Fiftieth Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
October 1 - 5, 2012

Revealing Stealthy Attacks in Control Systems

André Teixeira*, Iman Shames†, Henrik Sandberg*, and Karl H. Johansson*

Abstract— **In this paper the problem of revealing stealthy data-injection attacks on control systems is addressed. In particular we consider the scenario where the attacker performs zero-dynamics attacks on the system. First, we characterize and analyze the stealthiness properties of these attacks for linear time-invariant systems. Then we tackle the problem of detecting such attacks by modifying the system's structure. Our results provide necessary and sufficient conditions that the modifications should satisfy in order to detect the zero-dynamics attacks. The results and proposed detection methods are illustrated through numerical examples.**

I. Introduction

Critical-infrastructure security is of utmost importance in modern society and has been a major concern in recent years. The increasing complexity of these systems and the desire to improve their efficiency and flexibility has led to the use of heterogeneous IT infrastructures that support the timely exchange of data among and across different system layers, from the corporate level to the local control level. Furthermore, IT infrastructures are composed of heterogeneous components from several vendors and often use non-proprietary communication networks. Therefore the amount of cyber threats to these IT infrastructures has greatly increased over the past years, given the larger number of possible attack points across the several system layers. A good illustration of this phenomena may be found in the following article [1] about the search engine Shodan that successfully identified several devices connected to the internet, including components of industrial control systems.

Critical-infrastructures are also more vulnerable to cyber threats, given their tight coupling to IT infrastructures. There are several examples of cyber threats being exploited by attackers to disrupt the behavior of physical processes, for instance the staged attack on a power generator [2] or the more recent Stuxnet virus attack on centrifuges' control system [3], [4]. Hence monitoring and mitigating cyber attacks to these systems has become of the utmost importance, since they may bring disastrous consequences to society. This is well illustrated by recalling the consequences of the US-Canada 2003 blackout [5], partially due to lack of awareness in the control center.

*A. Teixeira, H. Sandberg, and K. H. Johansson are with the ACCESS Linnaeus Centre, KTH - Royal Institute of Technology, Automatic Control, Stockholm, Sweden.
{andretei,hsan,kallej}@kth.se
†I. Shames is with the Department of Electrical and Electronic Engineering, University of Melbourne, Australia
{iman.shames}@unimelb.edu.au
This work was supported in part by the European Commission through the HYCON2 project, the Swedish Research Council under Grants 2007-6350 and 2009-4565, the Swedish Foundation for Strategic Research, and the Knut and Alice Wallenberg Foundation.

A particular type of complex cyber attack is that of false-data injection, where the attacker introduces corrupted data in the communication network. Several instances of this scenario have been considered in the context of control systems, see [6], [7], [8] and references therein.

In this paper we address stealthy false-data injection attacks that are constructed so that they cannot be detected based on control input and measurement data. These attacks have been recently addressed from a system theoretic perspective. In [9] the author characterizes the set of attack policies for covert (stealthy) false-data injection attacks with detailed model knowledge and full access to all sensor and actuator channels, while [10] described the set of stealthy false-data injection attacks for omniscient attackers with full-state information, but possibly compromising only a subset of the existing sensors and actuators.

Recently, an instance of stealthy false-data injection attacks has been performed on an experimental networked control system testbed [11]. The experiment showed that, although the attack is initially hard to detect, it is in fact detected when the system dynamics change due to physical limitations such input saturation. Hence changes in the system dynamics could be used to reveal stealthy false-data attacks. In essence, this approach is similar to the method proposed in [12] to detect replay attacks, in which an auxiliary signal unknown to the attacker is used to excite the system.

Contributions and outline

The set of open-loop stealthy attacks is considered in this paper. The attack is open-loop in the sense that no online information is used to construct the attack. As such the attack policy is defined in terms of the available *a priori* information, namely the dynamical model of the system. This class of attacks is shown to be characterized by a property of the system known as zero-dynamics, thus we denote it as the class of zero-dynamics attacks.

Using a geometric control framework, the system under a zero-dynamics attack is characterized as an autonomous dynamical system with a given initial condition. Furthermore, the attack detectability is cast as an observability property of the autonomous system previously derived. These two steps provide the basis of our results.

It is shown that zero-dynamics attacks may not be completely stealthy since they require the system to be at a non-zero initial condition. The effects of initial condition mismatch are then characterized and it is shown that they can be made arbitrarily small. The problem of changing the system structure to reveal the attacks is then considered.

978-1-4673-4539-2/12/$31.00 ©2012 IEEE

Specifically, we analyze how separately changing the outputs, system dynamics, and inputs affects the attacks' stealthiness. For each component, we characterize classes of changes that reveal attacks, as well as those that do not. Regarding changes on the system outputs, we provide an algorithm to reveal all attacks by incrementally adding new measurements. As for the inputs, we characterize the output effect of a scalar multiplicative perturbation to the inputs, assuming it remains unknown to the attacker. This particular perturbation can be interpreted as a coding or encryption scheme between the controller and actuator, having the scalar factor as their shared private key. Moreover, the corresponding contribution to the output energy is quantified as a function of the augmented system state, which can be used to determine a suitable scaling factor.

The outline of the paper is as follows. The control system architecture and model under attack are described in Section II. Section III follows with a geometric control characterization of zero-dynamics attacks and the effects of non-zero initial conditions are analyzed in Section IV. Different strategies to reveal zero-dynamics attacks are then proposed and analyzed in Section V, followed by numerical examples illustrating our results. Summary and conclusions follow in Section VII.

II. Control System under False-Data Injection Attacks

In this section we describe the networked control system structure, where we consider three main components: the physical plant and communication network, the feedback controller, and the anomaly detector.

The physical plant is modeled in a discrete-time state-space form,

$$\mathcal{P}: \begin{cases} x_{k+1} = Ax_k + Bu_k + B_a a_k \\ y_k = Cx_k + D_a a_k \end{cases}, \tag{1}$$

where $x_k \in \mathbb{R}^n$ is the state variable, $u_k \in \mathbb{R}^q$ the control actions applied to the process, $y_k \in \mathbb{R}^p$ the measurements from the sensors, and $a_k \in \mathbb{R}^d$ the false-data injection attack vector at the sampling instant $k \in \mathbb{Z}$. The system is considered to be in nominal behavior if $a_k = 0$ for all $k \geq 0$.

In order to comply with performance requirements in the presence of the unknown process and measurement noises, we consider that the physical plant is controlled by an appropriate linear time-invariant output feedback controller [13] described as

$$u_k = \mathcal{F}(y_k). \tag{2}$$

An anomaly detector that monitors deviations from the nominal behavior is also considered. The anomaly detector is collocated with the controller and therefore it only has access to y_k and u_k to evaluate the behavior of the plant. The anomaly detector is then modeled as

$$r_k = \mathcal{D}(u_{k-1}, y_k), \tag{3}$$

where $r_k \in \mathbb{R}^m$ is the residue that is evaluated in order to detect and locate existing anomalies. In particular, an alarm is triggered if the residue meets

$$\|r_k\| \geq \tau, \tag{4}$$

where $\tau \in \mathbb{R}^+$ is chosen according to a suitable trade-off between detection and false alarm rates.

Since all the system components are linear and time-invariant, the state of the system can be decomposed as $x_k = \bar{x}_k + x_l^a$, where $\bar{x}_k$ is the component of the system under no attack and x_k^a the component induced by the attack. Furthermore, assuming the attack starts at $k = k_0$ and having $\bar{x}_{k_0} = x_{k_0}$ and $x_{k_0}^a = 0$, the state component under attack is modeled by

$$\mathcal{P}: \begin{cases} x_{k+1}^a = Ax_k^a + Bu_k^a + B_a a_k \\ y_k^a = Cx_k^a + D_a a_k \end{cases}, \; x_{k_0}^a = 0 \tag{5}$$

with $u_k^a = \mathcal{F}(y_k^a)$ and $u_{k_0}^a = 0$.

A. Stealthy attacks

Denoting $\mathcal{A}_{k_0}^{k_f} = \{a_{k_0}, \ldots, a_{k_f}\}$ as the attack signal, the set of stealthy attacks are defined as follows.

Definition 1: The attack signal $\mathcal{A}_{k_0}^{k_f}$ is α-stealthy with respect to $\mathcal{D}$ if $||r_k|| \leq \alpha \quad \forall k \geq k_0$.

The particular subset of 0-stealthy attacks is characterized in the following lemma:

Lemma 1: Let y_k^a be the output of the system (5) with $x_{k_0}^a = 0$ and $u_{k_0}^a = 0$. The attack signal $\mathcal{A}_{k_0}^{k_f}$ is 0-stealthy with respect to any output feedback controller $\mathcal{F}$ and anomaly detector $\mathcal{D}$ if $y_k^a = 0, \quad \forall k \geq k_0$.

The set of 0-stealthy attacks satisfying the conditions in Lemma 1 results in trajectories of the system that do not affect y_k^a, and thus result in $u_k^a = 0$ for all $k \geq k_0$. For linear systems the 0-stealthy attack signals are related to the output zeroing problem or zero-dynamics studied in the control theory literature [14], which we revisit in the next section. For the sake of notation, in the remainder of the paper we drop the superscript when referring to system variables under attack. Additionally, the results presented in the following sections do not consider the influence of the feedback controller. However the results can be generalized by considering the augmented system composed by the plant and controller dynamics, which is subject to future work.

B. Attacker model

In this work we consider the attacker model for zero-dynamics attacks described in [11]. In this model the attacker is also able to inject false data in the actuator channels, which is captured by having $B_a = B$ and $D_a = 0$. However, the attacker cannot eavesdrop on the sensor and actuator data. Hence the corresponding attack policy does not use any online data on the system and is further assumed to be computed *a priori*. Therefore it corresponds to an open-loop type of policy. The attacker also has access to the detailed model of the system $\Sigma = (A, B, C)$, which is used to compute the appropriate attack policy as described in the following section.

III. Geometric control characterization of zero-dynamics

Recalling Lemma 1, the zero-dynamics attacks can be analyzed by considering the plant dynamics due to the false-data injection attack as described in (5).

The set of zero-dynamics attacks to (5) with $B_a = B$ $D_a = 0$ are now characterized under a geometric control framework [15].

Remark 1: The case for $D_a \neq 0$ can be analyzed in a similar fashion.

The following assumptions on $\Sigma = (A, B, C)$ are considered.

Assumption 1: The matrix B is full column-rank and C is full row-rank. Moreover Σ is the minimal realization of the system.

We now introduce the necessary concepts from geometric control theory [15] to describe the zero dynamics. In the following we denote $A \subseteq C$ as the set inclusion of A by C and $A \subseteq B + C$ as the set inclusion of A by the union of B and C. Furthermore, the range space of B is denoted as $\mathrm{Im}(B)$ and the null-space of C as $\ker(C)$.

Controlled Invariants

The first concept is that of controlled invariant subspace.

Lemma 2: For a given non-empty subspace $\mathcal{V}$ for which $A\mathcal{V} \subseteq \mathcal{V} + \mathrm{Im}(B)$ holds, there exists a matrix F such that $(A + BF)\mathcal{V} \subseteq \mathcal{V}$. Furthermore, $\mathcal{V}$ is called an $(A, \mathrm{Im}(B))$-controlled invariant subspace.

The subset of controlled invariant subspaces contained in $\ker(C)$ is the basis for characterizing the system's zero-dynamics, as summarized in the next statement.

Lemma 3: There exists an initial condition $x_0 \neq 0$ and control input a_k such that $y_k = 0 \quad \forall k \geq 0$ if and only if there exists a non-empty $(A, \mathrm{Im}(B))$-controlled invariant subspace $\mathcal{V}$ such that $\mathcal{V} \subseteq \ker(C)$.

The set of all subspaces $\mathcal{V}$ satisfying the conditions of Lemma 3 admits a maximum, $\mathcal{V}^\star$, which we denote by the maximal output-nulling invariant subspace. A procedure to compute $\mathcal{V}^\star$ can be found in [15]. Furthermore we denote the eigenvalues of $A + BF$ restricted to the eigenspace spanned by $\mathcal{V}^\star$ as the zeros of the system Σ. Denoting $\lambda \in \mathbf{C}$ as one such eigenvalue, the zero is said to be unstable if $|\lambda| > 1$ and stable otherwise.

Output-nulling subspace

The output-nulling inputs of the system (5) can be characterized as the output of an autonomous dynamical system as stated in the following theorem.

Theorem 1: The input $a_k = Fz_k$ with $z_{k+1} = (A + BF)z_k$, $(A + BF)\mathcal{V}^\star \subseteq \mathcal{V}^\star \subseteq \ker(C)$ and $z_0 \in \mathcal{V}^\star$ yields $y_k = 0 \quad \forall k \geq 0$ for the initial condition $x_0 = z_0$.

In general the above theorem characterizes only a subset of the possible output-nulling inputs, as some inputs may be described by a forced dynamical system. The reader is referred to [14] for more details.

Remark 2: Note that the former definition of zero-dynamics requires the initial condition to be non-zero and belong to $\mathcal{V}^\star$. Such requirement contradicts the definition of 0-stealthy attacks where the initial condition of the system component under attack is the origin. The effect of having non-zero initial conditions is addressed in the next section. The zero-dynamics attack policy readily follows from Theorem 1.

Corollary 1: The zero-dynamics attack policy is characterized by

$$\begin{aligned} z_{k+1} &= (A + BF)z_k \\ a_k &= Fz_k, \end{aligned} \tag{6}$$

with $z_0 \in \mathcal{V}^\star$ and F such that $(A + BF)\mathcal{V}^\star \subseteq \mathcal{V}^\star$.

IV. Effects of non-zero initial condition

Note that the zero-dynamics do not match the definition of 0-stealthy attacks, since a non-zero initial condition in (5) is required. However, in some cases the effects of the initial condition may be made arbitrarily small as discussed below.

Using Corollary 1, the system under a zero-dynamics attack is described by

$$\begin{aligned} \begin{bmatrix} x_{k+1} \\ z_{k+1} \end{bmatrix} &= \begin{bmatrix} A & BF \\ 0 & A + BF \end{bmatrix} \begin{bmatrix} x_k \\ z_k \end{bmatrix} \\ y_k &= \begin{bmatrix} C & 0 \end{bmatrix} \begin{bmatrix} x_k \\ z_k \end{bmatrix} \end{aligned} \tag{7}$$

with $z_0 \in \mathcal{V}^\star$. For $x_0 = z_0$ it directly follows that $y_k = 0 \quad \forall k \geq 0$. Introducing the error variable $e_k = x_k - z_k$, the previous system may be rewritten as

$$\begin{aligned} \begin{bmatrix} e_{k+1} \\ z_{k+1} \end{bmatrix} &= \begin{bmatrix} A & 0 \\ 0 & A + BF \end{bmatrix} \begin{bmatrix} e_k \\ z_k \end{bmatrix} \\ y_k &= \begin{bmatrix} C & 0 \end{bmatrix} \begin{bmatrix} e_k \\ z_k \end{bmatrix} \end{aligned} \tag{8}$$

with $z_0 \in \mathcal{V}^\star$ and $e_0 = x_0 - z_0$. The next result readily follows.

Theorem 2: For a zero initial condition $x_0 = 0$, a zero-dynamics attack generated by $z_0 \in \mathcal{V}^\star$ yields the output characterized by

$$\begin{aligned} e_{k+1} &= Ae_k \\ y_k &= Ce_k \end{aligned},$$

with $e_0 = -z_0$.

The previous result allows us to characterize conditions on which the energy of the output of zero-dynamics attacks can be made arbitrarily small.

Corollary 2: The output of a zero-dynamics attack generated by $z_0 \in \mathcal{V}^\star$ with $x_0 = 0$ has finite energy if and only if z_0 is orthogonal to the eigenvectors of A associated with unstable eigenvalues.

Proof: Recall that the system is assumed to be observable and thus there are no unobservable modes. Thus any initial condition exciting an unstable mode affects the output. Furthermore initial conditions only exciting stables modes induce state trajectories decaying asymptotically to zero, thus having finite output energy. ■

Now we analyze the case where z_0 is orthogonal to the unstable eigenvectors of A. Consider the coordinate transform $e_k = Tv_k$ where $T = [T_s\ T_u]$ is a basis for the eigenspace of A and T_s is associated with the stable eigenvalues. The dynamics are thus described by $v_{k+1} = \Lambda v_k$ where Λ is the Jordan block matrix of A containing its eigenvalues. Given the structure of T, Λ can be written as

$$\Lambda = \begin{bmatrix} \Lambda_s & 0 \\ 0 & \Lambda_u \end{bmatrix}$$

where Λ_s contains all the stable eigenvalues. Supposing that z_0 only excites stable eigenvalues of A, the output may be characterized as

$$\begin{aligned} v_{s_{k+1}} &= \Lambda_s v_{s_k} \\ y_k &= CT_s v_{s_k} \end{aligned},$$

where $v_k = [v_{s_k}^\top\ v_{u_k}^\top]^\top$ with $v_{s_0} = [I_s\ 0_u]T^{-1}z_0$ and $v_{u_0} = [0_s\ I_u]T^{-1}z_0 = 0$. This leads to the following result.

Corollary 3: Consider a zero-dynamics attack generated by $z_0 \in \mathcal{V}^\star$ with z_0 orthogonal to the unstable eigenvectors of A and $x_0 = 0$. The output energy of such attack is given by $\|y\|_{\ell_2}^2 = z_0^\top \bar{Q} z_0$ where

$$\bar{Q} = T^{-\top} \begin{bmatrix} I_s \\ 0_u \end{bmatrix} Q_s \begin{bmatrix} I_s & 0_u \end{bmatrix} T^{-1}$$

and Q_s is the the solution to

$$\Lambda_s^\top Q_s \Lambda_s - Q_s - T_s^\top C^\top C T_s = 0$$

.

Proof: The proof is omitted. ∎

The output energy of zero-dynamic attacks can thus be made arbitrarily small by selecting a sufficiently small initial condition $z_0 \in \mathcal{V}^\star/\mathcal{T}_u$ to generate the attack, where $\mathcal{T}_u = \mathrm{Im}(T_u)$ and $\mathcal{V}^\star/\mathcal{T}_u$ denotes the quotient space of $\mathcal{V}^\star$ with respect to $\mathcal{T}_u$. Such attacks are particularly dangerous if the initial condition z_0 excites an unstable eigenvalue of $A + BF$, as illustrated in the numerical example in Section VI. This motivates us to broaden the scope and address all zero-dynamics attacks characterized by Theorem 1.

V. Revealing Zero-Dynamics Attacks

In this section we discuss possible methods to reveal the zero-dynamics attacks characterized in Section III. The following definition of revealed attacks is considered throughout this work.

Definition 2: Consider the system under attack as described in (7). The zero-dynamics attack signal $\mathcal{A}_{k_0}^{k_f}$ is said to be revealed if $y_k \neq 0$ for some $k \geq k_0$.

Remark 3: The former definition can be extended to require the output energy to be sufficiently large. Furthermore, it can also account for the output feedback controller and anomaly detector by considering the closed-loop dynamics in (7).

Given Definition 2, the attack can be revealed if the zero-dynamics of the system are changed. As it is well-known in the control literature [16], this cannot be achieved by state- or output-feedback policies. Instead, a possible method is to modify the system $\Sigma = (A, B, C)$ in a certain way to $\tilde{\Sigma} = (\tilde{A}, \tilde{B}, \tilde{C})$ so that the attack signal (6) is no longer an output-nulling input of the resulting system

$$\begin{aligned} \begin{bmatrix} x_{k+1} \\ z_{k+1} \end{bmatrix} &= \begin{bmatrix} \tilde{A} & \tilde{B}F \\ 0 & A + BF \end{bmatrix} \begin{bmatrix} x_k \\ z_k \end{bmatrix} \\ y_k &= \begin{bmatrix} \tilde{C} & 0 \end{bmatrix} \begin{bmatrix} x_k \\ z_k \end{bmatrix}. \end{aligned} \tag{9}$$

Since (9) is an autonomous system, the following result readily follows.

Lemma 4: Every zero-dynamics attack is revealed if and only if the system (9) is observable for all $x_0 = z_0 \in \mathcal{V}^\star$.

Proof: By definition of observability, a given subspace $\mathcal{M}$ is observable if and only if $Y = W_o w_0 \neq 0, \quad \forall w_0 \in \mathcal{M}$ where $Y = [y_0^\top \cdots y_n^\top]^\top$ and $W_o \in \mathbb{R}^{np \times n}$ is the observability matrix of the augmented system (9). Given Definition 2, $\mathcal{V}^\star$ being an observable subspace then implies that the attacks are revealed, since $Y \neq 0$. ∎

Attacks remaining stealthy after the perturbation can also be characterized using similar arguments.

Corollary 4: Consider a zero-dynamics attack generated by $x_0 \in \mathcal{V}^\star$. The former attack remains stealthy after the perturbation if and only if $w_0 = [x_0^\top\ x_0^\top]^\top$ belongs to the unobservable subspace of the system (9).

Proof: Suppose x_0 is an eigenvector of $A+BF$, without loss of generality, and consider the augmented system before the perturbation as in (7). Since the state trajectories of (7) generated by the attack are contained in span(w_0), the state when the perturbation occurs can be written as $\tilde{w}_0 = \alpha w_0$, for a given $\alpha \in \mathbb{R}$. The remaining of the proof follows from Definition 2. ∎

A less restrictive condition for revealing the set of zero-dynamics attacks associated with unstable zeros follows from the above theorem.

Corollary 5: Every unstable zero-dynamics attack is revealed if and only if the system (9) is detectable for all $x_0 = z_0 \in \mathcal{V}^\star$.

A procedure to verify the observability of (9) restricted to $x_0 = z_0 \in \mathcal{V}^\star$ is to use the corresponding observability matrix W_o and compute

$$\mathcal{X}_d = \ker(W_o)^\perp \cap \begin{pmatrix} I \\ I \end{pmatrix} \mathcal{V}^\star.$$

It follows that $[x_0^\top\ x_0^\top]^\top \in \mathcal{X}_d$ belongs to the observable subspace and hence x_0 can be estimated and the corresponding attack signal affects the output.

Next we propose schemes to reveal the zero-dynamics attacks by separately changing A, B, or C.

A. Modifying the output matrix C

Here we consider modifications on the output matrix C to reveal zero-dynamics attacks. In particular, we consider that a new output matrix $\tilde{C}$ is obtained by adding and removing measurements. The following result directly follows from Theorem 1.

Lemma 5: All the zero-dynamics attacks associated with a given $z_0 \in \mathcal{V}^\star$ remain stealthy with respect to $\tilde{\Sigma} = (A, B, \tilde{C})$ if and only if $\mathcal{V}^\star \subseteq \ker \tilde{C}$.

The former statement shows that only removing measurements does not reveal any attack. Moreover, attacks are revealed by adding measurements if only if $\mathcal{V}^\star \cap \ker \tilde{C}$ is empty or a strict subset of $\mathcal{V}^\star$.

Theorem 3: There exists a $z_0 \in \mathcal{V}^\star$ generating an stealthy attack to $\tilde{\Sigma} = (A, B, \tilde{C})$ if and only if there exists a non-empty $(A + BF)$-invariant subspace $\mathcal{X}$ that is contained in $\mathcal{V}^\star \cap \ker \tilde{C}$.

Proof: First we have that all attack are revealed if $\mathcal{V}^\star \cap \ker \tilde{C} = \emptyset$. Now suppose that $\mathcal{X} \subseteq \mathcal{V}^\star \cap \ker \tilde{C} \neq \emptyset$ and let $z_0 \in \mathcal{X}$. Observing that $\mathcal{X} \subseteq \ker \tilde{C}$, from Theorem 1 we have that the attack generated by z_0 remains stealthy if and only if $\mathcal{X}$ is $(A + BF)$-invariant. ■

The previous results indicate that one should add measurements such that the dimension of $\mathcal{X} = \mathcal{V}^\star \cap \ker(\tilde{C})$ is reduced as much as possible. In particular, $\mathcal{X} \subset \mathcal{V}^\star$ indicates that a set of the zero-dynamics attacks has been revealed, while $\mathcal{X} = \emptyset$ implies that none of the zero-dynamics attacks remains stealthy.

Based on these arguments, Algorithm 1 can be used to incrementally deploy measurements that reveal zero-dynamics attacks

Algorithm 1 Algorithm to deploy additional measurements revealing zero-dynamics attacks.

Initialize $\mathcal{M} \leftarrow \{C_i\}$ as the set of additional measurements available;
$j \leftarrow 0$;
$\mathcal{X}_0 \leftarrow \mathcal{V}^\star$;
repeat
  for all $C_i \in \mathcal{M}$ **do**
    $\mathcal{Y}_i \leftarrow \mathcal{X}_j \cap \ker C_i$;
  end for
  Choose $C_i \in \mathcal{M}$ such that $\dim(\mathcal{Y}_i)$ is minimized;
  Compute $\mathcal{X}_{j+1}$ as the maximal $(A + BF)$-invariant contained in $\mathcal{Y}_i$;
  $j \leftarrow j + 1$;
until $\mathcal{X}_j = \emptyset$ **or** $\mathcal{X}_{j-1} = \mathcal{X}_j$

Note that the proposed algorithm requires the addition of at most $N = \dim(\mathcal{V}^\star)$ new measurements. Furthermore, all the zero-dynamics attacks become revealed if and only if the output-nulling subspace is empty, i.e. $\mathcal{X}_j = \emptyset$.

B. Modifying the system matrix A

Perturbations to the system dynamics as $\tilde{A} = A + \Delta A$ are now considered, resulting in the system $\tilde{\Sigma} = (\tilde{A}, B, C)$. The following result provides the conditions under which an attack remains stealthy.

Theorem 4: All the zero-dynamics attacks associated with a given $z_0 \in \mathcal{V}^\star$ remain stealthy with respect to $\tilde{\Sigma} = (\tilde{A}, B, C)$ if and only if $\mathcal{V}^\star \subseteq \ker \Delta A$.

Proof: Let $z_0 \in \mathcal{V}^\star$ and recall that $w_0 = [z_0^\top \; z_0^\top]^\top$ belongs to the unobservable subspace of the augmented system (7). From Corollary 4, the attack remains stealthy if and only if w_0 is also in the unobservable subspace of the perturbed system (9). Using the PBH observability test [13], this means that there exists a complex number λ such that

$$\begin{bmatrix} \lambda I - \tilde{A} & -BF \\ 0 & \lambda I - (A + BF) \\ C & 0 \end{bmatrix} \begin{bmatrix} z_0 \\ z_0 \end{bmatrix} = 0.$$

Thus the attack is stealthy if and only if $\Delta A z_0 = 0$, which concludes the proof. ■

The above result indicates that ΔA should be designed so that $\mathcal{V} \not\subseteq \ker \Delta A$ for all $(A + BF)$-invariant subspaces $\mathcal{V} \subseteq \mathcal{V}^\star$, thus revealing all the zero-dynamics attacks. Below we provide a necessary and sufficient condition for all the attacks to be revealed.

Corollary 6: All the zero-dynamics attacks are revealed if and only if $\mathcal{V}^\star \cap \ker \Delta A = \emptyset$.

C. Modifying the input matrix B

Here we consider modifications on the input matrix B to reveal zero-dynamics attacks. A new input matrix $\tilde{B}$ is obtained by adding and removing actuators or perturbing the B with ΔB. The following result directly follows from Theorem 1.

Lemma 6: Suppose inputs are added to Σ, i.e. $\tilde{B} = [B \, B_i]$. Then all the zero-dynamics attacks on Σ remain stealthy with respect to $\tilde{\Sigma} = (A, \tilde{B}, C)$.

Proof: The proof is omitted. ■

The former statement shows that only adding inputs does not reveal any attack. On the other hand, although removing actuators might reveal the zero-dynamics attacks, it also reduces the controllability of the system. A less intrusive approach is to change the actuator gains i.e., have $\tilde{B} = BW$ and $\tilde{u}_k = W^{-1} u_k$ where W is a diagonal matrix unknown to the attacker. This can be interpreted as a coding or encryption scheme performed by the actuator and controller with W as their shared private key. Assuming W is unknown by the attacker, we then have the following result.

Theorem 5: All the zero-dynamics attacks on Σ remain stealthy with respect to $\tilde{\Sigma} = (A, BW, C)$ if and only if $B(W - I)F\mathcal{V}^\star = \emptyset$.

Proof: Let $z_0 \in \mathcal{V}^\star$ and recall that $w_0 = [z_0^\top \; z_0^\top]^\top$ is in the unobservable subspace of the perturbed system (9) if and only if there exists a complex number λ such that

$$\begin{bmatrix} \lambda I - A & -BWF \\ 0 & \lambda I - (A + BF) \\ C & 0 \end{bmatrix} \begin{bmatrix} z_0 \\ z_0 \end{bmatrix} = 0.$$

Thus the attack is stealthy if and only if $B(W - I)F z_0 = 0$, which concludes the proof. ■

A necessary and sufficient condition for zero-dynamics attacks to be revealed with such perturbations follows directly from the previous theorem.

Corollary 7: All the zero-dynamics attacks are revealed if and only if $\mathcal{V}^\star \cap \ker(B(W - I)F) = \emptyset$.

The former result and the assumption that the system is observable can be used to provide a method for choosing W.

Lemma 7: Assume that (A, C) is observable. For any matrix F such that $\mathcal{V}^\star$ is $(A+BF)$-invariant, it holds that $\mathcal{V}^\star \cap \ker(BF) = \emptyset$.

Proof: Recall the $\mathcal{V}^\star$ is $(A+BF)$-invariant and suppose that $\mathcal{V}^\star \cap \ker(BF) \neq \emptyset$ i.e., there exists $z_0 \in \mathcal{V}^\star$ such that $BFz_0 = 0$. This then implies that z_0 is A-invariant and generates an unobservable state trajectory, which is a contradiction since the system is observable. ■

Since $\ker(BF)$ is not affected by a uniform scaling, a possible weight for revealing zero-dynamics attacks is $W = \alpha I$ with $\alpha \in \mathbb{R}_+$ and $\alpha \neq 1$, resulting in $B(W-I)F = (\alpha-1)BF$. We now analyze the effects of such perturbation on the output energy of the system. Introducing the variable $\tilde{x}_k = \alpha^{-1}x_k$, the perturbed system (9) can be rewritten as

$$\begin{aligned} \begin{bmatrix} \tilde{x}_{k+1} \\ z_{k+1} \end{bmatrix} &= \begin{bmatrix} A & BF \\ 0 & A+BF \end{bmatrix} \begin{bmatrix} \tilde{x}_k \\ z_k \end{bmatrix} \\ y_k &= \begin{bmatrix} \alpha C & 0 \end{bmatrix} \begin{bmatrix} \tilde{x}_k \\ z_k \end{bmatrix}, \end{aligned} \tag{10}$$

with $\tilde{x}_0 = \alpha^{-1}z_0$ and $z_0 \in \mathcal{V}^\star$. The output of such system is characterized as follows.

Theorem 6: Suppose the augmented system under a zero-dynamics attack (9) is at the state $z_k = x_k = z$ when the perturbation $W = \alpha I$ is performed. After the perturbation the output is described by

$$\begin{aligned} e_{k+1} &= Ae_k \\ y_k &= \alpha Ce_k \end{aligned},$$

with $e_0 = (\alpha^{-1} - 1)z$.

Proof: The proof comes from introducing the variable $e_k = \tilde{x}_k - z_k$ and rewriting (10) with respect to e_k and z_k. ■

Note that the output energy after the perturbation is dependent on z and the scaling α, as summarized in the following statements.

Corollary 8: The perturbation $W = \alpha I$ results in a finite-energy output if and only if z is orthogonal to the eigenvectors of A associated with unstable eigenvalues.

Consider the eigenvalue decomposition

$$A = T\Lambda T^{-1} = \begin{bmatrix} T_s & T_u \end{bmatrix} \begin{bmatrix} \Lambda_s & 0 \\ 0 & \Lambda_u \end{bmatrix} \begin{bmatrix} T_s & T_u \end{bmatrix}^{-1},$$

where Λ_s contains all the stable eigenvalues of A and T_s is a basis of the corresponding eigenspace.

Corollary 9: Consider the output described in Theorem 6 with z orthogonal to the unstable eigenvectors of A. The energy of the output is given by $\|y\|_{\ell_2}^2 = z^\top \bar{Q} z$ where

$$\bar{Q} = T^{-\top} \begin{bmatrix} I_s \\ 0_u \end{bmatrix} Q_s \begin{bmatrix} I_s & 0_u \end{bmatrix} T^{-1}$$

and Q_s is the the solution to

$$\Lambda_s^\top Q_s \Lambda_s - Q_s - \alpha^2 T_s^\top C^\top C T_s = 0$$

.

VI. Illustrative Example

To better illustrate the results from the previous sections, here we provide an example of a zero-dynamics attack on a process control system. Our example consists of the Quadruple-Tank Process (QTP) [17]. The continuous-time nonlinear plant model is given by

$$\begin{aligned} \dot{h}_1(t) &= -\frac{a_1}{A_1}\sqrt{2gh_1(t)} + \frac{a_3}{A_1}\sqrt{2gh_3(t)} + \frac{\gamma_1 k_1}{A_1}u_1(t) \\ \dot{h}_2(t) &= -\frac{a_2}{A_2}\sqrt{2gh_2(t)} + \frac{a_4}{A_2}\sqrt{2gh_4(t)} + \frac{\gamma_2 k_2}{A_2}u_2(t) \\ \dot{h}_3(t) &= -\frac{a_3}{A_3}\sqrt{2gh_3(t)} + \frac{(1-\gamma_2)k_2}{A_3}u_2(t) \\ \dot{h}_4(t) &= -\frac{a_4}{A_4}\sqrt{2gh_4(t)} + \frac{(1-\gamma_1)k_1}{A_4}u_1(t) \end{aligned} \tag{11}$$

where h_i are the heights of water in each tank, A_i the cross-section area of the tanks, a_i the cross-section area of the outlet hole, k_i the pump constants, γ_i the flow ratios and g the gravity acceleration. The outputs are defined as the water levels of tanks 1 and 2, h_1 and h_2 respectively. The system has an adjustable zero with respect to u, which is unstable if $0 < \gamma_1 + \gamma_2 < 1$. In the simulation we consider the linearized model at a given operating point, which is sampled with a period of $T_s = 0.5s$. The resulting discrete-time system is given by (1) with

$$\begin{aligned} A &= \begin{bmatrix} 0.975 & 0 & 0.042 & 0 \\ 0 & 0.977 & 0 & 0.044 \\ 0 & 0 & 0.958 & 0 \\ 0 & 0 & 0 & 0.956 \end{bmatrix}, \\ B &= \begin{bmatrix} 0.0515 & 0.0016 \\ 0.0019 & 0.0447 \\ 0 & 0.0737 \\ 0.0850 & 0 \end{bmatrix}, \\ C &= \begin{bmatrix} 0.2 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 \end{bmatrix}. \end{aligned}$$

The corresponding maximal $(A, \operatorname{Im} B)$-controlled invariant subspace contained in $\ker(C)$, $\mathcal{V}^\star$, is spanned by $V^\star$ which is shown below together with a suitable F

$$V^\star = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad F = \begin{bmatrix} 0 & 0 & -0.8057 & 0.0302 \\ 0 & 0 & 0.0349 & -0.9844 \end{bmatrix}.$$

The system $\Sigma = (A, B, C)$ has two zeros, $\lambda = 0.89$ and $\lambda = 1.03$, and A has only stable eigenvalues. The unstable zero-dynamics corresponding to $\lambda = 1.03$ are excited by $z_0 = \epsilon[0 \quad 0 \quad -0.72 \quad 0.69]^\top$ with $\epsilon \neq 0$. The respective input signal is depicted in Figure 1. This attack is considered in the examples below.

A. Modifying the output matrix C

Consider that the possible measurements can be used to reveal zero-dynamics attacks

$$\begin{aligned} C_3 &= \begin{bmatrix} 0 & 0 & 0.2 & 0 \end{bmatrix} \\ C_4 &= \begin{bmatrix} 0 & 0 & 0 & 0.2 \end{bmatrix}. \end{aligned}$$

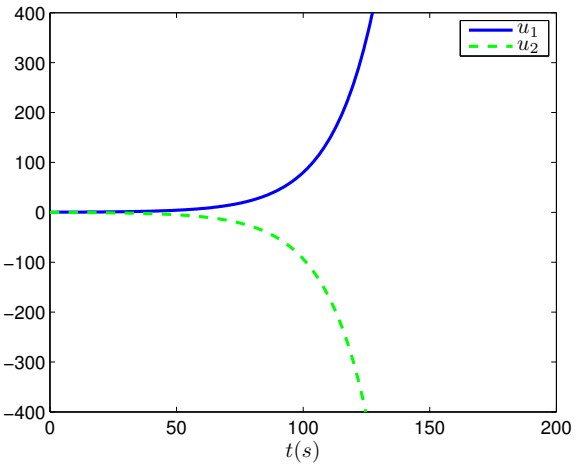

Fig. 1. Unstable zero-dynamics attack applied to the system from $t = 0s$.

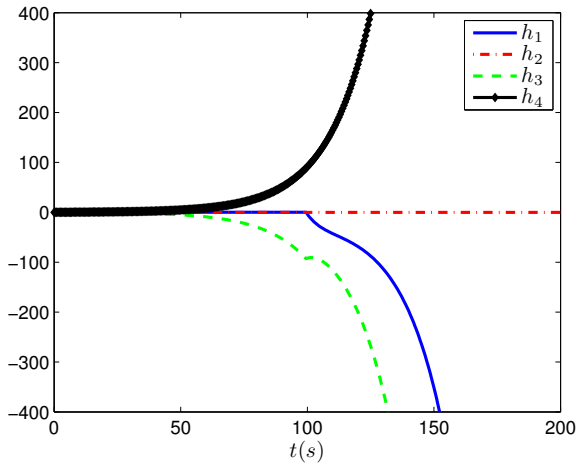

Fig. 2. State trajectories of the system under attack and active attack detection.

Applying the algorithm proposed in Section V-A we see that adding C_3 yields $\mathcal{Y} = \mathcal{V}^\star \cap \ker C_3 = \text{span}([0\,0\,0\,1]^\top)$, which is not $(A + BF)$-invariant subspace and thus all the zero-dynamics attacks to Σ are revealed. In fact $\tilde{\Sigma} = (A, B, \tilde{C})$ with $\tilde{C} = [C^\top\, C_3^\top]^\top$ has no zeros. In this particular example, adding C_4 instead of C_3 would also reveal all the zero-dynamics attacks.

B. Modifying the system matrix A

From Theorem 4 we have that any system perturbation of the type

$$\Delta A = \begin{bmatrix}\Delta & 0\end{bmatrix}$$

with $\Delta \in \mathbb{R}^{4\times 2}$ leaves all the zero-dynamics attacks stealthy. In fact, note that $(A + \Delta A + BF)\mathcal{V}^\star \equiv (A + BF)\mathcal{V}^\star$ and therefore the zero-dynamics of Σ and $\tilde{\Sigma}$ are identical. Therefore such perturbations should be avoided.

On the other hand, the zero-dynamics change for perturbations of the type

$$\Delta A = \begin{bmatrix}0 & \Delta\end{bmatrix}.$$

For instance, adding an extra connection from tank 3 to tank 1 corresponds to

$$\Delta A = \begin{bmatrix}0 & 0 & 0.0397 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & -0.0402 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}.$$

The outcome of such perturbation can be seen in Figure 2 and Figure 3. The attack begins at $t = 0s$ with a initial conditions mismatch, leading to a small increase in the output energy as initially seen in Figure 3. The change to the system dynamics occurs at $t = 100s$ and one immediately observes a perturbation in the state trajectory. The extra coupling between tanks 3 and 1 changes the zero-dynamics of the system and thus the current attack signal affects the water level of tank 1. As a result the attack is revealed in the output, as illustrated in Figure 3.

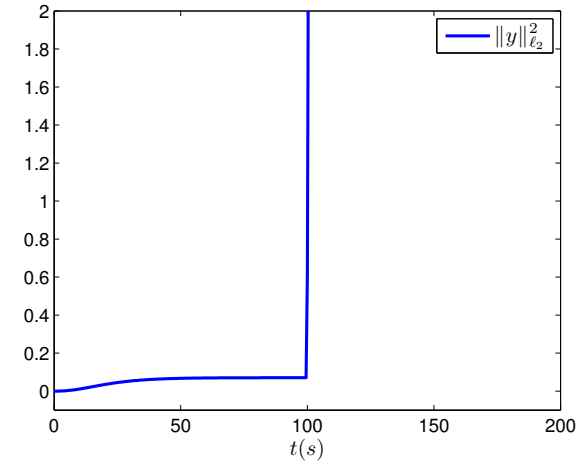

Fig. 3. Output energy of the system after connecting tank 3 to tank 1 at $t = 100s$.

C. Modifying the input matrix B

Consider the case where the uniform input scaling $W = 0.987I$ is applied to the system. From the results in Section V-C, all the zero-dynamics are revealed, since $\ker(BF) = \ker((1-\alpha)BF)$ and $\mathcal{V}^\star \cap \ker(BF) = \emptyset$. Moreover, as stated in Corollary 8 the scaling results in a finite energy output since A is stable. The output energy resulting from the attack an input scaling is depicted in Figure 4. As before, the attack begins at $t = 0s$ with a mismatch in the initial condition, resulting in a finite output energy. The input scaling is applied at $t = 100s$, which again results in a finite increment of the output energy since A is stable, as depicted in Figure 4.

VII. Conclusions and Future Work

The problem of revealing zero-dynamics attacks on control system was tackled. First we studied the effect of initial condition mismatch in terms of the resulting increase in the output energy. We concluded that for the subset of attacks

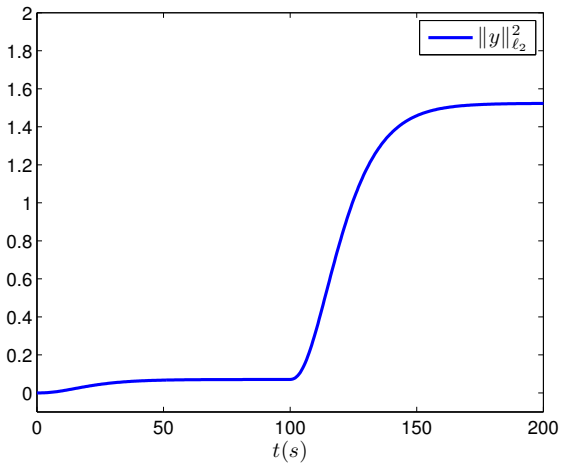

Fig. 4. Output energy of the system after introducing the input scaling $BW = 0.987B$ at $t = 100s$.

exciting unstable zero-dynamics, this effect can be made arbitrarily small while still affecting the system performance. Then we addressed the problem of revealing zero-dynamics attacks by modifying the system structure in terms of the respective outputs, inputs, and dynamics. For changes in each component, we provided necessary and sufficient conditions for all attacks to be revealed. Furthermore, we provided an algorithm to incrementally add measurements and thus reveal attacks. We also proposed a coordinated scaling of the inputs by the actuator and controller. For this particular change, we quantified the resulting increase in output energy in terms of the initial condition and scaling factor. Both these changes on the inputs and outputs are able to reveal attacks while not affecting the system performance when no attack is present.

REFERENCES

[1] W. Shefte, S. Al-Jamea, and R. O'Harrow. (2012, June) Cyber search engine shodan exposes industrial control systems to new risks. http://www.washingtonpost.com/investigations/cyber-search-engine-exposes-vulnerabilities/2012/06/03/gJQAIK9KCV_story.html. The Washington Post.

[2] J. Meserve, "Sources: Staged cyber attack reveals vulnerability in power grid," *CNN*, 2007, available at http://edition.cnn.com/2007/US/09/26/power.at.risk/index.html.

[3] Symantec, "Stuxnet introduces the first known rootkit for industrial control systems," *Symantec*, August 6th 2010, available at: http://www.symantec.com/connect/blogs/stuxnet-introduces-first-known-rootkit-scada-devices.

[4] T. Rid, "Cyber war will not take place," *Journal of Strategic Studies*, 2011.

[5] U.S.-Canada PSOTF, "Final report on the August 14th blackout in the United States and Canada," U.S.-Canada Power System Outage Task Force, Tech. Rep., April 2004.

[6] P. Esfahani, M. Vrakopoulou, K. Margellos, J. Lygeros, and G. Andersson, "Cyber attack in a two-area power system: Impact identification using reachability," in *American Control Conference, 2010*, jul 2010, pp. 962–967.

[7] A. Cárdenas, S. Amin, Z. Lin, Y. Huang, C. Huang, and S. Sastry, "Attacks against process control systems: risk assessment, detection, and response," in *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*, ser. ASIACCS '11. New York, NY, USA: ACM, 2011, pp. 355–366.

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

[9] R. Smith, "A decoupled feedback structure for covertly appropriating networked control systems," in *Proc. of the 18th IFAC World Congress*, Milano, Italy, August-September 2011.

[10] F. Pasqualetti, F. Dorfler, and F. Bullo, "Cyber-physical attacks in power networks: Models, fundamental limitations and monitor design," in *Proc. of the 50th IEEE Conf. on Decision and Control and European Control Conference*, Orlando, FL, USA, Dec. 2011.

[11] A. Teixeira, D. Pérez, H. Sandberg, and K. Johansson, "Attack models and scenarios for networked control systems," in *Proceedings of the 1st international conference on High Confidence Networked Systems*. ACM, 2012, pp. 55–64.

[12] Y. Mo and B. Sinopoli, "Secure control against replay attack," in *47th Annual Allerton Conference on Communication, Control, and Computing*, Oct. 2009.

[13] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1996.

[14] J. Tokarzewski, *Finite zeros in discrete time control systems*, ser. Lecture notes in control and information sciences. Springer, 2006.

[15] G. Basile and G. Marro, *Controlled and conditioned invariants in linear system theory*. Prentice Hall, 1992.

[16] S. Skogestad and I. Postlethwaite, *Multivariable Feedback Control: Analysis and Design*. John Wiley & Sons, 1996.

[17] K. Johansson, "The quadruple-tank process: a multivariable laboratory process with an adjustable zero," *IEEE Transactions on Control Systems Technology*, vol. 8, no. 3, pp. 456–465, May 2000.