2016 IEEE 55th Conference on Decision and Control (CDC)
ARIA Resort & Casino
December 12-14, 2016, Las Vegas, USA

System Identification in the Presence of Adversarial Outputs

Mehrdad Showkatbakhsh, Paulo Tabuada and Suhas Diggavi

***Abstract*— We consider the problem of system identification of linear time invariant systems when some of the sensor measurements are changed by a malicious adversary. We treat adversaries as omniscient and impose no restrictions (statistical or otherwise) on how they can alter the measurements of the sensors under attack. Given a bound on the number of attacked sensors, and under a certain observability condition, we show that we can construct models that are useful for certain control purposes, e.g., stabilization. We also provide a precise characterization of the equivalence relation that identifies which models cannot be distinguished in the presence of attacks.**

I. Introduction

The recent spate of publicized attacks on cyber-physical systems ranging from cars [1] to infrastructure [2] has led to significant research on security of cyber-physical systems, (see for example, [3], [4], [5], [6] and references therein).

One mechanism advocated in several recent works is to use the properties of the system dynamics to defend against sensor attacks (see for example [7], [8], [9], [10], [11] and references therein). The basic assumption in these papers is that the system model is accurately known to all parties.

In this paper we ask the question of whether one can identify the system despite attacks on sensor measurements. Clearly this can be an ill-posed problem. For instance, consider the system in Figure 1 labeled "attack free" and its attacked version labeled "under attack". The attack consists in changing the output of the p^{th} sensor from $c_p x$ to $c'_p x$. Since the resulting system is still Linear Time-Invariant (LTI), we cannot expect to distinguish the attacked system from the un-attacked system in the bottom of Figure 1 solely based on the (corrupted) measured data. Therefore, we seek to characterize the equivalent class of systems that cannot be distinguished in the presence of attacks. Moreover, we show that despite being impossible to distinguish between two systems in the same equivalence class, we can use the identified model for the purpose of stabilization.

The main result in this paper is a characterization of this equivalence class, for given bounds on the number of sensors attacked as well as a certain observability condition on the underlying system (see Sections II and III for more details). We also demonstrate that identification up to this class is indeed useful, as we can use it to stabilize the underlying system. These results generalize the classical results in system identification without any attacks, where there is a characterization of such an equivalence class (of "similar state-space representations") see for example [12].

This work was supported in part by NSF awards 1136174 and 1321120.

The authors are with the UCLA Electrical Engineering Department, University of California at Los Angeles, CA 90095-1594 { mehrdadsh, tabuada, suhas } @ucla.edu

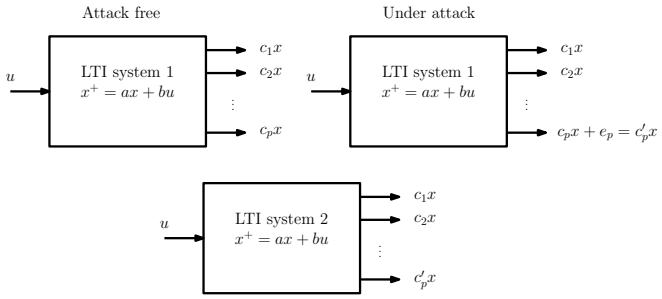

Fig. 1: An example that illustrates the impossibility of exact system identification under adversarial attacks. Consider the system labeled "attack free" and its attacked version labeled "under attack". The attack consists in changing the output of the p^{th} sensor from $c_p x$ to $c'_p x$. Since the resulting system is still LTI, we cannot expect to distinguish the attacked system from the un-attacked system 2 solely based on the (corrupted) measured data.

The paper is organized as follows. In Section II we give the precise formulation of the problem after establishing the notation used in the paper. Section III gives the main result which is proved in Section IV. In Section V we discuss a computational implementation of the proposed results and conclude with a discussion in Section VI. Proofs are given in [13] due to the space constraint.

A. Related work

Among the several different security problems reported in the literature, *e.g.,* denial-of-service [14], [15], [16], [17], man-in-the-middle [18], etc, the results in this paper are closest to the line of research on the secure state estimation problem, [7], [9], [19], [20], [21], [22] and [23]. The problem of secure and resilient state estimation in the presence of malicious agents has recently gained attention, [8], [10], [11], [21] and [24]. Fawzi et. al. [7] considered the problem of control and estimation of LTI systems under adversarial attacks. The authors exploit the dynamics of the system for the identification of attacks. As mentioned, in this paper we study the problem of identifying attacks when the plant is not known. Using a coding theoretic approach, Fawzi [7] investigated conditions under which attack detection is possible and showed this problem to be closely related to observability under the absence of several sensors. This notion was further refined by Shoukry et. al. [23] and called it sparse observability. Independently, Chong et. al. [19] investigated the same problem and introduced the notion of observability under attacks. Mishra et. al. [25] analyzed the noisy version of this problem and identified its optimal

solution for Gaussian noise. Secure state estimation for a class of non-linear plants has been explored recently *e.g.*, [9], [20], [24] and [26].

In another line of work, Tiwari et. al. [27] considered the problem of determining sensor spoofing attacks. Their proposed two-step method does not rely on the dynamics of the system. In the first step, they construct a safety envelope to be used for attack detection. This method relies on the attack-free stream of data for the first step. In contrast, our method can be applied directly to the corrupted data and does not rely on the existence of attack-free data.

II. Preliminaries and problem definition

A. Notation

We represent vectors and real numbers by lower case letters, such as u,x,y, and matrices with capital letters, such as A. The sets of non-negative integers, natural and real numbers are denoted by $\mathbb{N}_0$, $\mathbb{N}$ and $\mathbb{R}$, respectively. For a vector $x\in\mathbb{R}^n$ and $O\subseteq\{1,\dots,n\}$, we denote the vector obtained from x by removing all the elements except those indexed by O by $x|_O$. We denote the size of O by $|O|$. For a given vector space $\mathbb{Y}\subseteq\mathbb{R}^n$, we use the notation $\mathbb{Y}|_O=\cup_{y\in\mathbb{Y}}\{y|_O\}$. We denote the space of polynomial matrices in the indeterminate σ of dimension $a\times b$ by $\mathbb{R}^{a\times b}[\sigma]$.

B. Definitions

In this paper we use many ideas from the behavioral approach to system theory introduced by Willems, see, e.g., [28] and [29]. Since all we have access to is data generated by a system, i.e., behaviors, the behavioral framework provides a natural setting to investigate what can be inferred from data even in the presence of attacks.

Definition 1: **(Time series)** A time series is a map $\mathbf{w}:\mathbb{T}\to\mathbb{W}$, where $\mathbb{T}\subseteq\mathbb{N}_0$ is the time axis and $\mathbb{W}$ is the signal space.

In this paper, $\mathbb{T}=\{0,\dots,T\}$ and $\mathbb{W}=\mathbb{R}^d$, where $T\in\mathbb{N}\cup\{\infty\}$ and d is the dimension of signal space. We represent time series by lower case bold letters, such as $\mathbf{w}$, and the restriction of $\mathbf{w}$ to the i-th component as $\mathbf{w}_i$. For a set $O\subseteq\{1,\dots,d\}$, we denote the time series obtained from $\mathbf{w}$ by removing all the components except those indexed by O by $\mathbf{w}|_O$, i.e., $\mathbf{w}|_O(t):=\mathbf{w}(t)|_O$. In this paper, we use the terms "sequence", "time series" and "trajectory" interchangeably.

For times series $\mathbf{u}:\mathbb{T}\to\mathbb{U}$ and $\mathbf{y}:\mathbb{T}\to\mathbb{Y}$, we use the notation $(\mathbf{u},\mathbf{y})$ to denote the time series $\mathbf{w}:\mathbb{T}\to\mathbb{U}\times\mathbb{Y}$ where $\mathbf{w}(t):=(\mathbf{u}(t),\mathbf{y}(t))$ for $t\in\mathbb{T}$.

We denote the Hankel matrix of time series $\mathbf{w}$ by

$$\mathscr{H}_{i,j}(\mathbf{w}):=\begin{bmatrix}\mathbf{w}(0) & \mathbf{w}(1) & \dots & \mathbf{w}(j-1)\\ \mathbf{w}(1) & \mathbf{w}(2) & \vdots & \mathbf{w}(j)\\ \vdots & \vdots & \vdots & \vdots\\ \mathbf{w}(i-1) & \mathbf{w}(i) & \dots & \mathbf{w}(i+j-2)\end{bmatrix}. \quad (1)$$

We also denote a Hankel matrix by $\mathscr{H}_i(\mathbf{w})$ whenever j takes the maximal possible value, i.e., $\mathscr{H}_i(\mathbf{w})=\mathscr{H}_{i,T-i+1}(\mathbf{w})$.

Definition 2: **(Discrete-time dynamical system)** A discrete-time dynamical system S is a 3-tuple $S=(\mathbb{T},\mathbb{W},\mathscr{B})$, with $\mathbb{T}\subseteq\mathbb{N}_0$ the time axis, $\mathbb{W}$ the signal space, and $\mathscr{B}\subseteq\mathbb{W}^{\mathbb{T}}$ a collection of time series called its behavior. In the context of control systems the signal space is often decomposed into input and output spaces, i.e., $\mathbb{W}:=\mathbb{U}\times\mathbb{Y}$, where $\mathbb{U}$ and $\mathbb{Y}$ denote the input and the output spaces, respectively.

In the remainder of this paper, we refer to discrete-time dynamical systems simply as systems. We say that a system $S=(\mathbb{T},\mathbb{W},\mathscr{B})$ explains a time series $\mathbf{w}$ if $\mathbf{w}\in\mathscr{B}|_{[0,T]}$, where $\mathscr{B}|_{[0,T]}:=\{\Pi_{[0,T]}\mathbf{w}|\mathbf{w}\in\mathscr{B}\}$, and $\Pi_{[0,T]}:\mathbb{W}^{\mathbb{T}}\to\mathbb{W}^{T+1}$ is the natural projection mapping of the time series $\mathbf{w}:\mathbb{T}\to\mathbb{W}$ that restricts it to the subset $\{0,\dots,T\}$ of its domain. We say a map $\Pi:\mathbb{M}\to\mathbb{N}$ is a linear projection if it is linear and surjective.

The notion of system in Definition 2 is quite general. In this paper we focus on systems that are linear and time-invariant.

Definition 3: **(LTI system)** A system $S=(\mathbb{T},\mathbb{W},\mathscr{B})$ is linear when the signal space $\mathbb{W}$ is a vector space and $\mathscr{B}$ is a linear subspace of $\mathbb{W}^{\mathbb{T}}$. System S is time invariant if $\sigma\mathscr{B}\subseteq\mathscr{B}$, where σ is the backward shift operator on the time series $(\sigma\mathbf{w})(t):=\mathbf{w}(t+1)$ and $\sigma\mathscr{B}:=\{\sigma\mathbf{w}|\mathbf{w}\in\mathscr{B}\}$. We say S is Linear Time-Invariant (LTI) if it is both linear and time-invariant.

Consider the difference equation

$$R_0\mathbf{w}(t)+R_1\mathbf{w}(t+1)+\ldots+R_l\mathbf{w}(t+l)=0, \quad (2)$$

where $R_\tau\in\mathbb{R}^{g\times d}$, for $\tau\in\{0,\dots,l\}$. This difference equation (2) induces a dynamical system via the representation

$$\mathscr{B}=\{\mathbf{w}\in\left(\mathbb{R}^d\right)^{\mathbb{N}_0}| \quad (2) \quad \text{holds}\}. \quad (3)$$

One can write (2) compactly in terms of polynomial matrices as $R(\sigma)\mathbf{w}=0$, where $R(z):=R_0+R_1z+\ldots+R_lz^l=0$,

$$\ker\big(R(\sigma)\big):=\{\mathbf{w}\in\left(\mathbb{R}^d\right)^{\mathbb{N}_0}|R(\sigma)\mathbf{w}=0\}. \quad (4)$$

We call representation (4), a kernel representation of the behavior (3). The maximum lag of the kernel representation is the maximum degree of the polynomial matrix, which for (2) is equal to l.

For a given behavior the kernel representation exists[1] but it is not unique, however they are all related by an equivalence relation. In a shortest lag kernel representation, R is row proper, see chapter 7 in [29].

In order to define the complexity of an LTI system for the purpose of identification, one can define the following terms.

Definition 4: **(Lag and order of a behavior)** We define the lag of a behavior, $\mathscr{B}$, as the maximum lag of its shortest lag kernel representation and denote it by $l(\mathscr{B})$. We denote the order of the system by $n(\mathscr{B})$, which is defined as the state dimension of its minimal state-space realization.

We denote the class of LTI systems with m inputs, p outputs and minimal state dimension of at most n by $\mathscr{L}_m^{m+p,n}$

[1]Under the assumption of completeness of the behavior. We refer interested reader to chapter 7 in [29]

Not all behaviors can be identified solely based on input and output trajectories. In order to identify a behavior, a notion of controllability is necessary, see chapter 8 in [29].

Definition 5: **(Controllability)** The system $\mathscr{B}$ is controllable if for any two trajectories $\mathbf{w}_1, \mathbf{w}_2 \in \mathscr{B}$, there is a third trajectory $\mathbf{w} \in \mathscr{B}$ and $t_2 > t_1 > 0$, such that $\mathbf{w}(t) = \mathbf{w}_1(t)$, for all $t \leq t_1$, and $\mathbf{w}(t) = \mathbf{w}_2(t)$, for all $t \geq t_2$.

In order to address the resilience of a system to adversarial attacks, we need to formalize redundancy in the outputs. Observability in the behavioral framework captures this concept.

Definition 6: **(Observability)** Let $(\mathbb{T}, \mathbb{W}_1 \times \mathbb{W}_2, \mathscr{B})$ be a time-invariant dynamical system. Trajectories in $\mathscr{B}$ are written as $(\mathbf{w}_1, \mathbf{w}_2)$. We say $\mathbf{w}_2$ is observable from $\mathbf{w}_1$ if for all $(\mathbf{w}_1, \mathbf{w}_2), (\mathbf{w}_1, \mathbf{w}_2') \in \mathscr{B}$ we have $\mathbf{w}_2 = \mathbf{w}_2'$.

In general, state-space observability (see chapter 3 in [12]) and observability in the behavioral framework are different notions. However, in the special case when $(A, B, \begin{bmatrix} C_1^T & C_2^T \end{bmatrix}^T, D)$ is the minimal realization of the system $S = (\mathbb{T}, \mathbb{U} \times \mathbb{Y}_1 \times \mathbb{Y}_2, \mathscr{B})$, observability of y_2 from (u, y_1) is equivalent to state-space observablity of (A, C_1).

Definition 7: **(Quotient system)** We say $S_Q = (\mathbb{T}, \mathbb{U} \times \mathbb{Y}_Q, \mathscr{B}_Q)$ is a quotient of $S = (\mathbb{T}, \mathbb{U} \times \mathbb{Y}, \mathscr{B})$ if both have the same input space and there exists a linear projection denoted by $\Pi : \mathbb{U} \times \mathbb{Y} \to \mathbb{U} \times \mathbb{Y}_Q$ such that $\mathscr{B}_Q = \Pi\mathscr{B} := \{\mathbf{w} | \exists \mathbf{w}_0 \in \mathscr{B} \quad \text{s.t.} \quad \mathbf{w}(t) = \Pi\mathbf{w}_0(t), \forall t \in \mathbb{T}\}$.

When $\mathbb{Y} \subseteq \mathbb{R}^p$, $O \subseteq \{1, \ldots, p\}$, and Π is the natural projection mapping from $\mathbb{U} \times \mathbb{Y}$ to $\mathbb{U} \times \mathbb{Y}|_O$, we represent this specific quotient subsystem $(\mathbb{T}, \mathbb{U} \times \mathbb{Y}|_O, \Pi\mathscr{B})$ by $S|_O$.

C. Preliminaries

We define the notion of s-sparse observability in the behavioral setting by drawing inspiration from the state-space notion introduced in [23].

Definition 8: **(s-Sparse observability)** System S is s-sparse observable if any s outputs are observable from the input and the remaining outputs.

Given any minimal state-space realization of the system, this definition is equivalent to Definition 3.1 in [23].

Proposition 1: System S is s-sparse observable if for any minimal realization (A, B, C, D), $(A, B, C|_O, D|_O)$ is observable for any subset O of indices with $|O| = p - s$, where p is the number of outputs.

Proof: By definition of s-sparse observability. ∎

The following concepts will be used to characterize identifiability subject to attacks.

Definition 9: **(Parallel composition)** Consider systems $S_i = (\mathbb{T}, \mathbb{U} \times \mathbb{Y}_i, \mathscr{B}_i)$ for $i \in \{1, 2\}$ with the same input space. The parallel composition of S_1 with S_2 is the system $(\mathbb{T}, \mathbb{U} \times \mathbb{Y}_1 \times \mathbb{Y}_2, \mathscr{B})$, where $\mathscr{B}$ is defined by

$$\{(\mathbf{u}, \mathbf{y}_1, \mathbf{y}_2) \in \left(\mathbb{U} \times \mathbb{Y}_1 \times \mathbb{Y}_2\right)^{\mathbb{T}} | (\mathbf{u}, \mathbf{y}_1) \in \mathscr{B}_1, (\mathbf{u}, \mathbf{y}_2) \in \mathscr{B}_2\}. \tag{5}$$

Definition 10: **(Similar modulo outputs)** Two LTI systems, S_1 and S_2, with the same input space are called similar modulo outputs, if both of them have the same order n, and there exists an n-dimensional subspace which is invariant under the dynamics of the parallel composition of S_1 with S_2. We denote this relation by $\sim$.

To the best of the authors' knowledge, similarity modulo outputs has not been introduced before in the literature. This notion will be used in Section III for the purpose of characterizing systems under adversarial attacks.

The following lemma is useful in characterizing similar modulo outputs systems.

Lemma 1: Two systems S_1 and S_2 are similar modulo outputs if and only if for any minimal realizations of S_1 and S_2, denoted by (A, B, C, D) and (A', B', C', D'), respectively, there exists a linear change of coordinates, P, such that

$$\begin{aligned} A' &= PAP^{-1}, \\ B' &= PB, \end{aligned} \tag{6}$$

Proof: The proof is omitted due to the space constraint [13]. ∎

Now, we are ready to prove that similarity modulo outputs is an equivalence relation and divides $\mathscr{L}_m^{m+p,n}$ into equivalence classes.

Proposition 2: Similarity modulo outputs is an equivalence relation.

Proof: Reflexivity and symmetry hold based on the definition. Transitivity is clear by Lemma 1. ∎

In Information Theory, the Hamming distance is typically used for quantifying the error correction capability of codes. We define the Hamming distance between two time series similarly to classical coding theory for detecting the attacks on the outputs. We can think of each time series, such as $\mathbf{w}$, as a code and $\mathbf{w}_i$ for $i \in \{1, \ldots, d\}$ as symbols.

Definition 11: **(Hamming distance)** For two time series $\mathbf{y}$ and $\mathbf{z}$ the Hamming distance between $\mathbf{y}$ and $\mathbf{z}$ is the maximum number of indices, i, such that $\mathbf{y}_i \neq \mathbf{z}_i$.

Note that $\mathbf{y}_i$ is a function. Hence, the equality $\mathbf{y}_i = \mathbf{z}_i$ is to be understood as $\mathbf{y}_i(t) = \mathbf{z}_i(t)$ for all $t \in \{0, \ldots, T\}$.

D. Problem definition

We consider the problem of system identification of LTI systems when sensor measurements are subject to adversarial attacks. The adversary is omniscient and can arbitrarily alter sensor measurements. We impose no assumptions on the signals injected by the adversary. However, we assume that we know an upper bound on the number of attacked sensors.

Assumption 1 (Bound on the number of attacked sensors): We assume that an upper bound s on the number of attacked sensors is given.

The following assumption is required, otherwise even the identification of the system in the absence of attacks becomes an ill-posed problem.

Assumption 2 (Identifiablity given the input sequence): The behavior of the underlying system is identifiable from the unattacked sequence.

For LTI systems, this assumption can be further simplified.

Proposition 3: (see Theorem 8.16 in [29]) The behavior $\mathscr{B} \in \mathscr{L}^w$ is identifiable from the exact data $\mathbf{w} := (\mathbf{u}, \mathbf{y}) \in \mathscr{B}$ if $\mathscr{B}$ is controllable and $\mathscr{H}_{l(\mathscr{B})+n(\mathscr{B})+1}(\mathbf{u})$ is of full row rank.

We refer to the latter condition as persistency of excitation of the input sequence.

Note that applying proposition 3 requires the knowledge of $l(\mathscr{B})$ and $n(\mathscr{B})$ a priori. One should not expect to know these parameters exactly, however, upper bounds are assumed to be known, denoted by $l_{\max}$ and $n_{\max}$, respectively. We can then use these bounds in proposition 3 rather than using the exact values, see chapter 7 in [29].

Now, we are ready to precisely state the problem we are studying. The underlying LTI system is denoted by $S = (\mathbb{N}_0, \mathbb{U} \times \mathbb{Y}, \mathscr{B})$ with $\mathbb{U} = \mathbb{R}^m$ and $\mathbb{Y} = \mathbb{R}^p$. System S is controllable and upper bounds on the lag and order of its behavior are given by $l_{\max}$ and $n_{\max}$, respectively. The available data is the time series $(\mathbf{u}, \mathbf{y})$, where $\mathbf{u}$ is the input sequence and $\mathbf{y}$ is the attacked output sequence. Each output corresponds to one sensor measurement.

The sensor measurements are given by, $\mathbf{y} = \mathbf{y}_S + \mathbf{y}_{\text{attack}}$, where $\mathbf{y}_{\text{attack}}$ is the signal injected by the adversary that can attack a set of sensors $K \subset \{1, \ldots, p\}$ with $|K| \leq s$. We assume the attacker can choose a set K with $|K| \leq s$ of sensors to be attacked and to only attack sensors in this set. This assumption is motivated by the fact that the time it takes for the adversarial agent to attack new sensors is small compared to the time scale of the modern control systems. Although we know the upper-bound s on the cardinality of K, we do not know K. We do not impose any further restrictions on $\mathbf{y}_{\text{attack}}$, and $\mathbf{y}_{\text{attack}}|_K$ can be any arbitrary sequence. Furthermore, by Assumption 2, $\mathscr{H}_{l_{\max}+n_{\max}+1}(\mathbf{u})$ is of full row rank.

Problem statement: Given the sequence $(\mathbf{u}, \mathbf{y})$, we seek answers to the following problems:

1) Identify a model that explains the input-output behavior of the unattacked sensor measurements $(\mathbf{u}, \mathbf{y}|_{\{1,\ldots,p\}\setminus K})$. Note that such a model is not unique.
2) Characterize the equivalence class of models that can explain this sequence.
3) Stabilize the true underlying system using the identified model.

III. Main result

In this section, we briefly present our main result and its implications. In Section IV-B we prove our main result using tools that we develop in Section IV-A.

Theorem 1: The Hamming distance between output trajectories of two $2s$-sparse observable systems that are not similar modulo outputs is at least $2s+1$, provided that Assumption 2 holds for both of them.

In the introduction, we argued that one can only identify the system up to an equivalence class, using the corrupted data. Theorem 1 essentially states that under a $2s$-sparse observability assumption, it is possible to find a model which is closely related to the underlying system via the similarity modulo outputs relation. We elaborate more in Section IV-B.

Theorem 2: Let us denote an s-sparse observable system that explains $(\mathbf{u}, \mathbf{y}|_O)$ by S', where O is any subset of at least $p-s$ sensors such that $(\mathbf{u}, \mathbf{y}|_O)$ can be explained by an s-sparse observable system. System S is similar modulo outputs to S' and any controller that stabilizes S' also stabilizes S, provided that S is an $2s$-sparse observable system and Assumptions 1 and 2 hold.

Theorem 2 essentially states that given the corrupted time series, one can still identify a relevant model that is similar modulo outputs to the underlying system. Furthermore, this model suffices for stabilizing the system, i.e., one can design a stabilizing control law for this hypothetical system and use it for stabilizing the underlying system.

IV. Proofs

A. Preliminary results

In this section, we present several results that will be used to prove the main result in Section III.

The following relationship between observability and similarity modulo outputs will be used to prove the main result.

Proposition 4: The following are equivalent for any system $S = (\mathbb{N}_0, \mathbb{U} \times \mathbb{Y}_1 \times \mathbb{Y}_2, \mathscr{B})$:

1) y_2 is observable from (u, y_1).
2) S and $S_Q = (\mathbb{N}_0, \mathbb{U} \times \mathbb{Y}_1, \mathscr{B}_Q)$ are similar modulo outputs, where $\mathscr{B}_Q = \Pi\mathscr{B}$ and $\Pi : \mathbb{U} \times \mathbb{Y}_1 \times \mathbb{Y}_2 \to \mathbb{U} \times \mathbb{Y}_1$ is the natural projection mapping.

Proof: The proof is omitted due to the space constraint [13]. ■

The following corollary will be used in order to prove the main result.

Corollary 1: Consider an s-sparse observable system $S = (\mathbb{T}, \mathbb{U} \times \mathbb{Y}, \mathscr{B})$. Any quotient system of S, $S|_O := (\mathbb{T}, \mathbb{U} \times \mathbb{Y}|_O, \mathscr{B}|_O)$, where $O \subseteq \{1, \ldots, p\}$ and $|O| = p - s$, is similar modulo outputs to S.

Proof: Based on the definition of s-sparse observability and Proposition 4. See [13] for more detail. ■

B. Proof of Theorem 1

Similarity modulo outputs is an equivalence relation and it divides the set of all LTI systems into equivalence classes. Using the tools developed in Section IV-A, we are ready to prove Theorem 1.

Proof: Assume that the Hamming distance between output trajectories of S_1 and S_2 is less than $2s+1$. We show that S_1 and S_2 are similar modulo outputs. By assumpton the Hamming distance between output trajectories of S_1 and S_2 is less than $2s+1$ and it follows from this assumption that there exist at least $p-2s$ sensors in each system which have the same output sequence for the given input time series. Now we show these $p-2s$ sensors to be enough to conclude that both systems are similar modulo outputs. Without loss of generality we assume these $p-2s$ sensors to be indexed from 1 to $p-2s$ in both systems, we denote the restriction of the outputs to these indices by $\mathbb{Y}_1$. Therefore the output space can be decomposed as $\mathbb{Y}_1 \times \mathbb{Y}_2$, where $\mathbb{Y}_2$ represents the space of the remaining $2s$ outputs. Consider systems $S_i = (\mathbb{T}, \mathbb{U} \times \mathbb{Y}_1 \times \mathbb{Y}_2, \mathscr{B}_i)$ for $i \in \{1,2\}$ and their corresponding quotient systems $Q_i = (\mathbb{T}, \mathbb{U} \times \mathbb{Y}_1, \mathscr{B}_i^Q)$ for $i \in \{1,2\}$, where $\mathscr{B}_i^Q = \Pi_1\mathscr{B}_i$ and Π_1 is the projection map onto the first $p-2s$ coordinates. Proposition 4 implies that

$S_i \sim Q_i$ for $i \in \{1,2\}$. Since the input is sufficiently rich for identification, the quotient systems should have the same behavior, i.e., $Q_1 = Q_2$ and therefore $Q_1 \sim Q_2$. Similarity modulo outputs is an equivalence relation and we conclude that $S_1 \sim Q_1 \sim Q_2 \sim S_2$. ■

C. Proof of Theorem 2

In this section, we show that under a $2s$-sparse observability assumption, it is possible to construct a model that can be used for stabilization of system S. First we show that our method can identify all the attack-free sensors possibly with some sensors that are under a special kind of attacks that we call ineffective. The reason behind this terminology comes from the fact that such attacks cannot prevent us from stabilizing the plant. Using any identification algorithm, one can identify a model for $(\mathbf{u},\mathbf{y}|_O)$, where O is the subset chosen by our method. This model is guaranteed to be similar modulo outputs to S. Furthermore, we show that this model can be used to stabilize the underlying LTI system, i.e., if the controller stabilizes this hypothetical system, it also stabilizes S.

We now explain how this class can help us identify an appropriate set of sensors for the identification. Corollary 1 implies that for any set $O \subseteq \{1,\dots,p\}$ such that $|O| = p-s$, $S|_O$ lies in the same equivalence class as S, i.e., $S|_O \in [S]$. Note that $S|_O$ is an s-sparse observable system. Therefore for any subset of size $p-2s$ of O, denoted by O', we know that $S|_{O'} \sim S|_O$ and therefore $S|_{O'} \sim S$. If the data were attack-free, then any of these subsets are sufficient to reconstruct $[S]$, however, some sensors are under attack.

We start by finding a group of $p-s$ sensors, O, such that for all subsets O' of O with $|O'| = p-2s$, the time series $(\mathbf{u},\mathbf{y}|_{O'})$ can be explained by a system which is similar modulo outputs to $(\mathbf{u},\mathbf{y}|_O)$. Note that by proposition 4 and corollary 1 this is exactly the same as finding any subset, O, such that $(\mathbf{u},\mathbf{y}|_O)$ can be explained by an s-sparse observable system. We denote this system by S' and we claim that $S' \sim S$. At most s sensors are under attack so there exists a subset of size $p-2s$ of O, denoted by O'_{clean}, such that $(\mathbf{u},\mathbf{y}|_{O'_{\text{clean}}})$ is an attack-free time series. Therefore O suffices to find this equivalence class, and it contains at least $p-2s$ attack-free sensors. Note that this set may contain some of attacked sensors, however they are ineffective in misleading us, and still result in the same class. We can further consider other subsets that satisfy this condition and take the union of all such groups for identification. Clearly one subset corresponds to $p-s$ attack-free sensors which satisfies the test.

Now, we are ready to prove Theorem 2.

Proof: Since at most s sensors are under attacks, there exists a set $O \subseteq \{1,\dots,p\}$ with $|O| = p-s$ that $(\mathbf{u},\mathbf{y}|_O)$ can be explained by an s-sparse observable system. Furthermore, there exists a subset of O, denoted by O_{clean} that corresponds to $p-2s$ attack-free sensors. Note that assumption 2 implies that $S'|_{O_{\text{clean}}} = S|_{O_{\text{clean}}}$. Corollary 1 implies that $S' \sim S'|_{O_{\text{clean}}}$ and $S|_{O_{\text{clean}}} \sim S$, therefore $S \sim S'$.

Pick any arbitrary minimal state-space realizations of S and S' denoted by (A,B,C,D) and (A',B',C',D'), respectively. Lemma 1 implies that there exists a linear change of coordinates, P, such that $A' = PAP^{-1}$ and $B = PB'$. Let us denote the state sequences corresponding to these realizations by $\mathbf{x}$ and $\mathbf{x}'$, respectively. According to the definition of similarity modulo outputs and given the fact that both systems have same input sequences, we know that $\mathbf{x}'(t) = P\mathbf{x}(t)$. The controller makes S' asymptotically stable, i.e., $\lim_{t\to\infty} \|\mathbf{x}'(t)\| = 0$. Note that $\|\mathbf{x}(t)\| \leq \|P^{-1}\|\|\mathbf{x}'(t)\|$, so $\lim_{t\to\infty} \|\mathbf{x}(t)\| = 0$. We conclude that the same control input makes S asymptotically stable. ■

V. Computation

In this section we analyze the computational part of our method. As it was mentioned in the last section, we need to check if $(\mathbf{u},\mathbf{y}|_O)$ can be explained by an s-sparse observable system, i.e., any s outputs of the hypothetical system are observable given the input and the remaining outputs.

One way to do so is by constructing the kernel realization of the hypothetical system and then checking its observability.

Instead, we proceed by direct checking the behavioral definition of observability. Let us consider an LTI system $(\mathbb{N}_0, \mathbb{U}\times\mathbb{Y}_1\times\mathbb{Y}_2, \mathscr{B})$, we aim to check if $\mathbf{y}_2$ is observable from $(\mathbf{u},\mathbf{y}_1)$. According to the definition, we need to check the existence of a map from $(\mathbb{U}\times\mathbb{Y}_1)^{\mathbb{N}_0}$ to $\mathbb{Y}_2^{\mathbb{N}_0}$ that maps $(\mathbf{u},\mathbf{y}_1)$ to $\mathbf{y}_2$. Since the underlying system is LTI, this map should be linear and causal. We know the lag of the underlying behavior is upper bounded by $l_{\max}$, therefore we need to check the existence of linear mappings $\mathscr{L}_u : \mathbb{U}^{l_{\max}+1} \to \mathbb{Y}_2$ and $\mathscr{L}_{y_1} : \mathbb{Y}_1^{l_{\max}+1} \to \mathbb{Y}_2$ such that:

$$\begin{aligned}\mathbf{y}_2(t) = \mathscr{L}_u(\mathbf{u}(t),\dots,\mathbf{u}(t-l_{\max})) \\ + \mathscr{L}_{y_1}(\mathbf{y}_1(t),\dots,\mathbf{y}_1(t-l_{\max})), \\ \forall t \in \{l_{\max},\dots,T-1\},\end{aligned} \tag{7}$$

where T is the length of time series.

Lemma 2: Linear mappings $\mathscr{L}_u : \mathbb{U}^{l_{\max}+1} \to \mathbb{Y}_2$ and $\mathscr{L}_{y_1} : \mathbb{Y}_1^{l_{\max}+1} \to \mathbb{Y}_2$ satisfying (7) exist if and only if $\big[\mathbf{y}_2(l_{\max}),\dots,\mathbf{y}_2(T-1)\big]$ lies in the row space of $\mathscr{H}_{l_{\max}+1}(\mathbf{u},\mathbf{y}_1)$.

Proof: The proof is omitted due to the space constraint [13].

■

Algorithm 1 summarizes the proposed method. This algorithm has an exponential worst-case run-time complexity with respect to p and s.

VI. Conclusions and Future Work

A. Conclusions

We considered the problem of system identification of LTI systems under adversarial attacks. We imposed no restriction on the outputs attacked by the adversary. Given a bound on the number of attacked sensors, and under a suitable sparse observability assumption, we showed that it is possible to

Algorithm 1: Pseudo-code of the proposed method.

Let O_i for $i \in \{1,\dots,i_{\max}\}$ denote all possible subsets of size $p-s$ of $\{1,\dots,p\}$;
Let $O_{i,j}$ for $j \in \{1,\dots,j_{\max}\}$ denote all possible subsets of size $p-2s$ of O_i;
input : **u** (Input), **y** (output), $n_{\max}$, $l_{\max}$, s ;
output: $O \subseteq \{1,2,\dots,p\}$;
$O \leftarrow \emptyset$;
for $i \leftarrow 1$ ***to*** i_{max} **do**
  flag $\leftarrow \text{rank}\Big(\mathscr{H}_{l_{\max}+1}\big((\mathbf{u},\mathbf{y}|_{O_i})\big)\Big)$;
  j $\leftarrow 1$;
  while $flag = rank\Big(\mathscr{H}_{l_{max}+1}\big((\mathbf{u},\mathbf{y}|_{O'_{i,j}})\big)\Big)$ **do**
    if $j = j_{max}$ **then**
      $O \leftarrow O_i \cup O$;
      break;
    end
    increment j;
  end
end
return O

construct a meaningful model that enables stabilization of the unknown system. Although such a model is not unique we provided a precise characterization of which models can be distinguished by sensor measurements under attack. We defined the notion of similarity modulo outputs and showed that all of such models are similar modulo outputs. This generalizes the ideas of equivalent systems in classical linear systems theory to the case when there are sensor attacks.

B. Future work

The proposed algorithm has an exponential worst-case run-time complexity with respect to the number of sensors. One direction that we are currently pursuing is leveraging satisfiability modulo theory solvers [20] in order to improve the computational complexity.

REFERENCES

[1] A. Greenberg, "Hackers remotely kill a jeep on the highway, with me in it," *[online] http://www.wired.com/2015/07/hackers-remotely-kill-jeep-highway*, 2015.

[2] R. Langner, "Stuxnet: Dissecting a cyberwarfare weapon," *IEEE Security & Privacy*, vol. 9, no. 3, pp. 49–51, 2011.

[3] A. A. Cárdenas, S. Amin, and S. Sastry, "Research challenges for the security of control systems.," in *HotSec*, 2008.

[4] S. Sundaram, M. Pajic, C. N. Hadjicostis, R. Mangharam, and G. J. Pappas, "The wireless control network: monitoring for malicious behavior," in *49th IEEE Conference on Decision and Control (CDC)*, pp. 5979–5984, 2010.

[5] S. Amin, G. A. Schwartz, and A. Hussain, "In quest of benchmarking security risks to cyber-physical systems," *IEEE Network*, vol. 27, no. 1, pp. 19–24, 2013.

[6] Y. Mo, T. H.-J. Kim, K. Brancik, D. Dickinson, H. Lee, A. Perrig, and B. Sinopoli, "Cyber–physical security of a smart grid infrastructure," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 195–209, 2012.

[7] 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.

[8] Y. Mo, E. Garone, A. Casavola, and B. Sinopoli, "False data injection attacks against state estimation in wireless sensor networks," in *49th IEEE Conference on Decision and Control (CDC)*, pp. 5967–5972, 2010.

[9] A. Teixeira, K. C. Sou, H. Sandberg, and K. H. Johansson, "Secure control systems: A quantitative risk management approach," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 24–45, 2015.

[10] Y. Mo, J. P. Hespanha, and B. Sinopoli, "Resilient detection in the presence of integrity attacks," *IEEE Transactions on Signal Processing*, vol. 62, no. 1, pp. 31–43, 2014.

[11] F. Pasqualetti, F. Dorfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[12] P. J. Antsaklis and A. N. Michel, *Linear systems*. Springer Science & Business Media, 2006.

[13] M. Showkatbakhsh, P. Tabuada, and S. Diggavi, "System identification in the presence of adversarial outputs.," *Technical Report UCLA-CyPhyLab-2016-01, 2016. Electronically available at http://www.cyphylab.ee.ucla.edu/Home/publications*.

[14] M. Zhu and S. Martinez, "On the performance analysis of resilient networked control systems under replay attacks," *IEEE Transactions on Automatic Control*, vol. 59, no. 3, pp. 804–808, 2014.

[15] C. De Persis and P. Tesi, "Input-to-state stabilizing control under denial-of-service," *IEEE Transactions on Automatic Control*, vol. 60, no. 11, pp. 2930–2944, 2015.

[16] D. Senejohnny, P. Tesi, and C. De Persis, "A jamming-resilient algorithm for self-triggered network coordination," *arXiv preprint arXiv:1603.02563*, 2016.

[17] A. Gupta, C. Langbort, and T. Basar, "Optimal control in the presence of an intelligent jammer with limited actions," in *49th IEEE Conference on Decision and Control (CDC)*, pp. 1096–1101, 2010.

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

[19] M. S. Chong, M. Wakaiki, and J. P. Hespanha, "Observability of linear systems under adversarial attacks," in *American Control Conference (ACC)*, pp. 2439–2444, 2015.

[20] Y. Shoukry, M. Chong, M. Wakiaki, P. de Nuzzo, A. D. Sangiovanni-Vincentelli, S. Seshia, J. P. Hespanha, and P. Tabuada, "Smt-based observer design for cyber physical systems under sensor attacks," in *American Control Conference (ACC)*, pp. 2439–2444, 2015.

[21] M. Pajic, J. Weimer, N. Bezzo, P. Tabuada, O. Sokolsky, I. Lee, and G. J. Pappas, "Robustness of attack-resilient state estimators," in *ICCPS'14: ACM/IEEE 5th International Conference on Cyber-Physical Systems (with CPS Week 2014)*, pp. 163–174, 2014.

[22] Y. Mo and B. Sinopoli, "Secure estimation in the presence of integrity attacks," *Automatic Control, IEEE Transactions on*, vol. 60, no. 4, pp. 1145–1151, 2015.

[23] Y. Shoukry and P. Tabuada, "Event-triggered state observers for sparse sensor noise/attacks," *IEEE Transactions on Automatic Control*, 2013.

[24] S. Yong, M. Zhu, and E. Frazzoli, "Resilient state estimation against switching attacks on stochastic cyber-physical systems," in *IEEE International Conference on Decision and Control (CDC)*, 2015.

[25] S. Mishra, Y. Shoukry, N. Karamchandani, S. Diggavi, and P. Tabuada, "Secure state estimation: Optimal guarantees against sensor attacks in the presence of noise," in *IEEE International Symposium on Information Theory (ISIT)*, pp. 2929–2933, 2015.

[26] Y. Shoukry, P. Nuzzo, N. Bezzo, A. L. Sangiovanni-Vincentelli, S. A. Seshia, and P. Tabuada, "Secure state reconstruction in differentially flat systems under sensor attacks using satisfiability modulo theory solving," in *IEEE 54th Annual Conference on Decision and Control (CDC)*, pp. 3804–3809, 2015.

[27] A. Tiwari, B. Dutertre, D. Jovanović, T. de Candia, P. D. Lincoln, J. Rushby, D. Sadigh, and S. Seshia, "Safety envelope for security," in *ACM Proceedings of the 3rd international conference on High confidence networked systems*, pp. 85–94, 2014.

[28] J. C. Willems and J. W. Polderman, *Introduction to mathematical systems theory: a behavioral approach*, vol. 26. Springer Science & Business Media, 2013.

[29] I. Markovsky, J. C. Willems, S. Van Huffel, and B. De Moor, *Exact and approximate modeling of linear systems: A behavioral approach*, vol. 11. SIAM, 2006.