Automatica 97 (2018) 121–133

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Detection and mitigation of classes of attacks in supervisory control systems☆

Lilian Kawakami Carvalho [a,*], Yi-Chin Wu [b,c], Raymond Kwong [d], Stéphane Lafortune [b]

[a] Department of Electrical Engineering, Universidade Federal do Rio de Janeiro, Brazil
[b] Department of EECS, University of Michigan, USA
[c] Department of EECS, University of California at Berkeley, USA
[d] Department of ECE, University of Toronto, Canada

ARTICLE INFO

Article history:
Received 12 September 2016
Received in revised form 27 February 2018
Accepted 16 July 2018

Keywords:
Discrete event systems
Automata
Failure diagnosis
Cyber-attacks

ABSTRACT

The deployment of control systems with network-connected components has made feedback control systems vulnerable to attacks over the network. This paper considers the problem of intrusion detection and mitigation in supervisory control systems, where the attacker has the ability to enable or disable vulnerable actuator commands and erase or insert vulnerable sensor readings. We present a mathematical model for the system under certain classes of actuator enablement attacks, sensor erasure attacks, or sensor insertion attacks. We then propose a defense strategy that aims to detect such attacks online and disables all controllable events after an attack is detected. We develop an algorithmic procedure for verifying whether the system can prevent damage from the attacks considered with the proposed defense strategy, where damage is modeled as the reachability of a pre-defined set of unsafe system states. The technical condition of interest that is necessary and sufficient in this context, termed "GF-safe controllability", is characterized. We show that the verification of GF-safe controllability can be performed using diagnoser or verifier automata. Finally, we illustrate the methodology with a traffic control system example.

1. Introduction

The increasing amount of networked components in feedback control systems has made these systems vulnerable to cyber threats. Since control systems are often safety critical (e.g., avionics, power grid), it is imperative to embed defense mechanisms into them (Banerjee, Venkatasubramanian, Mukherjee, & Gupta, 2012; Cardenas, Amin, & Sastry, 2008).

In this paper, we consider the closed-loop control system architecture of Fig. 1, where the plant is controlled by the supervisor through sensors and actuators in the traditional feedback loop. The communication channels for the sensor and actuator signals are often unprotected, allowing attackers to potentially inject false sensor or actuator signals. We consider event-driven supervisory control systems where the plant is abstracted as a discrete event system. The supervisor monitors the plant behavior through the events generated by the sensors and it dynamically issues enable/disable actuator commands in order to enforce a given specification. We study the problem of intrusion detection and mitigation for control systems under four classes of attacks: *Actuator Enablement attacks* (AE-attacks), *Actuator Disablement attacks* (AD-attacks), *Sensor Erasure attacks* (SE-attacks) and *Sensor Insertion attacks* (SI-attacks). Specifically, in an attack scenario, some actuators or sensors are deemed vulnerable and the attacker can change the actuator commands (from disable to enable or vice-versa) or change the sensor readings (by erasing a genuine sensor event or inserting a fictitious one). We address the problem of protecting the system from reaching a pre-defined set of unsafe states under each of the above attack scenarios. Note that in general actuator attacks or sensor erasure attacks are not directly observable, while inserted fictitious sensor events are assumed to be indistinguishable from genuine ones for the supervisor. We leverage results from supervisory control and fault diagnosis of discrete event systems and propose a defense strategy that detects attacks online and disables all controllable actuator events

☆ This work was partially supported by the U.S. National Science Foundation (grant CNS-1421122), by Brazil's CNPq (National Council of Technological and Scientific Development), and by the Natural Sciences and Engineering Research Council of Canada (grant RGPIN-2015-04273). The material in this paper was partially presented at the 13th International Workshop on Discrete Event Systems (WODES 2016), May 30–June 1, 2016, Xi'an, China. This paper was recommended for publication in revised form by Associate Editor Christoforos Hadjicostis under the direction of Editor Christos G. Cassandras.

* Corresponding author.
E-mail addresses: lilian@dee.ufrj.br (L.K. Carvalho), yichin.wu@berkeley.edu (Y.-C. Wu), kwong@control.utoronto.ca (R. Kwong), stephane@umich.edu (S. Lafortune).

https://doi.org/10.1016/j.automatica.2018.07.017
0005-1098/

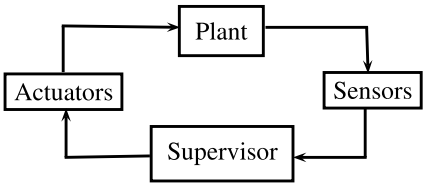

Fig. 1. The closed-loop control system architecture.

after detecting an attack with certainty. This defense strategy may not be sufficient in general to prevent damage. Hence, we characterize a property termed *General Form of safe controllability* (GF-safe controllability for short) that precisely captures the capability of preventing the system from reaching an unsafe state after an attack, using the proposed defense strategy. Here, GF stands for AE, SE, or SI. An algorithmic procedure is developed to verify whether the system is GF-safe controllable. For this purpose, diagnoser or verifier automata can be employed.

The key feature distinguishing this work from the large amount of work in cybersecurity is our focus on closed-loop control systems. We adopt a model-based approach to precisely capture the vulnerabilities and the effects of an attack on the control system. The model-based approach enables a formal characterization of the unsafe behavior that an attacker tries to induce and the resiliency that the system defender wants to achieve. The model-based approach also allows for monitoring deviations from the normal system behavior. Our work is complementary to the works on anomaly/intrusion detection in cyber systems (e.g., Hoffman, Zage, & Nita-Rotaru, 2009; Lazarevic, Kumar, & Srivastava, 2005; Modi et al., 2013; Zhou, Leckie, & Karunasekera, 2010) where detection is based on statistical analysis of network packets, for instance. We do not focus on how attackers infiltrate vulnerable actuators or sensors, but rather on the detection of attacks and on the modeling of their effects on the control system. Under each of the four types of attacks considered, we adopt a fairly simple attack model which can be paraphrased as "attack whenever possible". However, our methodology is general and more sophisticated attack models could be embedded in it. Similarly, our defense strategy upon detection of attacks is based on "safety first", by switching to a "safe mode" of operation, but more refined defense mechanisms could be embedded in our modeling methodology, if so desired.

Intrusion detection and prevention in the setting of supervisory control of discrete event systems have been previously studied in Thorsley and Teneketzis (2006), where the authors consider the design of a supervisor that achieves the specification both in normal operation and after an attack. The focus in Thorsley and Teneketzis (2006) is on finding language conditions under which the supervisor can prevent unsafe behavior in the presence of attacks while achieving a given specification, using a notion called *disable language*, which shares several similarities with the safe controllability condition used in this paper. Our focus is more explicit than Thorsley & Teneketzis (2006) in terms of modeling several classes of attacks, detecting them algorithmically using diagnoser automata, and switching to safe mode upon detection. The problem of intrusion detection and prevention is related to fault tolerant supervisory control problems, a well-studied problem in the literature (see, e.g., Moor, 2015; Nke & Lunze, 2011; Paoli, Sartini, & Lafortune, 2011; Rohloff, 2005; Sulek & Schmidt, 2014; Wen, Kumar, & Huang, 2014), where a robust supervisor is designed to maintain the specification even when the system becomes faulty. Our approach is closest to the work in Paoli et al. (2011), where the authors consider a strategy that detects faults online and reconfigures the control law when a fault is detected. Our notion of GF-safe controllability is a GF-attack variant of the *safe controllability* property introduced in Paoli et al. (2011).

The main contributions of this paper are as follows. First, we present a mathematical model for supervisory control systems under AE-attacks and propose a defense strategy that detects attacks online and, upon detection with certainty, disables all controllable events in order to prevent attack damage. We define the property of *AE-safe controllability* that characterizes the system's capability to prevent damage under AE-attacks and develop algorithmic procedures for verifying AE-safe controllability using diagnoser and verifier automata. Next, we consider other types of attacks. We only briefly discuss AD-attacks and focus instead on SE- and SI-attacks. Paralleling the case of AE-attacks, we model the effect of SE- and SI-attacks on the control system. For AE- and SE-attacks, we consider a worst-case scenario where the attacker may attack at every opportunity. For SI-attacks, we consider an attack strategy where the attacker never inserts a sensor reading that is not defined in the current state of the nominal supervisor. We then generalize AE-safe controllability to GF-safe controllability, the property that the system should satisfy in order to successfully prevent damage from either AE-, SE- or SI-attacks, and finally we develop a test to verify GF-safe controllability. In the case of SE- and SI-attacks, in addition to testing the corresponding version of GF-safe-controllability, it is also necessary to test if the control system under attack has a deadlock.

The remainder of this paper is organized as follows. We define the types of attacks we deal with in Section 2. Section 3 introduces our mathematical framework. Section 4 studies the effect of actuator enablement attacks. Then, in Section 5, we define the property of AE-safe controllability and discuss its verification. We present the model of the system under sensor erasure and insertion attacks in Sections 6 and 7, respectively. In Section 8, we define the property of GF-safe controllability and present an algorithm for its verification. Finally, in Section 9, we illustrate our methodology with a traffic control system example and in Section 10, we conclude the paper.

A preliminary and partial version of the results in Sections 4 and 5 was presented in Carvalho, Wu, Kwong, and Lafortune (2016). The results in Sections 5.4, 6, 7, and 8 are new. Owing to space limitations, proofs have been omitted. The extended version of this paper with proofs can be found in Carvalho, Wu, Kwong, and Lafortune (2018).

2. Types of attacks

We depict in Fig. 2 the attack model under consideration. The control system architecture under attack has a plant G equipped with a set of potentially vulnerable sensors and actuators, and G is controlled by a partial-observation supervisor (or P-supervisor) S_P. Let E be the event set of G. The actuators are modeled by the set of *controllable* events E_c, with $E_c \subseteq E$, while the sensors are modeled by the set of *observable* events E_o, with $E_o \subseteq E$. The supervisor observes the occurrences of the plant's observable events through projection P_o from set E to set E_o. The attacker, represented by block A, has access to subsets of E_c and E_o, representing *vulnerable* actuators and sensors and denoted by $E_{c,v} \subseteq E_c$ and $E_{o,v} \subseteq E_o$, respectively. The sets $E_{c,v}$ and $E_{o,v}$ are predefined based on system knowledge and are application dependent. They can, for example, reflect the capability of the attacker to exploit vulnerabilities of the system. Finally, block G_D is the module that detects attacks, which we call the *intrusion detection module*.

The fact that the attacker can compromise either sensors or actuators is captured by the two outputs of A that affect: (i) the actual observations of S_P and G_D, which consist of the genuine sensor readings affected by the attacks on them; and (ii) the actual control actions that are applied to G, which consist of the combination of the genuine control actions of S_P with those of A. The combination of the attacks of A with genuine sensor readings