Fifty-fourth Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
September 27 - 30, 2016

Recursive Reachable Set Computation for On-line Safety Assessment of the Cyber-Physical System against Stealthy Cyber Attacks

Cheolhyeon Kwon and Inseok Hwang

***Abstract*—With recent progress in networked embedded control technology, cyber attacks have become one of the major threats to Cyber-Physical Systems (CPSs) due to their close integration of logical and physical processes. Since the conventional computer security techniques are unable to assure the underlying physical behavior against cyber attacks, this paper considers the safety of the compromised CPS from a controls domain perspective. Specifically, we propose an on-line algorithm to assess the safety of the CPS in the presence of stealthy cyber attacks which can be designed intelligently to avoid detection. The main idea is based on a reachability analysis that computes the reachable set of CPS states possibly reached by all potential stealthy cyber attacks. The reachable set computation typically demands a large computation cost and has mostly relied on the approximation techniques. However, our algorithm analytically derives the exact reachable set solution and further establishes a recursive computation structure that can perform in the real-time CPS operation. This significantly enhances the quality of the on-line safety assessment, enabling more reliable, less conservative, and computationally efficient process. The proposed algorithm is demonstrated with an illustrative example of an unmanned aircraft system (UAS) application.**

I. Introduction

Recent technological advances have introduced the world to Cyber Physical Systems (CPSs) whose computation and communication capabilities afford significant advantages in many dynamical applications, however, at the cost of possible vulnerability to cyber attacks. Traditionally, cyber attacks have been studied in the computer science field, focusing on issues such as trustworthiness of data flow [1], [2], but these methods alone are not sufficient to diagnose a CPS's physical processes such as its real-time dynamic behaviors [3], [4].In seeking to address this problem, rather than focus on the computer security of the system, we propose a controls domain framework to include the compromised physical behavior [5], [6], [7].

It is also important to note that conventional control approach cannot assure the safety of a compromised CPS. This is mainly due to the unpredictable nature of cyber attacks and an inability to distinguish between attack features and common disturbances and faults. Accounting for the potential cyber threats to CPSs, we have done preliminary work that thoroughly analyzed the attack severity [8], [9], [10] and detectability [11], [12]. Specifically, the attack detectability analysis provides some insight into which types of cyber attacks are more difficult to detect and how such attacks can be designed. This paper further extends the attack detectability work to enhance the safety of a CPS subject to stealthy cyber attacks capable of significantly altering CPS behavior while avoiding detection by the existing detection schemes.

The objectives of this paper are to diagnose CPS safety while considering the possibility of stealthy cyber attacks, and to develop an on-line safety assessment algorithm to help increase system safety in cases where there are stealthy cyber attacks. To determine whether the CPS is safe or not, reachability analysis is used without needing any specification to the injected cyber attack. Setting the safe region based on the current state estimate and environment, we compute the reachable set of all CPS states that can be reached by all possible stealthy cyber attacks and compare it with the safe region. Then, if the entire reachable set lies within the safe region, we can assure with a certain probability that the current CPS is safe even if there is a cyber attack. Otherwise, if there are any parts of the reachable set outside the safe region, the CPS could potentially be unsafe even if no attack is detected.

The main difficulty is to compute the reachable set, which is computationally complex. The existing research considers the over-approximated reachable set rather than dealing with the intensive computation cost of calculating the accurate reachable set [13], [14]. However, such an over-approximation approach can be excessively conservative and may not provide informative enough results, especially for real-time CPS operations. Here, our algorithm analytically derives the exact reachable set via linear matrix inequalities and further establishes a recursive computation structure. Hence, the reachable set computation developed in this paper is computationally efficient, significantly enhancing the quality of the on-line safety assessment while augmenting the security-monitoring systems already in place in the CPS.

The rest of this paper is organized as follows: In Section II, we describe a mathematical model of the CPS dynamics subject to cyber attacks and the attack detection mechanism. Additionally, the safety assessment of the CPS subject to stealthy cyber attacks is posed as a mapping function. Section III presents an on-line safety assessment algorithm to verify the safety of the CPS, followed by a recursive computation structure of an analytical reachable set solution in Section IV. An illustrative CPS example under cyber attack is presented to demonstrate the proposed algorithm in Section V. Conclusions

The authors are with the School of Aeronautics and Astronautics, Purdue University, West Lafayette, IN, 47907 USA (email: kwonc@purdue.edu, ihwang@purdue.edu)

are given in Section VI.

II. Problem Formulation

In this section, we present the mathematical models to describe the dynamics of the CPS subject to cyber attacks and detail an embedded attack detection scheme that leads to the safety assessment problem. The CPS dynamics under a cyber attack is modeled as a linear time-invariant system. In the CPS, actuators and sensors are usually connected by communication channels which are susceptible to malicious data injection by cyber attacks. The CPS model subject to cyber attacks is given by:

$$\begin{aligned} x_a(k+1) &= Ax_a(k) + Bu(k) + B_c a_c(k) + w(k) \\ y_a(k) &= Cx_a(k) + B_o a_o(k) + v(k) \end{aligned} \tag{1}$$

where $x_a(k) \in \mathbb{R}^n$ (the subscript 'a' means the system under attack), $u(k) \in \mathbb{R}^p$, $y_a(k) \in \mathbb{R}^m$ are the system's state, input, and output respectively; and $w(k) \in \mathbb{R}^n$, $v(k) \in \mathbb{R}^m$ are the process and measurement noise. It is assumed that the input $u(k)$ is known and w, v are the zero mean Gaussian white noise with constant covariance matrices Q and R respectively. The *attack sequences* $a_c \in \mathbb{R}^s(k)$ and $a_o(k) \in \mathbb{R}^q$ are injected into actuators and sensors, respectively, with the *attack matrices* B_c and B_o of compatible dimensions. Note that the system matrix pairs (A, B) and (C, A) satisfy the controllability and observability conditions.

With the sensor observations, the monitoring system keeps track of the CPS's state using a linear state estimator. Let $\hat{x}_a(k)$ be the estimate of the system's state under attack. Then the dynamics of the linear estimator can be represented as:

$$\begin{aligned} \hat{x}_a(k+1) = A\hat{x}_a(k) + Bu(k) + L\big(y_a(k+1) \\ - CA\hat{x}_a(k) - CBu(k)\big) \end{aligned} \tag{2}$$

where L is the estimator gain. In this work we consider the steady-state Kalman filter (KF). Then, L is the steady-state Kalman gain given by $L = \Sigma_P C^{\mathrm{T}}(C\Sigma_P C^{\mathrm{T}} + R)^{-1}$ where Σ_P is the predicted error covariance matrix which is the solution to the following discrete-time algebraic Riccati equation:

$$\begin{aligned} -\Sigma_P + A\Sigma_P A^{\mathrm{T}} + Q - A\Sigma_P C^{\mathrm{T}} \left(C\Sigma_P C^{\mathrm{T}} + R\right)^{-1} \\ \times C\Sigma_P A^{\mathrm{T}} = 0 \end{aligned}$$

Since a cyber attack can be regarded as a fault in the CPS, existing fault diagnosis algorithms could be used for cyber attack detection. We consider one of the most common fault detection algorithms using the residuals generated by the steady-state KF. The residual is defined as:

$$r(k+1) := \; y(k+1) - CA\hat{x}(k) - CBu(k) \tag{3}$$

Without cyber attacks, the residual has a zero-mean Gaussian distribution with a constant covariance matrix $\Sigma_r = C\Sigma_P C^T + E_2 R E_2^T$. Therefore, cyber attacks can be diagnosed by testing the following two incompatible statistical hypotheses.

$$\mathcal{H}_0 : \; r(k) \sim \mathcal{N}(\mathbf{0}, \Sigma_r) \quad \text{and} \quad \mathcal{H}_1 : \; r(k) \not\sim \mathcal{N}(\mathbf{0}, \Sigma_r)$$

where $\mathcal{N}(\mathbf{a}, \Sigma)$ represents the Gaussian distribution with mean $\mathbf{a}$ and covariance Σ. There are various statistical hypothesis testing algorithms that can be used such as the Sequential Probability Ratio Test (SPRT) [15], the cumulative sum (CUSUM) [16], and Generalized Likelihood Ratio (GLR) test [17]. In this paper, we consider the hypothesis test by checking the 'power' of the residual, $r^{\mathrm{T}}(k)\Sigma_r^{-1}r(k)$, known as the *Compound Scalar Testing* [18]:

$$\begin{cases} \text{Accept} \;\; \mathcal{H}_0 & \text{If,} \quad r^{\mathrm{T}}(k)\Sigma_r^{-1}r(k) \le h \\ \text{Accept} \;\; \mathcal{H}_1 & \text{If,} \quad r^{\mathrm{T}}(k)\Sigma_r^{-1}r(k) > h \end{cases} \tag{4}$$

where $h > m$ (note that m is the dimension of r) is a threshold value. If there are no faults or attacks, $r^{\mathrm{T}}(k)\Sigma_r^{-1}r(k)$ follows a χ^2 distribution and is highly likely to accept the hypothesis $\mathcal{H}_0$ due to the statistical characteristics of the χ^2 distribution. If $r^{\mathrm{T}}(k)\Sigma_r^{-1}r(k) > h$, $\mathcal{H}_1$ is accepted and the algorithm declares a fault in the system which may be induced by cyber attacks. Since the entire testing process is stochastic, there is a probability of false alarm according to the preset threshold value h. Therefore, the attack detection performance and false attack alarm rate depend on the design parameter h, i.e., the higher threshold lowers the false alarm rate while making the system less sensitive to cyber attacks. Such security holes can be exploited for a more dangerous attack strategy, as having the attack be less detectable provides a clear advantage to the attacker. In particular for the detection scheme (4), the stealthy attacker can avoid triggering an alarm by not causing a large increase in the residual power.

Definition 1. Let a_c^k, a_o^k denote the vectors that contain individual attack vector sequences up to time step k:

$$\begin{aligned} a_c^k &:= \left[a_c^{\mathrm{T}}(0) \; a_c^{\mathrm{T}}(1) \cdots a_c^{\mathrm{T}}(k-1)\right]^{\mathrm{T}} \\ a_o^k &:= \left[a_o^{\mathrm{T}}(1) \; a_o^{\mathrm{T}}(2) \cdots a_o^{\mathrm{T}}(k)\right]^{\mathrm{T}} \end{aligned}$$

Then, $\mathcal{A}(k) \subseteq \mathbb{R}^{sk} \times \mathbb{R}^{qk}$ is the set of cyber attack sequences, respectively injected into the actuators and the sensors while not being detected such that:

$$\mathcal{A}(k) := \left\{ \left(a_c^k, a_o^k\right) \in \mathbb{R}^{sk} \times \mathbb{R}^{qk} \middle| \mathbb{E}[r^{\mathrm{T}}(k)\Sigma_r^{-1}r(k)] \le h \right\} \tag{5}$$

where $\mathbb{E}[\bullet]$ denotes the expected value.

If the CPS does not recognize the injected attacks, it cannot stop the actual state from deviating from the normal condition, and thus its safety could be compromised. Specifically, this paper considers the attacks affecting the state estimator in the CPS which is a critical component and used for various subsystems. Therefore, the attacker's intent is to fail the KF by causing a state estimation error. As the measure of whether such an error is safe or not, we consider the safe region as follows:

Definition 2. At each time step k, a subset $G(k) \subset \mathbb{R}^n$ is assigned as the set of safe states whose boundary is described by an ellipsoid given by:

$$G(k) := \left\{ x \in \mathbb{R}^n \middle| (x - \hat{x}_a(k))^{\mathrm{T}} P(k) (x - \hat{x}_a(k)) \le 1 \right\} \tag{6}$$

where $P(k) \in \mathbb{R}^{n\times n}$ is a time-varying positive definite matrix.

Against stealthy cyber attacks, the safety property of the CPS at time step k, denoted by the map $F_k : \mathcal{A}(k) \to \{\text{safe}, \text{unsafe}\}$, is defined as:

$$F_k\left(a_c^k, a_o^k\right) := \begin{cases} \text{safe} & \text{if } \ x_a(k) \in G(k) \\ \text{unsafe} & \text{otherwise} \end{cases} \tag{7}$$

Therefore, the safety of the CPS is guaranteed with a certain probability dependent on the stochastic properties of the system if the attacked state $x_a(k)$ always remains in the set $G(k)$. Unfortunately, neither the cyber attack sequence nor the actual state is accurately traceable, making (7) difficult to evaluate. In the next section, we present an algorithm to verify the safety of a CPS based on reachability analysis.

III. On-line Safety Assessment Algorithm

This section presents an algorithm that can verify the safety of the CPS subject to cyber attacks. Note that detectable attacks are not of particular concern in this research. Rather, we focus on the stealthy cyber attacks in Definition 1. Let $\text{Reach}(x_a, k) \subset \mathbb{R}^n$ be the reachable set of the CPS state driven by stealthy cyber attacks from time step 0 to k:

$$\text{Reach}(x_a, k) := \left\{x_a(k) \middle| x_a(0) = \hat{x}_a(0), \left(a_c^k, a_o^k\right) \in \mathcal{A}(k)\right\}$$

Considering the estimation error defined as $e_a := x_a - \hat{x}_a$, $\text{Reach}(x_a, k)$ can be rewritten by:

$$\text{Reach}(x_a, k) \equiv \left\{\hat{x}_a(k) + e_a \middle| e_a \in \text{Reach}(e_a, k)\right\}$$

where

$$\text{Reach}(e_a, k) = \left\{e_a(k) \middle| \mathbb{E}[e_a(0)] = \mathbf{0}, \left(a_c^k, a_o^k\right) \in \mathcal{A}(k)\right\}$$

Then, recalling Definition 2 and (7), the safety property of the CPS always holds under the following equivalence:

$$\begin{aligned} F_k\left(a_c^k, a_o^k\right) = \text{safe}, \quad & \forall \left(a_c^k, a_o^k\right) \in \mathcal{A}(k) \\ \iff \ & \hat{x}_a(k) + e_a \in G(k), \quad \forall e_a \in \text{Reach}(e_a, k) \end{aligned} \tag{8}$$

Consequently, the safety of the CPS depends on the reachable set of the estimation error $\text{Reach}(e_a, k)$. Note that, even if $\mathbb{E}[e_a(k)] = \mathbf{0}$, that does not necessarily imply $\text{Reach}(e_a, k) = \{\mathbf{0}\}$ due to the stochastic nature of the CPS dynamics (1). To be more precise, $\text{Reach}(e_a, k)$ is the reachable set of a stochastic system with unbounded uncertainties (in this case, Gaussian random variables), yet computing it is not practically feasible because the unbounded errors give a probability (which could be small) that the system could be in any possible state. One typical way to deal with this problem is the probabilistic approach in which the reachable set is computed according to the statistical characteristics of the system, i.e. the reachable set is defined as the set of states that the system has a probability greater than a certain threshold value of being in. Specifically, if $\mathbb{E}[e_a(k)] = \mathbf{0}$, we set $\text{Reach}(e_a, k)$ as the interior points of a bounded region derived from the error covariance such that:

$$\text{Reach}(e_a, k) = \left\{e_a \middle| e_a e_a^{\mathrm{T}} \preceq \alpha \text{Cov}[e_a]\right\} \tag{9}$$

where the system is assumed to have entered the steady state (stationary), i.e., $\text{Cov}[e_a] = (I - LC_G)\Sigma_P$. Here, $\alpha > 0$ is a design parameter directly related to the chosen threshold probability that determines the confidence level of the safety assessment algorithm as follows:

$$\alpha = \chi_n^2\left(p\left[e_a(k) \in \text{Reach}(e_a, k) \middle| \mathbb{E}[e_a(k)] = \mathbf{0}\right]\right) \tag{10}$$

where $p[\bullet|\bullet]$ denotes the conditional probability and $\chi_n^2 : \mathbb{R} \to \mathbb{R}$ is the quantile function of the chi-squared distribution with n degrees of freedom [19]. Therefore, the on-line safety assessment algorithm can be posed as two steps: (i) computing the reachable set $\text{Reach}(e_a, k)$; and (ii) examining the resulting set by comparing with the safe region $G(k)$, as illustrated in Figure 1.

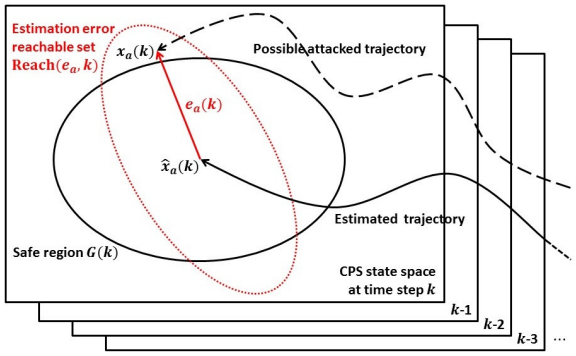

Fig. 1: CPS Safety Assessment via Reachability Analysis

To compute $\text{Reach}(e_a, k)$, we first derive the explicit form of $\mathcal{A}(k)$ in the attack space $\mathbf{R}^{sk\times qk}$ from the residual condition $\mathbb{E}[r^{\mathrm{T}}(k)\Sigma_r^{-1}r(k)] \le h$, and project it onto the estimation error space $\mathbf{R}^n$. This results in the following compact form of the exact reachable set solution:

$$\text{Reach}(e_a, k) = \left\{e_a \middle| e_a^{\mathrm{T}}\Upsilon^{-1}(k)e_a \le 1\right\} \tag{11}$$

where $\Upsilon(k) \in \mathbb{R}^{n\times n}$ is defined by:

$$\begin{aligned} \Upsilon(k) :=& (h-m)\begin{bmatrix}\Psi_c(k) & \Psi_o(k)\end{bmatrix}\begin{bmatrix}\Theta_c(k) & \mathbf{0}_{sk,qk} \\ \mathbf{0}_{qk,sk} & \Theta_o(k)\end{bmatrix}^{-1} \\ & \times \begin{bmatrix}\Psi_c(k) & \Psi_o(k)\end{bmatrix}^{\mathrm{T}} + \alpha(I - LC)\Sigma_P \end{aligned} \tag{12}$$

In (12), $\mathbf{0}_{n,m}$ denotes an $n \times m$ dimensional null matrix and block matrices $\Psi_c(k) \in \mathbb{R}^{n\times sk}$, $\Psi_o(k) \in \mathbb{R}^{n\times qk}$, $\Theta_c(k) \in \mathbb{R}^{sk\times sk}$, and $\Theta_o(k) \in \mathbb{R}^{qk\times qk}$ are respectively given by:

$$\begin{aligned} \Psi_c(k) :=& \Big[(A - LCA)^{k-1}(I - LC)B_c \\ & (A - LCA)^{k-2}(I - LC)B_c \ \cdots \ (I - LC)B_c\Big] \\ \Psi_o(k) :=& \Big[-(A - LCA)^{k-1}LB_o \quad -(A - LCA)^{k-2}LB_o \\ & \cdots \ -LB_o\Big] \end{aligned} \tag{13}$$

$$\Theta_c(k) := \begin{bmatrix} CA\Psi_c(k-1) & CB_c \end{bmatrix}^{\mathrm{T}} \Sigma_r^{-1} \begin{bmatrix} CA\Psi_c(k-1) & CB_c \end{bmatrix}$$
$$\Theta_o(k) := \begin{bmatrix} CA\Psi_o(k-1) & B_o \end{bmatrix}^{\mathrm{T}} \Sigma_r^{-1} \begin{bmatrix} CA\Psi_o(k-1) & B_o \end{bmatrix} \tag{14}$$

Due to the space limitation, the detailed derivations are omitted and will be presented in an upcoming journal paper containing the results of this paper. Using the computed reachable set, the safety of the CPS can be determined by:

$$\begin{aligned} F_k\left(a_c^k, a_o^k\right) = \text{safe}, \quad \forall \left(a_c^k, a_o^k\right) \in \mathcal{A}(k) \\ \iff e_a^{*\mathrm{T}}(k) P(k) e_a^*(k) \leq 1 \end{aligned} \tag{15}$$

where e_a^* is the solution to the following optimization problem at each time step k:

$$\begin{aligned} &\max_{e_a} e_a^{\mathrm{T}} P(k) e_a(k) \\ &\text{s.t. } e_a^{\mathrm{T}} \Upsilon^{-1}(k) e_a \leq 1 \end{aligned} \tag{16}$$

IV. Recursive Reachable Set Computation

The presented reachability analysis is meant to perform the on-line safety assessment, being able to run with the admissible computation cost that ensures the real-time operation. However, the result of Section III, i.e., the analytical reachable set solution (11) is still computationally demanding as it takes into account the entire past attack history at once. Specifically, the intermediate matrix terms such as Θ_c, and Θ_o are growing their dimensions with time, i.e., one has to carry out more intensive computation as time goes by. This prompts the need of a recursive computation structure where the update for each time step is repeated in a self-similar way with an equal computation cost.

Algorithmically, we want to induce the matrices for the time step $k+1$ from the time step k. From (13), Ψ_c, Ψ_o can be formulated as the following recursive forms:

$$\begin{aligned} \Psi_c(k+1) &= \begin{bmatrix} (A - LCA)^k (I - LC) B_c & \Psi_c(k) \end{bmatrix} \\ \Psi_o(k+1) &= \begin{bmatrix} -(A - LCA)^k L B_o & \Psi_o(k) \end{bmatrix} \end{aligned} \tag{17}$$

Applying (17) to (14), we have:

$$\begin{aligned} \Theta_c(k+1) &= \begin{bmatrix} X_c(k) & U_c(k) \\ U_c^{\mathrm{T}}(k) & \Theta_c(k) \end{bmatrix} \\ \Theta_o(k+1) &= \begin{bmatrix} X_o(k) & U_o(k) \\ U_o^{\mathrm{T}}(k) & \Theta_o(k) \end{bmatrix} \end{aligned} \tag{18}$$

where the block matrices X_c, X_o, U_c, and U_o are:

$$\begin{aligned} X_c(k) &= \left(CA(A - LCA)^{k-1}(I - LC)B_c\right)^{\mathrm{T}} \Sigma_r^{-1} \\ &\quad \times \left(CA(A - LCA)^{k-1}(I - LC)B_c\right) \\ X_o(k) &= \left(CA(A - LCA)^{k-1} L B_o\right)^{\mathrm{T}} \Sigma_r^{-1} \\ &\quad \times \left(CA(A - LCA)^{k-1} L B_o\right) \\ U_c(k) &= \left(CA(A - LCA)^{k-1}(I - LC)B_c\right)^{\mathrm{T}} \Sigma_r^{-1} \\ &\quad \times \begin{bmatrix} CA\Psi_c(k-1) & CB_c \end{bmatrix} \\ U_o(k) &= \left(CA(A - LCA)^{k-1} L B_o\right)^{\mathrm{T}} \Sigma_r^{-1} \\ &\quad \times \begin{bmatrix} CA\Psi_o(k-1) & B_o \end{bmatrix} \end{aligned} \tag{19}$$

To compute the large matrix inversion in (12), let us consider the following two matrix lemmas:

Lemma 1. If an invertible matrix is partitioned into four blocks, it can be inverted blockwise as follows:

$$\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}^{-1} = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$$

where

$$\begin{aligned} V_{11} &= \left(P_{11} - P_{12} P_{22}^{-1} P_{21}\right)^{-1} \\ V_{12} &= -\left(P_{11} - P_{12} P_{22}^{-1} P_{21}\right)^{-1} P_{12} P_{22}^{-1} \\ V_{21} &= -P_{22}^{-1} P_{21} \left(P_{11} - P_{12} P_{22}^{-1} P_{21}\right)^{-1} \\ V_{22} &= \left(P_{22} - P_{21} P_{11}^{-1} P_{12}\right)^{-1} \end{aligned}$$

Lemma 2. Let X, Y, and $(Y^{-1} + VX^{-1}U)$ be nonsingular square matrices. Then $X + UYV$ is invertible, and

$$(X + UYV)^{-1} = X^{-1} - X^{-1} U \left(Y^{-1} + VX^{-1}U\right)^{-1} VX^{-1}$$

The proofs of both lemmas can be found in [20]. From Lemma 1, the following equality holds:

$$\begin{bmatrix} \Theta_c(k) & \mathbf{0}_{sk,qk} \\ \mathbf{0}_{qk,sk} & \Theta_o(k) \end{bmatrix}^{-1} \equiv \begin{bmatrix} \Theta_c^{-1}(k) & \mathbf{0}_{sk,qk} \\ \mathbf{0}_{qk,sk} & \Theta_o^{-1}(k) \end{bmatrix} \tag{20}$$

And applying Lemmas 1 and 2 to (18), the individual inverse matrices can be given by:

$$\begin{aligned} \Theta_c^{-1}(k+1) &= \begin{bmatrix} V_{c1}(k) & V_{c2}(k) \\ V_{c2}^{\mathrm{T}}(k) & V_{c3}(k) \end{bmatrix} \\ \Theta_o^{-1}(k+1) &= \begin{bmatrix} V_{o1}(k) & V_{o2}(k) \\ V_{o2}^{\mathrm{T}}(k) & V_{o3}(k) \end{bmatrix} \end{aligned} \tag{21}$$

where

$$\begin{aligned} V_{c1}(k) &= \left(X_c(k) - U_c(k)\Theta_c^{-1}(k)U_c^{\mathrm{T}}(k)\right)^{-1} \\ V_{c2}(k) &= -\left(X_c(k) - U_c(k)\Theta_c^{-1}(k)U_c^{\mathrm{T}}(k)\right)^{-1} U_c(k)\Theta_c^{-1}(k) \\ V_{c3}(k) &= \Theta_c^{-1}(k) - \Theta_c^{-1}(k)U_c^{\mathrm{T}}(k)\Big(X_c(k) - \\ &\qquad U_c(k)\Theta_c^{-1}(k)U_c^{\mathrm{T}}(k)\Big)^{-1} U_c(k)\Theta_c^{-1}(k) \\ V_{o1}(k) &= \left(X_o(k) - U_o(k)\Theta_o^{-1}(k)U_o^{\mathrm{T}}(k)\right)^{-1} \\ V_{o2}(k) &= -\left(X_o(k) - U_c(k)\Theta_o^{-1}(k)U_o^{\mathrm{T}}(k)\right)^{-1} U_o(k)\Theta_o^{-1}(k) \\ V_{o3}(k) &= \Theta_o^{-1}(k) - \Theta_o^{-1}(k)U_o^{\mathrm{T}}(k)\Big(X_o(k) - \\ &\qquad U_o(k)\Theta_o^{-1}(k)U_o^{\mathrm{T}}(k)\Big)^{-1} U_o(k)\Theta_o^{-1}(k) \end{aligned} \tag{22}$$

With the inverse of $\Theta_c(k)$ and $\Theta_o(k)$ available, it is only necessary to find the inverse of $X_c(k) - U_c(k)\Theta_c^{-1}(k)U_c^{\mathrm{T}}(k)$ and $X_o(k) - U_o(k)\Theta_o^{-1}(k)U_o^{\mathrm{T}}(k)$ (i.e., $V_{c1}(k)$ and $V_{o1}(k)$) in order to obtain $\Theta_c^{-1}(k+1)$ and $\Theta_o^{-1}(k+1)$ respectively. Since $V_{c1}(k) \in \mathbb{R}^{s \times s}$ and $V_{o1}(k) \in \mathbb{R}^{q \times q}$ remain small dimensions regardless of the time steps, computing them is more efficient than directly inverting $\Theta_c(k+1)$ and $\Theta_o(k+1)$.

Finally, combining (17), (20), and (21) all together, the intensive reachable set computation with the large matrix

inversions can be converted into the following recursive form in terms of Υ that can be used in the examination (16):

$$\begin{aligned}\Upsilon(k+1) = (h-m)\Big((A-LCA)^k(I-LC)B_cV_{c1}(k) \\ \times B_c^{\mathrm{T}}(I-LC)^{\mathrm{T}}(A-LCA)^{k\mathrm{T}} - (A-LCA)^k \\ \times (I-LC)B_cV_{c1}(k)U_c(k)\Theta_c^{-1}(k)\Psi_c^{\mathrm{T}}(k) \\ - \Psi_c(k)\Theta_c^{-1}(k)U_c^{\mathrm{T}(k)}V_{c1}(k)B_c^{\mathrm{T}}(I-LC)^{\mathrm{T}} \\ \times (A-LCA)^{k\mathrm{T}} - \Psi_c(k)\Theta_c^{-1}(k)U_c^{\mathrm{T}(k)}V_{c1}(k) \\ \times U_c(k)\Theta_c^{-1}(k)\Psi_c^{\mathrm{T}}(k) + (A-LCA)^kLB_o \\ \times V_{o1}(k)B_o^{\mathrm{T}}L^{\mathrm{T}}(A-LCA)^{k\mathrm{T}} + (A-LCA)^k \\ \times LB_oV_{o1}(k)U_o(k)\Theta_o^{-1}(k)\Psi_o^{\mathrm{T}}(k) + \Psi_o(k) \\ \times \Theta_o^{-1}(k)U_o^{\mathrm{T}(k)}V_{o1}(k)B_o^{\mathrm{T}}L^{\mathrm{T}}(A-LCA)^{k\mathrm{T}} \\ - \Psi_o(k)\Theta_o^{-1}(k)U_o^{\mathrm{T}(k)}V_{o1}(k)U_o(k)\Theta_o^{-1}(k) \\ \times \Psi_o^{\mathrm{T}}(k)\Big) + \Upsilon(k)\end{aligned} \tag{23}$$

The converted recursion (23) consists of a few matrix multiplications with the output from the previous time step, and thus significantly reduces the computation cost compared to (12).

The overall recursive structure for the on-line safety assessment algorithm is summarized as Algorithm 1.

Algorithm 1: On-line Safety Assessment Algorithm

Initialization: given the design parameters h and α,

- $\Theta_c(1) = B_c^{\mathrm{T}}C^{\mathrm{T}}\Sigma_r^{-1}CB_c$
- $\Theta_o(1) = B_o^{\mathrm{T}}\Sigma_r^{-1}B_o$
- $\Upsilon(1) = (h-m)\big(B_c^{\mathrm{T}}(I-LC)^{\mathrm{T}}\Theta_c^{-1}(1)(I-LC)B_c + B_o^{\mathrm{T}}L^{\mathrm{T}}\Theta_o^{-1}(1)LB_c\big) + \alpha(I-LC)\Sigma_P$

for $k = 1$ to the termination time N

a) Reachable set computation

1) Compute X_c, X_o, U_c, and U_o using (19)
2) Compute V_{c1}, V_{c2}, V_{c3}, V_{o1}, V_{o2}, and V_{o3} using (22)
3) Update Θ_c^{-1} and Θ_o^{-1} using (21)
4) Update Υ using (23)
5) **Output** $\mathrm{Reach}(e_a, k)$ from (11)

b) Reachable set examination

6) Compute e_a^* through the optimization (16)
7) **Output**
   - *safe* $\leftarrow e_a^{\mathrm{T}}P(k)e_a \leq 1$
   - *unsafe* $\leftarrow e_a^{\mathrm{T}}P(k)e_a > 1$

end for

V. Simulation

In this section the effectiveness of the proposed safety assessment algorithm is demonstrated with one of the popular CPS examples, an Unmanned Aircraft System (UAS). For the simulation, we focus on the navigation behavior of a UAS and consider the motion of the UAS in two dimensions, referred to as X and Y. The UAS state is represented by the vector $x = \begin{bmatrix} X & \dot{X} & Y & \dot{Y} \end{bmatrix}^{\mathrm{T}}$ that consists of the position and velocity, and the system matrices from (1) are given by:

$$A = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix}, \; B = \begin{bmatrix} T^2/2 & 0 \\ T & 0 \\ 0 & T^2/2 \\ 0 & T \end{bmatrix}$$

where the sampling time T is set to $T = 1$ second for the simulations presented in this section. The position measurements are fed into the state estimator through the following observer matrix:

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

All the disturbance and noise terms are assumed to be zero mean Gaussian white noise with the covariance as an identity matrix. More details about the UAS model can be found in [8], [21]. The simulations below present the behaviors of a UAS subject to two different attack scenarios: controller attack and sensor attack. For each case we set a different attack matrix B_c and B_o, and inject an attack sequence capable of impacting the UAS trajectory. The result of each attack scenario is averaged over $1,000$ Monte Carlo simulations to which our safety assessment algorithm uses the identical safe region, a circle of radius 40 about the current position estimate. In the reachable set computation, the attack detection threshold value is set to $h = 3$, and the computed reachable set are shown with the safety confidence level $\alpha = 13.28$ corresponding to the four degree of freedom χ-squared quantile function with a probability of 99% that the actual state is within the reachable set.

The considered attack scenario allows the attacker to compromise the sensor measurements using $B_o = \begin{bmatrix} 0.5 & 1 \end{bmatrix}^{\mathrm{T}}$, with $B_c = \mathbf{0}$. In the simulation the attacker specifically injects a linearly increasing false data over time: $a_o(k) = k$. Then, the actual state deviates from the estimated trajectory as soon as the attack begins and the computed reachable set boundary continues to grow with time and eventually goes over the boundary of the safe region by time step 50, as shown in Figure 2. When this happens there is a possibility that a stealthy attack (i.e., residual power under the detection threshold h) has driven the UAS to an unsafe state, which is what happens in this scenario. Figure 3 clearly shows that this sensor attack is undetectable, with the residual power holding steady below the threshold while $h = 3$. Despite this fact, the attack still drives the system to an unsafe state. As can be seen, the actual state is outside the safe region and inside the reachable set, so the undetectable sensor attack has successfully driven the UAS to an unsafe state. Our safety assessment algorithm detects the potential loss of safety as soon as the reachable set has any part which extends beyond the safe region, so the system would be aware of this possibility despite the fact that the attack is undetected.

VI. Conclusions

This paper has investigated the safety implications of cyber attacks on Cyber Physical Systems (CPSs) from a controls

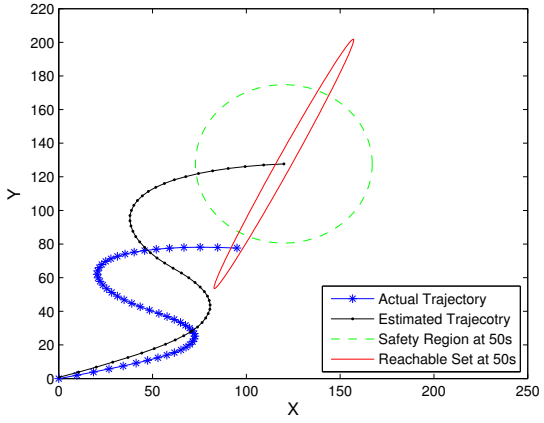

Fig. 2: Trajectory, safe region, and reachable set for controller attack at time steps 50.

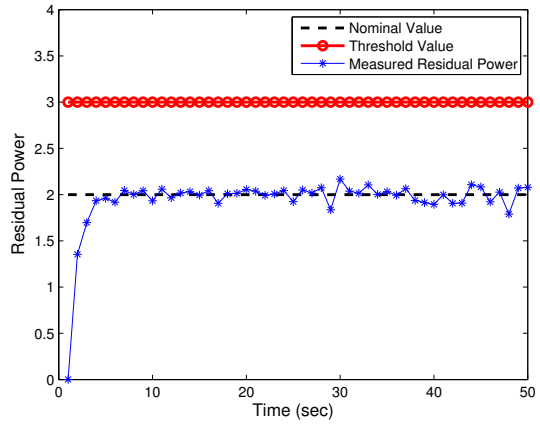

Fig. 3: Residual power statistics under sensor attack.

domain perspective. Focusing on the safety of a CPS subject to stealthy cyber attacks, we have proposed an on-line safety assessment algorithm based on a reachability analysis. The basic idea is to compute the reachable set of the CPS state by taking into consideration all possible stealthy cyber attacks, and compare this with the predefined safe region to determine the CPS's safety status. Our major contribution lies in the computationally demanding reachable set computation, for which we have analytically derived an exact reachable set solution and developed a corresponding recursive algorithm. This allows more reliable safety assessment for the real-time CPS operation, whereas similar research has only considered an over-approximated reachable set. Illustrative simulation examples with different attack scenarios have demonstrated the effectiveness of the proposed algorithm. Clearly, the point at which such attacks become dangerous, i.e. drive the CPS state estimation error outside the safe region, has been shown to be detected by our safety assessment algorithm.

REFERENCES

[1] J. Saltzer and M. Schroeder, "The protection of information in computer systems," in *Proceedings of the IEEE*, vol. 63, no. 9, Sep. 1975, pp. 1278–1308.

[2] A. Avizenis, J. Laprie, B. Randell, and C. Landwehr, "Basic concepts and taxonomy of dependable and secure computing," *IEEE Transactions on Dependable and Secure Computing*, vol. 1, no. 1, pp. 11–32, 2004.

[3] A. Cardenas, S. Amin, and S. Sastry, "Research challenges for the security of control systems," in *3rd USENIX Workshop on Hot topics in security*, Jul. 2008, p. Article 6.

[4] ——, "Secure control: Towards survivable cyber-physical systems," in *The 28th International Conference on Distributed Computing Systems Workshop*, Jun. 2008, pp. 495–500.

[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, Nov. 2013.

[6] C. Ten, G. Manimaran, and C. Liu, "Cybersecurity for critical infrastructures: Attack and defense modeling," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 40, no. 4, pp. 853–865, Jul. 2010.

[7] M. Pajic, J. Weimer, N. Bezzo, P. Tabuada, O. Sokolsky, I. Lee, and G. Pappas, "Robustness of attack-resilient state estimators," in *5th International Conference on Cyber-Physical Systems*, Apr. 2014, pp. 163–174.

[8] J. Goppert, A. Shull, N. Sathyamooethy, W. Liu, V. Sciandra, I. Hwang, and H. Aldridge, "Hardware/software-in-the-loop analysis of cyberattacks on unmanned aerial systems," *AIAA Journal of Aerospace Information System*, vol. 11, no. 5, pp. 337–343, 2014.

[9] J. Goppert, W. Liu, A. Shull, V. Sciandra, I. Hwang, and H. Aldridge, "Numerical analysis of cyber attacks on unmanned aerial systems," in *AIAA Conference on Infotech@Aerospace*, Jun. 2012.

[10] C. Kwon and I. Hwang, "Analytical analysis of cyber attacks on unmanned aerial systems," in *AIAA Conference on Guidance, Navigation, and Control*, Aug. 2013.

[11] C. Kwon, W. Liu, and I. Hwang, "Security analysis for cyber-physical systems against stealthy deception attacks," in *AACC American Control Conference*, Jun. 2013.

[12] ——, "Analysis and design of stealthy cyber attacks on unmanned aerial systems," *AIAA Journal of Aerospace Information Systems*, vol. 11, no. 8, pp. 525–539, Aug. 2014.

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

[14] Y. Mo and B. Sinopoli, "Integrity attacks on cyber-physical systems," in *1st international conference on High Confidence Networked Systems*, Apr. 2012, pp. 47–54.

[15] D. P. Malladi and J. L. Speyer, "A generalized shiryayev sequential probability ratio test for change detection and isolation," *IEEE Transactions on Automatic Control*, vol. 44, no. 8, pp. 1522–1534, Aug. 1999.

[16] I. V. Nikiforov, "A generalized change detection problem," *IEEE Transactions on Information Theory*, vol. 41, no. 1, pp. 171–187, 1995.

[17] D. Dionne, Y. Oshman, and D. Shinar, "Novel adaptive generalized likelihood ratio detector with application to maneuvering target tracking," *AIAA Journal of Guidance, Control, and Dynamics,*, vol. 29, no. 2, pp. 465–474, Mar. 2006.

[18] J. J. Gertler, "Survey of model-based failure detection and isolation in complex plants," *IFAC Proceedings Series*, vol. 7, 1987.

[19] M. Slotani, "Tolerance regions for a multivariate normal population," *Annals of the Institute of Statistical Mathematics*, vol. 16, no. 1, pp. 135–153, Dec. 1964.

[20] W. W. Hager, "Updating the inverse of a matrix," *SIAM Review*, vol. 31, no. 2, pp. 221–239, Jun. 1989.

[21] C. Kwon, S. Yantek, and I. Hwang, "Safety assessment of unmanned aerial systems subject to stealthy cyber attacks," *AIAA Journal of Aerospace Information Systems*, vol. 13, no. 1, pp. 27–45, Jan. 2016.