

CYBER-PHYSICAL SYSTEMS: DYNAMIC SENSOR ATTACKS AND STRONG OBSERVABILITY

Yuan Chen, Soumya Kar, and José M. F. Moura

Carnegie Mellon University
Department of Electrical and Computer Engineering
Pittsburgh, PA 15213 USA

ABSTRACT

We study cyber-physical systems subject to dynamic sensor attacks, relating them to the system's strong observability. First, we find necessary and sufficient conditions for an attacker to create a dynamically undetectable sensor attack and relate these conditions to properties of the system dynamics eigenvectors. Next, we provide an index that gives the minimum number of sensors that must be attacked in order for an attack to be undetectable. Finally, we illustrate our results with a numerical example on the Quadropole Tank Process.

Index Terms— Cyber-Physical Systems, Security, Sensor Attacks

1. INTRODUCTION

The security of cyber-physical systems – systems that integrate sensing, control, actuation components via a communication network – has received increased attention due to notable incidents of cyber-physical attacks. Events such as the Maroochy Shire Council Sewage control incident [1] and the Stuxnet Malware [2] have demonstrated the security vulnerabilities of cyber-physical systems monitoring large critical infrastructure. More recently, smaller scale cyber-physical systems such as commercial automobiles have become targets of similar forms of attack [3].

In a sensor attack, an attacker exploits vulnerabilities in the communication scheme to send falsified sensor data to the controller. Controllers depend on sensor data to perform tasks such as state estimation and output feedback, and consequently, an attacker can manipulate the physical behavior of a system simply by modifying its sensor measurements. In order to limit damaging behavior, cyber-physical systems are equipped with attack detectors. Static attack detectors verify

the consistency of sensor measurements at a single time step to determine the presence of a sensor attack [4]. Dynamic attack detectors incorporate knowledge of system dynamics to perform attack detection over multiple time steps and detect certain attacks that are undetectable to static detectors [5]. Attack reconstruction algorithms identify the specific sensors that fall under attack [6], [7], but have more restricted limitations than attack detectors since there are certain attacks that can be detected but not reconstructed.

This paper focuses on the fundamental limitations of dynamic sensor attack detection. We use the strong observability property of a dynamical system [8], [9] to determine the existence of dynamically undetectable sensor attacks against a particular system. The strong observability property is a general framework for analyzing limitations of dynamic attack detection that extends to systems under both actuator and sensor attacks. In this paper, we derive results for the case of systems under sensor attacks. We give a necessary and sufficient condition for the attacker to be undetectable in terms of the system dynamics eigenvectors. We provide an index that determines the minimum number of sensors that must be attacked in order for an attack to be undetectable and use this index to demonstrate a design guideline for improving the resilience of the system to sensor attacks. Finally, we illustrate our results with a numerical example.

The rest of this paper is organized as follows. Section 2 specifies the system and attack model, reviews attack detection, and formalizes the problem. Section 3 provides fundamental limitations of dynamic sensor attack detection and relates undetectable attacks to strong observability. We provide a numerical example in Section 4 and conclude in Section 5.

2. BACKGROUND

2.1. System and Attack Model

We use the following linear, time invariant, state space model for the cyber-physical system under sensor attack:

$$\begin{aligned}x(k+1) &= Ax(k), \\y(k+1) &= Cx(k) + Da(k),\end{aligned}\tag{1}$$

This material is based on research sponsored by DARPA under agreement number DARPA FA8750-12-2-0291. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

where $x \in \mathbb{R}^n$ is the system state, $y \in \mathbb{R}^p$ is the system output (sensor measurements), $k \in \mathbb{Z}$ is the time index and $a(k) \in \mathbb{R}^s$ is the sensor attack, which is unknown to the system. The system has an unknown initial state $x(0)$. The matrix $A \in \mathbb{R}^{n \times n}$, which represents the system dynamics, and the matrix $C \in \mathbb{R}^{p \times n}$, which represents the system sensing topology, are known to the system. The matrix $D \in \mathbb{R}^{p \times s}$ represents the capabilities of the attacker and is unknown to the system. Without loss of generality, we assume that D is a full rank matrix. Furthermore, we assume that the pair (A, C) of the system in (1) is observable. Equation (1) is a standard model for a cyber-physical system under sensor attack [6], [7], [10].

The attacker has full knowledge of the system dynamics, represented by the A matrix, and sensing topology, represented by the C matrix, and can choose the attack $a(k)$ arbitrarily at each time k . The attacker is restricted to attacking only a subset $K \subset \{1, 2, \dots, p\}$ of all sensors. The set K is known as the attack set. It has cardinality $|K| = s$, which is unknown to the system. This attacker model follows the sparse sensor attack model presented in [6] and [7]. For each attack set K , there is a corresponding D_K matrix to represent the attacker's capabilities in equation (1):

$$D_K = \begin{bmatrix} e_{K_1} & e_{K_2} & \cdots & e_{K_s} \end{bmatrix}, \quad (2)$$

where $K = \{K_1, K_2, \dots, K_s\}$ and e_j , $j = 1, \dots, p$, is the j^{th} canonical vector of \mathbb{R}^p . We assume that the attacker knows D_K . As notational shorthand, let $\Sigma_K = (A, C, D_K)$ represent the system in equation (1) with attack set K .

2.2. Attack Detection

Attack detection algorithms use knowledge of the system sensing mechanism and system output to determine whether or not a sensor attack has occurred. Static attack detectors monitor the consistency of the system output with the sensing mechanism at a single time step [5]. An example of a static attack detector is the residual-based bad measurement detector used in power system state estimation [4]. As the authors of [4] and [5] show, any attack that satisfies $Da(k) \in \mathcal{R}(C)$, where $\mathcal{R}(C)$ is the range space of C , is undetectable to a static detector.

A dynamic attack detector uses knowledge of the system dynamics A to perform attack detection over multiple time steps. We assume that dynamic attack detectors know the A and C matrices and the system output $y(k)$ exactly over all time steps. That is, a dynamic attack detector uses the output trajectory $Y(T) = \begin{bmatrix} y(0)^T & y(1)^T & \cdots & y(T)^T \end{bmatrix}^T$ to determine whether or not a nonzero attack $E(T) = \begin{bmatrix} a(0)^T & a(1)^T & \cdots & a(T)^T \end{bmatrix}^T$ has occurred over the time period $0, \dots, T$. Specific implementations of dynamic attack detectors are discussed in [10] and [11]. In [10], the authors determine the types of attacks that are undetectable in the presence of sensor and process noise, and, in [11], the

authors consider undetectable attacks against a residual based dynamic detector.

In this paper, we study sensor attacks that are undetectable or stealthy to *any* dynamic attack detector and do not assume that the detector has any particular implementation. According to [5], a dynamically undetectable attack is one that causes the system to have an output trajectory that corresponds to the output trajectory of the system not under attack. Given the system $\Sigma_K = (A, C, D_K)$, the output trajectory $Y(T)$ is exactly determined by the unknown initial state $x(0)$ and the unknown attack sequence $E(T)$. Specifically, we have

$$Y(T) = \mathcal{O}_T x(0) + (I_{T+1} \otimes D_K) E(T), \quad (3)$$

where I_{T+1} is the $(T+1) \times (T+1)$ identity matrix, \otimes is the Kronecker product, and \mathcal{O}_T is the extended system observability matrix,

$$\mathcal{O}_T = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^T \end{bmatrix}. \quad (4)$$

Reference [5] states that an attack $E(T)$ is dynamically undetectable if and only if it satisfies

$$\mathcal{O}_T x(0) + (I_{T+1} \otimes D_K) E(T) = \mathcal{O}_T x'(0), \quad (5)$$

where $x(0)$ is the system's initial state and $x'(0)$ is an arbitrary state. The stealth of an attack $E(T)$ is not affected by known inputs to the system (e.g., if the system in equation (1) has actuator inputs $u(k)$ such that $x(k+1) = Ax(k) + Bu(k)$) since the contribution of the input to the trajectory can be calculated exactly. For this reason, without loss of generality, we consider the system in equation (1) with no actuator inputs.

In [6] and [7], the authors present dynamic attack *reconstruction* algorithms that use the system trajectory $Y(T)$ to identify the attack set K and compute the attack sequence $E(T)$. There are, however, certain attacks that are unrecoverable, i.e., attacks that cannot be reconstructed, but are dynamically detectable [5]. Reference [6] states that no unrecoverable attacks on s sensors exist against a system (A, C, D_K) with $|K| = s$ if and only if $|\text{supp}(Cv)| > 2s$ for all eigenvectors v of A , where $\text{supp}(x)$ is the set of indices of the nonzero components of x . The authors of [6] and [7] assume that the attacker is restricted to attacking at most s sensors of the system (A, C, D_K) where $|\text{supp}(Cv)| > 2s$ for all eigenvectors v of A . We provide a result that states that no *undetectable* attacks exist against (A, C, D_K) with $|K| = s$ if and only if $|\text{supp}(Cv)| > s$, which means that there are attack sequences that are not reconstructed by algorithms in [6] and [7] but are detected by dynamic attack detectors.

2.3. Problem Definition

This paper studies the existence of dynamically undetectable sparse sensor attacks over the time period $0, \dots, T$ against

cyber-physical systems. Given a system $\Sigma_K = (A, C, D_K)$ as in equation (1), we provide a necessary and sufficient condition on K and D_K for the existence of a dynamically undetectable attack $E(T)$. Furthermore, we find the minimum number of sensors that an attacker must attack in order to create a dynamically undetectable attack.

3. UNDETECTABLE SENSOR ATTACKS

We consider a system $\Sigma_K = (A, C, D_K)$ and provide necessary and sufficient conditions for the existence of undetectable sensor attack sequences. Specifically, we seek conditions for the existence of attack sequences $E(T)$ against Σ_K that are undetectable over the time period $0, \dots, T$ with $T = n-1$. In order to determine the existence of undetectable attack sequences over any time period $0, \dots, T$, it is sufficient to examine the time period $0, \dots, n-1$.

Lemma 1. *If there exists a sensor attack $E(n-1)$ against the system $\Sigma_K = (A, C, D_K)$ that is undetectable over the time period $0, \dots, n-1$, then there exists a sensor attack $\bar{E}(T)$ that is undetectable over the time period $0, \dots, T$ for any $T = 0, 1, \dots$.*

Proof. The proof is omitted. \square

3.1. Strong Observability

We derive the conditions for the existence of undetectable attacks $E(n-1)$ against a system Σ_K using the system's strong observability property. In this subsection, we review the definition of a system's weakly unobservable subspace and a system's strong observability presented in [8] and [12]. Let $\Sigma_K = (A, C, D_K)$. The input-unobservable subspace over k steps for Σ_K , \mathcal{L}_k , is the space of all $x \in \mathbb{R}^n$ such that, for the system Σ_K with initial state $x(0) = x$, there exists an input $E(k-1)$ so that the system output trajectory satisfies $Y(k-1) = 0$. That is, \mathcal{L}_k is the subspace of all $x \in \mathbb{R}^n$ for which there exists an attack sequence $E(k-1)$ that satisfies

$$\mathcal{O}_{k-1}x + (I_k \otimes D_K) E(k-1) = 0. \quad (6)$$

The input unobservable subspaces $\mathcal{L}_1, \mathcal{L}_2, \dots$ satisfy $\mathcal{L}_{k+1} \subseteq \mathcal{L}_k$ for all k and $\mathcal{L}_n = \mathcal{L}_{n+j}$ for all j [12]. The *weakly unobservable subspace* of a system Σ_K , denoted as $\mathcal{V}(\Sigma_K)$, is defined as its input unobservable subspace over n steps, \mathcal{L}_n [12]. We call a system Σ_K *strongly observable* if its weakly unobservable subspace is trivial, i.e., $\mathcal{V}(\Sigma_K) = 0$ [8]. References [8], [9], and [12] provide methods to calculate a system's weakly unobservable subspace.

3.2. Existence of Stealthy Sensor Attacks

The strong observability of a system Σ_K determines the existence of sensor attacks $E(n-1)$ that are undetectable over the time period $0, \dots, n-1$.

Theorem 1 (Existence of Undetectable Sensor Attacks). *There exists a sensor attack $E(n-1)$ against the system Σ_K that is undetectable over the time period $0, \dots, n-1$ if and only if Σ_K is not strongly observable.*

Proof. (If) Let Σ_K be a system that is not strongly observable. By definition of strong observability, $\mathcal{V}(\Sigma_K) \neq 0$. Let $\theta \in \mathcal{V}(\Sigma_K)$, $\theta \neq 0$. We decompose θ into the sum of the system initial state $x(0)$ and another state $-x'(0)$, i.e., $\theta = x(0) - x'(0)$. Since $\theta \in \mathcal{V}(\Sigma_K)$, there exists $E(n-1)$ that satisfies

$$\mathcal{O}_{n-1}\theta + (I_n \otimes D_K) E(n-1) = 0. \quad (7)$$

Substituting for $\theta = x(0) - x'(0)$ and rearranging equation (7), we have

$$\mathcal{O}_{n-1}x(0) + (I_n \otimes D_K) E(n-1) = \mathcal{O}_{n-1}x'(0), \quad (8)$$

which shows that there exists a sensor attack $E(n-1)$ against Σ_K that is undetectable over the time period $0, \dots, n-1$.

(Only If) Let $E(n-1)$ be a nonzero undetectable sensor attack against Σ_K . Let $x(0)$ be the initial state of Σ_K . Then, there exists $x'(0) \in \mathbb{R}^n$ such that equation (8) is satisfied. Let $\theta = x(0) - x'(0)$. Rearranging equation (8) and substituting for θ gives equation (7). Since there exists $E(n-1)$ such that θ and $E(n-1)$ satisfy (7), we have $\theta \in \mathcal{V}(\Sigma_K)$. What remains is to show that $\theta \neq 0$. Because D_K is injective, so is $I_n \otimes D_K$, and because $E(n-1) \neq 0$, we have $(I_n \otimes D_K) E(n-1) \neq 0$. In order to satisfy equation (7), we then have $\mathcal{O}_{n-1}\theta \neq 0$, which shows that $\theta \neq 0$. Thus, there exists a nonzero θ that belongs to the subspace, $\mathcal{V}(\Sigma_K)$, and by definition, the system Σ_K is not strongly observable. \square

Reference [8] gives an equivalent condition for system strong observability: a system $\Sigma_K = (A, C, D_K)$ is strongly observable if and only if the system $\Sigma_{K,F} = (A, C + D_K F, D_K)$ is observable for all $F \in \mathbb{C}^{s \times n}$. This property provides a necessary and sufficient condition on D_K such that there exists an undetectable attack against $\Sigma_K = (A, C, D_K)$.

Theorem 2 (Undetectable Attack Sets). *An undetectable attack $E(n-1)$ against Σ_K exists if and only if there exists an eigenvector v of A for which $Cv \in \mathcal{R}(D_K)$, where $\mathcal{R}(D_K)$ is the range space of D_K .*

The proof of Theorem 2 requires the Popov-Belevitch-Hautus (PBH) criterion for observability: the system $\Sigma_K = (A, C, D_K)$ is observable if and only if the matrix $\begin{bmatrix} A - \lambda I \\ C \end{bmatrix}$ has full rank (i.e. rank n) for all $\lambda \in \mathbb{C}$ [13].

Proof. (If) Let v_0 be an eigenvector of A with eigenvalue λ_0 such that $Cv_0 \in \mathcal{R}(D_K)$. Then there exists $\theta \in \mathbb{R}^s$ such that $D_K\theta = -Cv_0$, and there exists $F \in \mathbb{C}^{s \times n}$ such that $Fv_0 = \theta$. For such a choice of F , we have $\begin{bmatrix} A - \lambda_0 I \\ C + D_K F \end{bmatrix} v_0 = 0$,

which means that the system $\Sigma_{K,F} = (A, C + D_K F, D_K)$ does not satisfy the PBH criterion for observability. Since there exists $F \in \mathbb{C}^{s \times n}$ for which $\Sigma_{K,F}$ is not observable, Σ_K is not strongly observable, and by Theorem 1, there exists an undetectable sensor attack $E(n-1)$ against Σ_K .

(Only If) Let there be an undetectable attack $E(n-1)$ against Σ_K . Then, by Theorem 1, Σ_K is not strongly observable, and there exists $F \in \mathbb{C}^{s \times n}$ such that $\Sigma_{K,F}$ is not observable. Applying the PBH criterion for observability to $\Sigma_{K,F}$, we have that there exists $\lambda_0 \in \mathbb{C}$ and $v_0 \in \mathbb{C}^n \setminus \{0\}$ such that $\begin{bmatrix} A - \lambda_0 I \\ C + D_K F \end{bmatrix} v_0 = 0$. Thus we have $(A - \lambda_0 I) v_0 = 0$, which shows that v_0 is an eigenvector of A , and $C v_0 = D_K \theta$ for $\theta = -F v_0$, which shows that $C v_0 \in \mathcal{R}(D_K)$. \square

Theorem 2 is in the spirit of the result provided in [10]. The result presented in this paper differs from the result in [10] in that we derive Theorem 2 by explicitly connecting the existence of undetectable attacks to the strong observability property. One advantage of the strong observability framework is that it is a more general framework that can be extended to apply to systems in which both the sensors and actuators fall under attack.

Using Theorem 2, we provide an index s_0 that specifies the minimum number of sensors an attacker must attack in order to be undetectable:

$$s_0 = \min_{v \in \mathbb{C}^n \setminus \{0\}, Av = \lambda v} |\text{supp}(Cv)|. \quad (9)$$

Theorem 3 (Smallest Attack Set). *There exist an undetectable attack on s sensors if and only if $s \geq s_0$.*

Proof. The proof is omitted and provided elsewhere. \square

Calculating s_0 is combinatorial for a general A matrix and is infeasible for a large number of sensors p . If, however, the matrix A is simple, one calculates s_0 by computing $s_{0_i} = |\text{supp}(Cv_i)|$ for each eigenvector v_i of A (there are n eigenvectors) and finding the minimum amongst the s_{0_i} values (there are n s_{0_i} values). Theorem 3 states that attacking at least s_0 sensors is a necessary but insufficient condition for a sensor attack to be undetectable. An attack set K with cardinality $|K| \geq s_0$ and associated matrix D_K must still satisfy the condition in Theorem 2 to have an undetectable attack.

The index s_0 and Theorem 3 provide sensor design guidelines to improve the resilience of the system to sparse sensor attacks. Given the system dynamics matrix A , one examines $\text{supp}(Cv)$ for all eigenvectors v of A and constructs the sensing matrix C to ensure that all attacks whose sparsity falls below a certain threshold are detectable. Placing additional sensors into a system is equivalent to concatenating rows to the system sensing matrix C . The choice of sensor determines the nonzero components of the row, and by proper sensor placement, one increases s_0 and improves the system resilience to sensor attacks.

4. NUMERICAL EXAMPLE

To illustrate our results, we provide numerical example of a sensor attack against a cyber-physical system. The example uses the dynamics of the Quadruple-Tank Process from [11] and [14]. The system consists of four interconnected water tanks equipped with sensors to measure the height of the water in each tank. Following [11], the dynamics matrix of the system is as follows:

$$A = \begin{bmatrix} 0.975 & 0 & 0.042 & 0 \\ 0 & 0.977 & 0 & 0.044 \\ 0 & 0 & 0.958 & 0 \\ 0 & 0 & 0 & .956 \end{bmatrix}. \quad (10)$$

We let the system have the following sensing matrix:

$$C = [I_4 \quad I_4]^T. \quad (11)$$

The minimum value of $|\text{supp}(Cv)|$ is achieved by two eigenvectors of A . The vectors e_1 and e_2 (e_i is the i^{th} canonical vector in \mathbb{R}^4) are eigenvectors of A and the resulting value of s_0 is $s_0 = 2$. Any sensor attack on a single sensor is detectable. Let $K = \{1, 5\}$, $\hat{K} = \{3, 7\}$, and D_K and $D_{\hat{K}}$ be as defined in equation (2). The matrix D_K satisfies the condition in Theorem 2 since $Ce_1 \in \mathcal{R}(D_K)$, so there exists an undetectable attack against the system $\Sigma_K = (A, C, D_K)$. One particular undetectable attack sequence against Σ_K is $a(k)$ such that $D_K a(k) = \alpha(.975)^k e_1$, where $\alpha \in \mathbb{R}$ determines the magnitude of the attack and .975 is the eigenvalue associated with e_1 . On the other hand, $D_{\hat{K}}$ does not satisfy the condition in Theorem 2, so there is no undetectable attack against the system $\Sigma_{\hat{K}} = (A, C, D_{\hat{K}})$. Finally, consider the problem of adding two additional sensors to C to improve detection resilience to attacks. If the additional sensors measure states 1 and 2, the s_0 index of the system increases to $s_0 = 3$ and a dynamic detector can detect all attacks on two or fewer sensors. While this is by no means a comprehensive algorithm for sensor placement, the s_0 index offers a quantitative consideration of detector attack resilience in the design of the sensing matrix C .

5. CONCLUSION

In this paper, we study the dynamic detection of sensor attacks against cyber-physical systems. We relate the existence of dynamically undetectable sensor attacks to the system's strong observability, and we use properties of strong observability to provide a necessary and sufficient condition for attack sets to have undetectable sensor attacks. In addition, we provide an index s_0 that gives the minimum number of sensors an attacker must attack in order to be undetectable. Finally, we illustrate our results with a numerical example and demonstrate a design guideline using the s_0 index for improving the system resiliency to sensor attacks.

6. REFERENCES

- [1] Á. A. Cárdenas, S. Amin, and S. Sastry, “Research challenges for the security of control systems,” in *Proceedings of the 3rd Conference on Hot Topics in Security*, San José, CA, 2008, pp. 1–6.
- [2] Á. A. Cárdenas, S. Amin, Z. Lin, Y. Huang and C. Huang, and S. Sastry, “Attacks against process control systems: Risk assessment, detection, and response,” in *Proc. 6th ACM Symposium on Information, Computer and Communications Security*, Hong Kong, Mar. 2011, pp. 355–366.
- [3] S. Checkoway, D. McCoy, B. Kantor, D. Anderson, H. Shacham, S. Savage, K. Koscher, A. Czeskis, F. Roesner, and T. Kohno, “Comprehensive experimental analyses of automotive attack surfaces,” in *Proc. 2011 USENIX Security Symposium*, San Francisco, CA, Aug. 2011, pp. 1–14.
- [4] Y. Liu, M. K. Reiter, and P. Ning, “False data injection attacks against power systems in electric power grids,” in *Proc. 16th ACM Conference on Computer and Communications Security*, Chicago, IL, Nov. 2009, pp. 21–32.
- [5] 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.
- [6] 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, June 2014.
- [7] Y. Shoukry and P. Tabuada, “Event-triggered state observers for sparse sensor noise/attack,” *Arxiv e-prints*, Sept. 2013.
- [8] H. L. Trentelman, A. A. Stoorvogel, and M. Hautus, *Control Theory for Linear Systems*, chapter 7, Springer, 2001.
- [9] B. P. Molinari, “A strong controllability and observability in linear multivariate control,” *IEEE Transactions on Automatic Control*, vol. 21, no. 5, pp. 761–764, Oct. 1976.
- [10] Y. Mo and B. Sinopoli, “False data injection attacks in control systems,” in *Proc. 1st Workshop on Secure Control Systems*, Stockholm, Sweden, Apr. 2010, pp. 56–62.
- [11] A. Teixeira, I. Shames., H. Sandberg, and K. H. Johansson, “Revealing stealthy attacks in control systems,” in *Proc. 50th Annual Allerton Conference*, Urbana, IL, Oct. 2012, pp. 1806–1813.
- [12] B. P. Molinari, “Extended controllability and observability for linear systems,” *IEEE Transactions on Automatic Control*, vol. 21, no. 1, pp. 136–137, Feb. 1976.
- [13] C. T. Chen, *Linear System Theory and Design*, chapter 6, Oxford University Press, 1999.
- [14] K. H. Johansson, “The quadruple tank process: A multivariable laboratory process with an adjustable zero,” *IEEE Transactions on Control Systems Technology*, , no. 4, pp. 456–465, May 2000.