, $||z(t)||$ will be unbounded if $t \to \infty$. This implies that, based on the Corollary 2 and Definition 2, the system will be perfectly attackable over time.

(*Only If*) For the sake of contradiction, let us assume that the system is perfectly attackable over time while the matrix

A is stable. From Definition 2, it holds that $\forall M > 0$ there exists a time step t' such that for any $t \geq t'$, $||\Delta x(t)|| > M$.

On the other hand, since the matrix $F(\mathcal{K}, N)$ is full rank, using Corollary 3 it holds that the estimation error should be bounded when the attack starts at time t_0, i.e., $||\Delta x(t_0)|| \leq \delta$ for some $\delta > 0$. Now, for the interval $t_0 < t < t'$ using Lemma 2 we have $\Delta x(t) = A^{t-t_0}\Delta x(t_0) + \sum_{i=t_0}^{t-1} A^{t-i-1}p(i)$. Since the eigenvectors of the matrix A can span the space $\mathbb{R}^n$ (here we assume that the matrix A is diagonizable; however, the results can be easily extended to the undiagonizable case), we have $\Delta x(t_0) = d_1 v_1 + ... + d_n v_n$, $p(i) = d'_{i,1}v_1 + ... + d'_{i,n}v_n$. Now, it holds that

$$\begin{aligned}\|\Delta x(t')\| &= \left\| A^{t'-t_0}\Delta x(t_0) + \sum_{i=t_0}^{t'-1} A^{t'-i-1}p(i) \right\| \\ &= \left\| \sum_{j=1}^{n} d_j \lambda_j^{t'-t_0} v_j + \sum_{i=t_0}^{t'-1}\sum_{j=1}^{n} d'_{i,j}\lambda_j^{t'-i-1} v_j \right\| \\ &\leq \delta + \frac{1}{1-|\lambda_{max}|}||p_{max}||\end{aligned}$$

where λ_{max} is the eigenvalue with the largest absolute value. Therefore, $\|\Delta x(t')\|$ must be bounded, which contradicts our first assumption that the system is perfectly attackable over time. □

IV. Simulation Results

To illustrate the derived conditions for PA of LTI systems with bounded noise, let us consider the system model (4), with $-.1 < w(t) < .1$ and

$$A = \begin{bmatrix} 1.1 & 1 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 1 \end{bmatrix}. \tag{23}$$

The pair (A, C) is observable and the system is unstable with eigenvalues 1.1 and 1. Assume that the system is equipped with the RSE from (11) with $N = 2$. Clearly, the matrix $F(\mathcal{K}, N) = C$ for the RSE is not full rank.

If the attacker starts compromising the sensor values at time 0, it is sufficient to inject $a(0) = CAz(-1)$ where $z(-1) \in \mathcal{N}(C)$. By having $z(-1) = c\begin{bmatrix} 1 & -1 \end{bmatrix}^T$ (where c denotes a scalar with any arbitrary value), the attack vector at time zero and the following time steps is $a(t) = Cz(t)$ and $z(t+1) = Az(t) + \mathcal{N}(C)$ for $t \geq 0$. Therefore, the system is perfectly attackable over time.

Now, let us assume that $N = 3$, meaning that the RSE is using one more sensor measurement to estimate the system state. In this case, the matrix $F(\mathcal{K}, N) = \begin{bmatrix} C^T & A^T C^T \end{bmatrix}^T$ is full rank. However, since the matrix A has an eigenvalue outside the unit circle, from Theorem 2 it follows that the system is perfectly attackable over time.

Figure 2 shows the evolution $\Delta x(t)$ when the attack starts at time 20, and the output of the anomaly detector is zero during the attack. Before the start of the attack, the estimation error is less than a threshold δ, which is the maximum state estimation error when the system is in normal condition and can be obtained from [7]. On the other hand, after the attack starts, the error increases over time.

V. Conclusion

In this paper, we have considered the problem of perfect attackability (PA) of linear-time invariant (LTI) dynamical systems with bounded noise, when attacks can compromise sensor measurements from any subset of plant sensors. First, we have defined PA at single time step and PA over period of time. Then, we have derived necessary and sufficient conditions for these two notions of PA. We have showed that PA over time needs stronger condition than PA at single time step for some class of LTI systems, as well that a system does not have to be unstable to be perfectly attackable.

References

[1] T. Chen and S. Abu-Nimeh, "Lessons from stuxnet," *Computer*, vol. 44, no. 4, pp. 91–93, 2011.

[2] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2009, pp. 911–918.

[3] Mo, Yilin and Sinopoli, Bruno, "False data injection attacks in control systems," in *First workshop on Secure Control Systems*, 2010, pp. 1–6.

[4] R. S. Smith, "Covert misappropriation of networked control systems: Presenting a feedback structure," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 82–92, 2015.

[5] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Transactions on Automatic control*, vol. 59, no. 6, pp. 1454–1467, 2014.

[6] Y. Shoukry, P. Nuzzo, A. Puggelli, A. L. Sangiovanni-Vincentelli, S. A. Seshia, and P. Tabuada, "Secure state estimation for cyber-physical systems under sensor attacks: A satisfiability modulo theory approach," *IEEE Transactions on Automatic Control*, vol. 62, no. 10, pp. 4917–4932, 2017.

[7] M. Pajic, I. Lee, and G. J. Pappas, "Attack-resilient state estimation for noisy dynamical systems," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 1, pp. 82–92, March 2017.

[8] M. Pajic, J. Weimer, N. Bezzo, O. Sokolsky, G. J. Pappas, and I. Lee, "Design and implementation of attack-resilient cyberphysical systems: With a focus on attack-resilient state estimators," *IEEE Control Systems Magazine*, vol. 37, no. 2, pp. 66–81, April 2017.

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

[10] C. Kwon, W. Liu, and I. Hwang, "Analysis and design of stealthy cyber attacks on unmanned aerial systems," *Journal of Aerospace Information Systems*, vol. 11, no. 8, pp. 525–539, 2014.

[11] Jovanov, Ilija and Pajic, Miroslav, "Relaxing integrity requirements for attack-resilient cyber-physical systems," *IEEE Transactions on Automatic Control*, 2019.

[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] M. Pajic, J. Weimer, N. Bezzo, P. Tabuada, O. Sokolsky, I. Lee, and G. Pappas, "Robustness of attack-resilient state estimators," in *ACM Int. Conf. on Cyber-Physical Systems (ICCPS)*, 2014, pp. 163–174.

[14] A. Teixeira, D. Pérez, H. Sandberg, and K. H. Johansson, "Attack models and scenarios for networked control systems," in *First Int. Conf. on High Confidence Networked Systems*, 2012, pp. 55–64.

[15] I. Jovanov and M. Pajic, "Relaxing integrity requirements for attack-resilient cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 64, no. 12, pp. 4843–4858, Dec 2019.

[16] Y. Shoukry and P. Tabuada, "Event-triggered state observers for sparse sensor noise/attacks," *IEEE Transactions on Automatic Control*, vol. 61, no. 8, pp. 2079–2091, 2016.

[17] S. Sundaram, M. Pajic, C. Hadjicostis, R. Mangharam, and G. Pappas, "The Wireless Control Network: Monitoring for Malicious Behavior," in *49th IEEE Conf. on Decision and Control*, 2010, pp. 5979–5984.

[18] S. Sundaram and C. N. Hadjicostis, "Distributed function calculation via linear iterative strategies in the presence of malicious agents," *IEEE Trans. on Automatic Control*, vol. 56, no. 7, pp. 1495–1508, 2011.

[19] A. Khazraei and M. Pajic, "Attack-Resilient State Estimation with Intermittent Data Authentication," Duke University, Tech. Rep., 2020, available at https://cpsl.pratt.duke.edu/publications.