Cyber Attack and Machine Induced Fault Detection and Isolation Methodologies for Cyber-Physical Systems

Mahdi Taheri, Khashayar Khorasani, Iman Shames, and Nader Meskin

Abstract— In this paper, the problem of simultaneous cyber attack and fault detection and isolation (CAFDI) in cyber-physical systems (CPS) is studied. The proposed solution methodology consists of two filters on the plant and the command and control (C&C) sides of the CPS and an unknown input observer (UIO) based detector on the plant side. Conditions under which the proposed methodology can detect deception attacks, such as covert attacks, zero dynamics attacks, and replay attacks are characterized. An advantage of the proposed methodology is that one does not require a fully secured communication link which implies that the communication link can be compromised by the adversary while it is used to transmit the C&C side observer estimates. Also, it is assumed that adversaries have access to parameters of the system, filters, and the UIO-based detector, however, they do not have access to all the communication link channels. Conditions under which, using the communication link cyber attacks, the adversary cannot eliminate the impact of actuator and sensor cyber attacks are investigated. To illustrate the capabilities and effectiveness of the proposed CAFDI methodologies, simulation case studies are provided and comparisons with detection methods that are available in the literature are included to demonstrate the advantages and benefits of our proposed solutions.

I. INTRODUCTION

Cyber-physical systems (CPS) are monitored and controlled by distributed sensors, actuators, and embedded computers that are connected via communication networks [1]. Our today's life massively depends on CPS due to their wide range of applications in different areas, such as power systems and smart grid, next generation aerospace and transportation systems, and process control and water treatment networks [2]. Through employing CPS for these applications provide us with unique capabilities to accomplish high level performance and reliability performing complex tasks [3].

Anomalies and machine induced faults as well as malicious cyber attacks in physical components of CPS do occur and are observed in actuators and sensors. In recent years, cyber security challenges in CPS, that include cyber attacks on communication networks have attracted significant interest [2]–[7]. Nevertheless, the problem of *simultaneous* diagnosis of cyber attacks and faults has not been fully addressed in the literature.

Mahdi Taheri (m_eri@encs.concordia.ca) and Khashayar Khorasani (kash@ece.concordia.ca) are with the Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada.

Iman Shames (iman.shames@unimelb.edu.au) is with the Department of Electrical and Electronic Engineering, University of Melbourne, Melbourne, Australia.

Nader Meskin (nader.meskin@qu.edu.qa) is with the Department of Electrical Engineering, Qatar University, Doha, Qatar.

A special type of cyber attack is defined as the deception attack in which an adversary changes the transmitted information of the system's input or output by compromising the CPS network communication channels. This paper studies the cyber attack and fault detection and isolation (CAFDI) problem of CPS in presence of machine induced faults as well as malicious deception cyber attacks, such as covert attacks, zero dynamics attacks, and replay attacks. Covert attacks and zero dynamics attacks are defined as undetectable attacks [8]–[10], since they have no impact on the received output measurements on the command and control (C&C) side of the CPS.

A number of researchers have attempted to directly apply fault detection methods to detect cyber attacks, however, there is an inherent difference between machine induced faults and cyber attacks anomalies. Faults represent structural physical anomalies in the system, whereas cyber attacks are injected intentionally by an intelligent adversary with the purpose of damaging the nominal behavior of the system. Standard fault detection algorithms, such as unknown input observer (UIO) [11], have been used as tools to detect cyber attacks. There is an inherent differences between faults and cyber attacks, where faults follow and are governed by laws of physics and are associated with physical system properties. On the other hand, cyber attacks are intelligently designed and do not necessarily follow physical system degradations. Consequently, conventional fault diagnosis algorithms should be fundamentally generalized to accommodate the malicious intelligent adversary cyber attacks threats.

As a brief overview, the geometric-based fault detection methodologies were proposed in [12], [13] to obtain necessary and sufficient conditions for existence of observers that can be used to generate a residual signal for the purpose of fault detection and isolation (FDI). In addition to geometric approaches, many algebraic model-based FDI methods have been introduced, such as UIO [14], [15], interacting multiple model [16], multiple model [17], [18], distributed detection algorithms [19]–[21], and parity equation based approaches [22], [23].

For the cyber attack detection problem, a periodic modulation scheme with the idea of changing the behavior of the control input was proposed in [24] to detect covert and zero dynamics attacks in CPS. However, by using this method a fault in the system can misleadingly be detected as a cyber attack. A method to detect covert attacks in a network of interconnected subsystems using the received information from subsystems was introduced in [25]. However, it was assumed that the communication links among the subsystems are *fully secured*, which is not always feasible in real-world systems.

In [26] geometric theory was used to define zero dynamics attacks and show their impact on the system, and proposed to add perturbations to the system matrices of the system (A, B, C) to change the zero dynamics of the system so that the adversary can no longer excite these new zero dynamics modes. However, in a zero dynamics cyber attack, the adversary has a complete knowledge of the system,

therefore, after changing the characteristics of the system one would still be able to discover the new matrices and dynamics.

In [27], a sensor coding method was proposed that reveals stealthy false data injection attacks by changing the direction of cyber attacks where an algorithm to compute the coding matrices was designed, and finally, a time-varying coding approach was developed for the case when the adversary is capable of estimating a static coding matrix. As a drawback of this approach, it should be noted that one is also not capable of isolating faults and cyber attack signals and anomalies.

The authors in [28] developed a moving target approach in which certain time-varying external dynamics are added to the system. Leveraging the moving target approach, the extended dynamics of the system become unknown to adversaries and they no longer are capable of executing covert attacks and replay attacks. However, zero dynamics attacks cannot be detected by using the moving target approach. In [29], the system was augmented by adding switching auxiliary dynamics that are unknown to the adversary and a switched Luenberger observer was designed to detect covert and zero dynamics attacks, however, for implementation purposes the extended system and the switched observer need to be synchronized.

Due to stealthiness of covert and zero dynamics attacks, it is of paramount importance to develop methods that can be used to detect and isolate them. In addition, due to existence of physical component faults in CPS, one needs to also clearly detect and isolate both faults and cyber attacks in these systems. This paper aims at addressing the problem of CAFDI in CPS.

In our proposed methodology, two filters are designed on both the plant side and the C&C side of the CPS that are interconnected via communication links that can be compromised by the adversary. Moreover, on the plant side UIO-based detectors are designed to generate residuals for detecting and isolating actuator cyber attacks, sensor cyber attacks, as well as actuator faults, and sensor faults while the adversary have a complete knowledge of the filters and UIO-based detectors. Any type of detectable and undetectable cyber attacks can be detected by using our proposed methodology, however, we have assumed that the adversary does not have access to all the communication channels among the filters.

By utilizing both the filters and detectors, we propose and derive conditions under which an adversary that performs cyber attack on the communication link channels cannot eliminate the impacts of actuator and sensor attacks.

To summarize, the main contributions of this paper are stated as follows:

1) A distributed filter design methodology based on observing the system from both the plant side and the C&C side is introduced and developed that can be utilized to detect and isolate both cyber attacks and machine induced faults.
2) By utilizing our proposed methodology, undetectable cyber attacks such as covert attacks and zero dynamics attacks, as well as detectable attacks such as replay attacks can be detected and isolated.
3) Based on both the plant side and the C&C side estimation and observation methodology, conditions under which isolation among actuator cyber attacks and sensor cyber attacks are provided and developed.

The remainder of the paper is organized as follows. A mathematical model of the system that takes into account faults and cyber attacks, the definition of undetectable attacks, and the main objective of this paper are provided in Section II. In Section III, our proposed CAFDI methodology that consists of two side filters, the UIO-based detector and residual signals are developed and investigated. Design conditions for the filters and detector are proposed and developed. To illustrate and demonstrate the capabilities of our analytical results, numerical simulation case studies are presented in Section IV. Conclusions are provided in Section V.

II. PROBLEM STATEMENT AND FORMULATION

A. The Cyber-Physical System (CPS) Model

In this paper, a strictly proper linear time-invariant (LTI) CPS of the form given below is studied:

$$\dot{x}^{\mathrm{s}}(t) = A^{\mathrm{s}}x^{\mathrm{s}}(t) + B^{\mathrm{s}}u^*(t) + L_1 f_1(t) + N^{\mathrm{s}}\omega^{\mathrm{s}}(t),$$
$$y_{\mathrm{p}}(t) = C^{\mathrm{s}}x^{\mathrm{s}}(t) + L_2 f_2^{\mathrm{s}}(t) + \nu^{\mathrm{s}}(t), \qquad (1)$$

where $x^{\mathrm{s}}(t) \in \mathbb{R}^n$ represents the state, $y_{\mathrm{p}}(t) \in \mathbb{R}^p$ denotes the measured output on the plant side, $u^*(t) \in \mathbb{R}^m$ denotes the control input, $f_1(t) \in \mathbb{R}^{m_{\mathrm{f}}}$ and $f_2^{\mathrm{s}}(t) \in \mathbb{R}^{p_{\mathrm{f}}}$ correspond to actuator and sensor faults, respectively. Moreover, $\omega^{\mathrm{s}}(t) \in \mathbb{R}^m$ and $\nu^{\mathrm{s}}(t) \in \mathbb{R}^p$ denote zero mean wide-sense stationary (WSS) random Gaussian processes that represent process and measurement noise with the covariance matrices Q and R, respectively. The quadruple $(A^{\mathrm{s}}, C^{\mathrm{s}}, B^{\mathrm{s}}, N^{\mathrm{s}})$ has appropriate dimensions and describe the CPS characteristics, and the known pair (L_1, L_2) capture the fault signatures.

In case of injection of a cyber attack on actuators, the control input is expressed and changed to

$$u^*(t) = u(t) + S_{\mathrm{a}} a_{\mathrm{u}}(t), \qquad (2)$$

where $u(t) \in \mathbb{R}^m$ represents the control command which is the output of the C&C, $a_{\mathrm{u}}(t) \in \mathbb{R}^{m_{\mathrm{a}}}$ denotes a vector describing the effects of unknown cyber attacks on actuators, and S_{a} is a matrix of appropriate dimension which indicates the control input channels that are under attack.

The output of the CPS on the C&C side when sensors are under cyber attack can be expressed as

$$y^*(t) = C^{\mathrm{s}}x^{\mathrm{s}}(t) + L_2 f_2^{\mathrm{s}}(t) + D_{\mathrm{a}} a_y(t) + \nu^{\mathrm{s}}(t), \qquad (3)$$

where $y^*(t) \in \mathbb{R}^p$ denotes the output, $a_y(t) \in \mathbb{R}^{p_{\mathrm{a}}}$ denotes the attack signal, and the known matrix D_a describes the sensor attack signature. A CPS in presence of both the actuator and sensor cyber attacks is depicted in Fig. 1.

Equations (1) and (2) provide a state space realization of the CPS from the C&C side in the following form:

$$\dot{x}^{\mathrm{s}}(t) = A^{\mathrm{s}}x^{\mathrm{s}}(t) + B^{\mathrm{s}}u(t) + B_{\mathrm{a}}^{\mathrm{s}}a_{\mathrm{u}}(t) + L_1 f_1(t) + N^{\mathrm{s}}\omega^{\mathrm{s}}(t), \qquad (4)$$

where $B_{\mathrm{a}}^{\mathrm{s}} = B^{\mathrm{s}}S_{\mathrm{a}}$ is to be interpreted as the actuator cyber attack signature.

In (2) and (3), $a_{\mathrm{u}}(t)$ and $a_{\mathrm{y}}(t)$ denote the impacts of the adversary's attack on the control input and output of the CPS, respectively. The signals $a_{\mathrm{u}}(t)$ and $a_{\mathrm{y}}(t)$ can be arbitrarily changed by the malicious adversary. In presence of $a_{\mathrm{u}}(t)$ and $a_{\mathrm{y}}(t)$, the adversary intends to inflict maximum possible damage on the components of the system while simultaneously remaining undetected. The following definitions are needed in the remainder of the paper.

Definition 1 (Weakly Unobservable Subspace [30]): Let us denote the CPS by $\Sigma = (A^{\mathrm{s}}, B^{\mathrm{s}}, B_a^{\mathrm{s}}, L_1, N^{\mathrm{s}}, C^{\mathrm{s}}, L_2, D_{\mathrm{a}})$. Under the fault free scenario $f_1(t) = 0$ and $f_2^{\mathrm{s}}(t) = 0$, the noise free scenario $\omega^{\mathrm{s}}(t) = 0$ and $\nu^{\mathrm{s}}(t) = 0$, and the
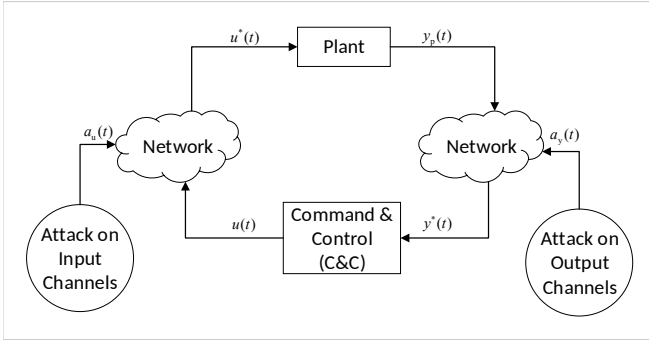
Fig. 1. Cyber-physical system under deception attack on both input and output channels, where $u(t)$ denotes the control command, $a_u(t)$ represents the cyber attack signal on the input channel, $u^*(t)$ represents the control input of the plant, $y_p(t)$ denotes the output on the plant side, $a_y(t)$ denotes the attack signal on the output channel, and $y^*(t)$ denotes the output on the C&C side.
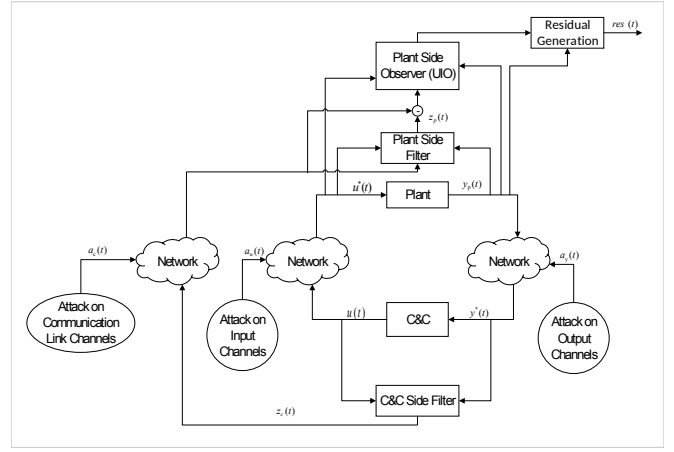

Fig. 2. Observers/filters on both the plant side and the C&C side of the CPS, where $z_c(t)$ represents the states of the C&C side filter, $z_p(t)$ denotes the states of the plant side filter, $a_c(t)$ denotes the cyber attack on the communication link channels, and $res(t)$ denotes the residual signals that are generated on the plant side.

cyber attack free scenario $a_u(t) = 0$ and $a_y(t) = 0$, a point $x^s(0) = x_0^s \in \mathbb{R}^n$ is called weakly unobservable if there exists an input function $u(t)$ such that the output satisfies $y^*(t) = 0$, $\forall t \geq 0$. The set of all weakly unobservable points is called weakly unobservable subspace and is denoted by $\mathcal{V}(\Sigma)$.

Let us denote $X^s(x^s(0), u(t), a_u(t), a_y(t))$ as the solution to (4) under the fault free condition, and $Y(x^s(0), u(t), a_u(t), a_y(t)) = C^s X^s(x^s(0), u(t), a_u(t), a_y(t))$ as the corresponding output of the CPS, $\forall t \geq 0$.

Definition 2 (Undetectable Cyber Attacks [9]): Given $x^s(0) = x_0^s$, in the CPS (4) under the fault free scenario, the cyber attack on actuators and sensors using $a_u(t) \neq 0$ and $a_y(t)$, is designated as undetectable if $Y(x_0^s, u(t), a_u(t), a_y(t)) = Y(x_0^s, u(t), 0, 0)$, $\forall t \geq 0$.

In the same manner as described in [13], [31], the sensor fault and sensor noise can be represented by pseudo actuator fault and pseudo process noise, respectively. It is worth noting that in this representation, as described below, sensor faults are mapped into and represented by pseudo actuator faults.

Towards the above end, the following auxiliary invertible LTI system that is driven by the appropriate $f_2(t)$, which represents the pseudo actuator fault, and $\omega^a(t)$, which captures the pseudo process noise, is expressed as:

$$\dot{x}^a(t) = A^a x^a(t) + L_2^a f_2(t) + N^a \omega^a(t), \\ C^a x^a(t) = L_2 f_2^s(t) + \nu^s(t), \tag{5}$$

where $x^a(t) \in \mathbb{R}^{p_f + p}$, $f_2(t) \in \mathbb{R}^{p_f}$, and $\omega^a(t) \in \mathbb{R}^p$. By incorporating the dynamics of (4) and (5), one can obtain the augmented and extended CPS in the following form:

$$\dot{x}(t) = Ax(t) + Bu(t) + B_a a_u(t) + F_1 f_1(t) + F_2 f_2(t) \\ + N\omega(t), \\ y^*(t) = Cx(t) + D_a a_y(t), \tag{6}$$

where $x(t) = [x^s(t)^\top, x^a(t)^\top]^\top$, $A = \text{diag}(A^s, A^a)$, $B = [B^{s\top}, 0_{m \times (p_f + p)}]^\top$, $B_a = [B_a^{s\top}, 0_{m_a \times (p_f + p)}]^\top$, $F_1 = [L_1^\top, 0_{m_f \times (p_f + p)}]^\top$, $F_2 = [0_{p_f \times n}, L_2^{a\top}]^\top$, $N = \text{diag}(N^s, N^a)$, $\omega(t) = [\omega^s(t)^\top, \omega^a(t)^\top]^\top$, and $C = [C^s, C^a]$. It should be noted that the defined output $y^*(t)$ in (3) is equal to the one that is given by (6), however, the representations are different.

B. Objectives

Our main objective in this paper is to address the simultaneous cyber attack and fault detection and isolation (CAFDI) problem for the CPS (6) by designing a bank of observers such that each set of residual signals corresponding to observers is sensitive and specified to detect one specific type of anomaly, namely either an actuator cyber attack $a_u(t)$, a sensor cyber attack $a_y(t)$, an actuator fault $f_1(t)$, and/or a pseudo actuator fault $f_2(t)$, while each residual is decoupled from all the other anomalies.

Decoupling the residuals from one another implies that occurrence of anomalies only affect those residual signals that are designated to them. We also do not limit our focus to detecting only detectable attacks, such as replay attacks. Our goal and objective is to further detect the so-called undetectable cyber attacks in sense of Definition 2, namely cyber attacks such as covert and zero dynamics. To accomplish our objectives we assume that the adversary cannot compromise all the communication channels among the proposed plant side and C&C side filters, although they have a complete knowledge of parameters of the filters and detectors.

III. PROPOSED METHODOLOGY

The presence of network layer in CPS has enabled malicious adversaries to perform cyber attacks on the entire system. On the other hand, due to existence of this network layer, it is possible to observe the CPS from both the plant side and its C&C side. The idea of observing the CPS from both the plant side and the C&C side is illustrated in Fig. 2. Our goal in this framework is to utilize information from the designed filters on both sides via a communication link and generate residuals that are specifically sensitive to faults and cyber attacks. Using these residuals, the isolation between faults or cyber attacks can also be achieved.

Two filters having the same characteristics on both sides are designed in Subsections III-A and III-B. By using the communication link, states of the C&C side filter are transmitted to the plant side to generate residual signal that is sensitive to only cyber attacks while this communication link may still be compromised by an adversary.

A detector on the plant side that utilizes an unknown input observer (UIO) is designed in the Subsection III-C. The detector utilizes the previously generated residuals as

additional input so that they are sensitive to both cyber attacks and faults. The reason for selecting UIO as the main detector is that it enables one to utilize a general design structure to simultaneously address the considered CAFDI problems.

Other algebraic-based observer design techniques, such as eigenstructure assignment and Kalman filters have certain limitations such as not having a flexible structure and requiring high computational cost. For instance, to isolate different types of cyber attacks and faults using Kalman filters, one needs to design and associate a large number of multiple models of Kalman filters on both sides of the CPS, which is computationally excessive and increases the risks and vulnerabilities exploited by intelligent malicious adversaries to inject cyber attacks.

Our proposed methodology is presented in the Subsection III-D. It is worth noting that by utilizing the proposed methodology, one is still capable of detecting any kind of stealthy cyber attacks on the system, such as covert attacks and zero dynamics attacks.

A. Command & Control side filter

From the C&C side and according to (6), the output of the CPS is governed by

$$y^*(t) = Cx(t) + D_a a_y(t). \quad (7)$$

We have the following standing assumption to be considered throughout this paper.

Assumption 1: Only the communication channels can be compromised and attacked. Consequently, on the C&C side one has access to the control signal, $u(t)$, before its manipulation by the adversary.

The proposed filter on the C&C side can be expressed as follows:

$$\dot{z}_c^\ell(t) = F_p^\ell z_c^\ell(t) + T_p Bu(t) + K_p^\ell y^*(t), \quad (8)$$

where $z_c^\ell(t) \in \mathbb{R}^n$ represents the filter state that estimates $x^s(t)$ from the C&C side, and the matrices F_p^ℓ, T_p^ℓ, and K_p^ℓ are of appropriate dimensions that are designed and selected subsequently. The index $\ell \in \{SA, AA, SF, AF\}$, designates if the filter is designed for detecting sensor attacks, actuator attacks, sensor faults, and actuator faults, respectively.

B. Plant side filter

On the plant side, sensor measurements are carried out before sensor attacks, and the output of CPS can be expressed as follows:

$$y_p(t) = Cx(t).$$

Moreover, on this side one has access to the potentially manipulated control signal $u^*(t) = u(t) + S_a a_u(t)$.

The proposed filter on the plant side is expressed in the following form:

$$\dot{z}_p^\ell(t) = F_p^\ell z_p^\ell(t) + T_p^\ell Bu^*(t) + K_p^\ell y_p(t) + L_p^\ell(z_p^\ell(t) - (z_c^\ell(t) + D_{ac}a_c(t))), \quad (9)$$

where $z_p^\ell(t) \in \mathbb{R}^n$ denotes the filter state estimating $x^s(t)$ from the plant side, $a_c(t) \in \mathbb{R}^{n_c}$ denotes the cyber attack on the communication link between the two filters with the signature D_{ac}. Similar to the C&C side filters, the index $\ell \in \{SA, AA, SF, AF\}$, indicates if the filter is designed for detecting sensor attacks, actuator attacks, sensor faults, and actuator faults, respectively.

The error signals between estimated states for both sides can be defined as $e_p^\ell(t) = z_p^\ell(t) - z_c^\ell(t)$. The state-space representation of the error dynamics between the two filter states can be derived as follows:

$$\dot{e}_p^\ell(t) = (F_p^\ell + L_p^\ell)e_p^\ell(t) + T_p^\ell B_a a_u(t) - K_p^\ell D_a a_y(t) - L_p^\ell D_{ac}a_c(t). \quad (10)$$

It follows from (10) that the error dynamics is only sensitive to cyber attacks.

C. UIO-based detector and residual signal generation

Consider a UIO-based detector on the plant side having the following representation:

$$\dot{z}^\ell(t) = F^\ell z^\ell(t) + T^\ell Bu^*(t) + K^\ell y_p(t) + L^\ell(z_p^\ell(t) - (z_c^\ell(t) + D_{ac}a_c(t))), \quad (11)$$
$$\hat{x}^\ell(t) = z(t)^\ell + H^\ell y_p(t),$$

where $z^\ell(t) \in \mathbb{R}^{(n+p_f+p)}$, and $\hat{x}(t) \in \mathbb{R}^{(n+p_f+p)}$ denotes the estimated states by the detector. The matrices F^ℓ, T^ℓ, K^ℓ, L^ℓ, and H^ℓ are of appropriate dimensions and will be specified subsequently, with $\ell \in \{SA, AA, SF, AF\}$, denoting the categories defined previously.

The error between the states of the detector and the CPS is defined as $e^\ell(t) = x(t) - \hat{x}^\ell(t)$. Let

$$res_\ell(t) = y_p(t) - C\hat{x}^\ell(t) = Ce^\ell(t), \quad (12)$$

denote a residual signal. By selecting $K^\ell = K_1^\ell + K_2^\ell$, $F^\ell = A - H^\ell CA - K_1^\ell C$, K_1^ℓ of appropriate dimension, and $K_2^\ell = FH^\ell$, the dynamics associated with $e^\ell(t)$ can now be expressed in the following form:

$$\dot{e}^\ell(t) = (A - H^\ell CA - K_1^\ell C)e^\ell(t) + (I - T^\ell - H^\ell C)(Bu(t) + B_a a_u(t)) + (I - H^\ell C)F_1 f_1(t) + (I - H^\ell C)F_2 f_2(t) + (I - H^\ell C)N\omega(t) - L^\ell e_p^\ell(t) - L^\ell D_{ac}a_c(t). \quad (13)$$

Definition 3: A cyber attack/fault is detected if the residual signal $res_\ell(t)$ given by (12) exceeds a pre-specified threshold $\eta > 0$ as follows:

$$\|res_\ell(t)\|_2 > \eta.$$

where $\|.\|_2$ indicates the Euclidean norm.

Remark 1: To select the threshold η, one may need to perform Monte Carlo simulation runs for the healthy system, i.e., for the fault free and cyber attack free system in presence of external disturbances and noise and choose the maximum value of $\|res(t)_\ell\|_2$ as η.

Definition 4 (Decoupled Residual): The residual signal $res_\ell(t)$ given by (12) is decoupled from an anomalous signal in the set $\{a_u(t), a_y(t), f_1(t), f_2(t)\}$ if the dynamics and trajectory of $res_\ell(t)$ is not affected by that anomalous signal.

D. Filters and detector design for cyber attack and fault detection and isolation objectives

The error dynamics in (10) and (13) can now be augmented as follows:

$$\dot{\check{e}}^\ell(t) = \check{F}^\ell \check{e}^\ell(t) + \check{B}^\ell u(t) + \check{B}_a^\ell a_u(t) + \check{F}_1^\ell f_1(t) + \check{F}_2^\ell f_2(t) - \check{K}_p^\ell a_y(t) - \check{L}^\ell a_c(t) + \check{N}^\ell \omega(t), \quad (14)$$

where $\check{e}^\ell(t) = [e^\ell(t)^\top \; e_p^\ell(t)^\top]^\top$, and

$$\check{F}^\ell = \begin{bmatrix} F^\ell & -L^\ell \\ 0 & F_p^\ell + L_p^\ell \end{bmatrix}, \; \check{B} = \begin{bmatrix} (I - T^\ell - H^\ell C)B \\ 0 \end{bmatrix},$$

$$\check{B}_a^\ell = \begin{bmatrix} (I - T^\ell - H^\ell C)B_a \\ T_p^\ell B_a \end{bmatrix}, \; \check{F}_1^\ell = \begin{bmatrix} (I - H^\ell C)F_1 \\ 0 \end{bmatrix},$$

$$\check{F}_2^\ell = \begin{bmatrix} (I - H^\ell C)F_2 \\ 0 \end{bmatrix}, \; \check{K}_p^\ell = \begin{bmatrix} 0 \\ K_p^\ell D_a \end{bmatrix}, \; \check{L}^\ell = \begin{bmatrix} L^\ell D_{ac} \\ L_p^\ell D_{ac} \end{bmatrix},$$

$$\check{N}^\ell = \begin{bmatrix} (I - H^\ell C)N \\ 0 \end{bmatrix},$$

(15)

where $\ell \in \{\text{SA}, \text{AA}, \text{SF}, \text{AF}\}$.

Assumption 2: The malicious adversary is aware of the parameters of filters in (8), (9), and the UIO-based detector in (11).

Assumption 3: The malicious adversary does not have access to <u>all</u> the communication channels between the two side filters, i.e., $\text{rank}(D_{ac}) < n$.

In the following, it is shown that how one can generate four residual signals $res_{AA}(t)$, $res_{SA}(t)$, $res_{AF}(t)$, and $res_{SA}(t)$ to detect the actuator cyber attack, the sensor cyber attack, the actuator fault, and the sensor fault, respectively, by using a bank of filters and four UIO-based detectors.

Proposition 1: Under Assumption 3, the residual signal $res_{AA}(t) = y_p(t) - C\hat{x}^{AA}(t)$ is affected by the actuator cyber attack $a_u(t)$ and is decoupled from $a_y(t)$, $f_1(t)$, and $f_2(t)$ in the sense of Definition 4, if the following conditions for the augmented dynamics (14) hold for $\ell = $ AA, namely:

1) $T^\ell = I - H^\ell C$;
2) $(I - H^\ell C)F_1 = 0$;
3) $(I - H^\ell C)F_2 = 0$;
4) $L^\ell D_{ac} = 0$;
5) $L_p^\ell D_{ac} = 0$;
6) $K_p^{AA} D_a = 0$;
7) the triplet (C, F^ℓ, L^ℓ) is left-invertible;
8) the Rosenbrock system matrix

$$P_{\Sigma_u}(s) = \begin{bmatrix} sI - (F_p^{AA} + L_p^{AA}) & -T_p^{AA}B_a \\ L^{AA} & 0_{(n+p_f+p) \times m_a} \end{bmatrix},$$

does not have any non-minimum phase zero dynamics;
9) $\text{rank}\,(L^{AA}T_p^{AA}B_a) = \text{rank}\,(T_p^{AA}B_a)$;
10) $\check{F}^\ell$ is Hurwitz.

Proof: The augmented governing error dynamics associated with $e^{AA}(t)$ and $e_p^{AA}(t)$ are governed by (14) where $\ell = $ AA. Under Conditions 1) to 6), the dynamics (14) become

$$\dot{\check{e}}^{AA}(t) = \check{F}^{AA}\check{e}^{AA}(t) + \check{B}_a^{AA}a_u(t) + \check{N}^{AA}\omega(t). \quad (16)$$

Consequently, the error signal $\check{e}(t)$ is not affected by the control command $u(t)$, the actuator fault $f_1(t)$, the sensor fault $f_2(t)$, the sensor attack $a_y(t)$, and the communication link attack signal $a_c(t)$. Furthermore, (16) can be partitioned into the following two subsystems:

$$\dot{e}_p^{AA}(t) = (F_p^{AA} + L_p^{AA})e_p^{AA}(t) + T_p^{AA}B_a a_u(t), \quad (17)$$

and

$$\dot{e}^{AA}(t) = Fe^{AA}(t) - L^{AA}e_p^{AA}(t) + (I - H^{AA}C)N\omega(t),$$
$$res_{AA}(t) = Ce^{AA}(t).$$

(18)

Based on Condition 7) and according to (18), the impact of $e_p^{AA}(t)$ will appear in $res_{AA}(t)$ for any $a_u(t) \neq 0$.

Consider $e_p^{AA}(t)$ in (17) with the output $L^{AA}e_p^{AA}(t)$ in order to construct the Rosenbrock system matrix $P_{\Sigma_u}(s)$. To prevent stealthy attacks on the plant side filter, one needs to design this filter and L^{AA} such that the Rosenbrock system matrix $P_{\Sigma_u}(s)$ has no non-minimum phase zero dynamics and is left-invertible [8].

The Rosenbrock system matrix $P_{\Sigma_u}(s)$ being left-invertible is equivalent to the largest controllability subspace of the system $(L^{AA}, F_p^{AA} + L_p^{AA}, T_p^{AA}B_a)$ contained in $\text{ker}(L^{AA})$, and designated as $\mathscr{R}^*(\Sigma_u)$ being null [30]. One has (refer to Theorem 8.22 in [30] and Theorem 5.6 in [32])

$$\mathscr{R}^*(\Sigma_u) = \mathscr{V}(\Sigma_u) \cap \mathscr{W}^*(\Sigma_u), \quad (19)$$

where $\mathscr{V}(\Sigma_u)$ is the weakly unobservable subspace that is equivalent to the largest output-nulling subspace of the triplet $(L^{AA}, F_p^{AA} + L_p^{AA}, T_p^{AA}B_a)$, and $\mathscr{W}^*(\Sigma_u)$ is the smallest conditioned invariant subspace containing $\text{Im}(T_p^{AA}B_a)$ [9].

As described in [30] and [32], these subspaces can be computed by using the following algorithm

$$\mathscr{V}_0 = \text{Ker}(L^{AA}),$$
$$\mathscr{V}_k = \mathscr{V}_0 \cap F_p^{AA-1}(\mathscr{V}_{k-1} + \text{Im}(T_p^{AA}B_a)), \quad (20)$$

and

$$\mathscr{W}_0 = \text{Im}(T_p^{AA}B_a),$$
$$\mathscr{W}_k = \mathscr{W}_0 + F_p^{AA}(\mathscr{W}_{k-1} \cap \text{Ker}(L^{AA})), \quad (21)$$

where $\mathscr{V}_k$ and $\mathscr{W}_k$ converge to $\mathscr{V}(\Sigma_u)$ and $\mathscr{W}^*(\Sigma_u)$, respectively, in at most $k = n$ steps.

Given (19), $\mathscr{R}^*(\Sigma_u) = 0$, if $\mathscr{V}_0 \cap \mathscr{W}_0 = 0$, or equivalently,

$$\text{Ker}(L^{AA}) \cap \text{Im}(T_p^{AA}B_a) = 0. \quad (22)$$

The equation (22) implies that $\text{Im}(T_p^{AA}B_a)$ should not be in the null space of L^{AA}, which is equivalent to

$$\text{rank}\,(L^{AA}T_p^{AA}B_a) = \text{rank}\,(T_p^{AA}B_a).$$

The Rosenbrock system matrix $P_{\Sigma_u}(s)$ being left-invertible implies that for any $a_u(t) \neq 0$, $L^{AA}e_p^{AA}(t) \neq 0$.

Finally, in order to detect actuator cyber attacks, the governing dynamics in (16) should be stable. This completes the proof of the Proposition 1. ∎

Remark 2: It should be emphasized that as per Assumption 3, there exists a nonzero L^{AA} that satisfies the Condition (4) in the above proposition.

Proposition 2: Under Assumption 3, the residual signal $res_{SA}(t) = y_p(t) - C\hat{x}^{SA}(t)$ is affected by the sensor cyber attacks $a_y(t)$ and is decoupled from $a_u(t)$, $f_1(t)$, and $f_2(t)$ in the sense of Definition 4, if Conditions 1)-5), 7), and 10) of the Proposition 1 for $\ell = $ SA, and the following conditions for the augmented error dynamics (14) hold:

1) $T_p^{SA}B_a = 0$;
2) the Rosenbrock system matrix

$$P_{\Sigma_y}(s) = \begin{bmatrix} sI - (F_p^{SA} + L_p^{SA}) & K_p^{SA}D_a \\ L^{SA} & 0_{(n+p_f+p) \times p_a} \end{bmatrix},$$

does not have any non-minimum phase zero dynamics; and
3) $\text{rank}\,(L^{SA}K_p^{SA}D_a) = \text{rank}\,(K_p^{SA}D_a)$.

Proof: The proof follows along similar lines to that of Proposition 1 and is omitted for sake of brevity. ∎

Remark 3: Suppose Condition (9) of the Proposition 1 is not satisfied and $\check{P}_{\Sigma_u}(s)$ is not left-invertible. In this case, it has been shown in [8] that one can find an actuator cyber attack $a_u(t) \neq 0$ such that $L^{AA}e_p^{AA}(t) = 0$. This type of cyber attack has been represented in [8] and has been defined as "undetectable controllable attack" in [9]. According to (17) and (18) the actuator cyber attack signal $a_u(t)$ can affect the error $e^{AA}(t)$ only through $L^{AA}e_p^{AA}(t)$. Hence, the adversary has the capability of injecting a stealthy cyber attack by using $a_u(t)$ that does not affect the residual signal $res_{AA}(t) = Ce^{AA}(t)$. Similarly, it can be shown that if Condition (3) of Proposition 2 is not satisfied and $\check{P}_{\Sigma_y}(s)$ is not left-invertible, the adversary can inject stealthy attack using $a_y(t)$ which does not affect the residual $res_{SA}(t)$.

Remark 4: In Propositions 1 and 2, there is no assumption on the nature, characteristics, and type of sensor and actuator cyber attacks. This implies that by using the proposed method, one is capable of detecting and isolating detectable attacks, such as replay attacks, as well as undetectable attacks (refer to Definition 2), such as covert attacks and zero dynamics attacks.

Proposition 3: Let ℓ = AF . The residual signal $res_{AF}(t) = y_p(t) - C\hat{x}^{AF}(t)$ is affected by the <u>actuator fault</u> $f_1(t)$ and is decoupled from $a_u(t)$, $a_y(t)$, and $f_2(t)$ in the sense of Definition 4, if $L^{AF} = 0$ and the following conditions hold:
1) $T^{AF} = I - H^{AF}C$;
2) $(I - H^{AF}C)F_2 = 0$;
3) $\check{F}^{AF}$ is Hurwitz.

Proof: In light of Conditions 1) and 2), and setting ℓ = AA, (14) yields

$$\dot{\check{e}}^{AF}(t) = \check{F}^{AF}\check{e}^{AF}(t) + \check{B}_a^{AF}a_u(t) + \check{F}_1^{AF}f_1(t) - \check{K}_p^{AF}a_y(t) - \check{L}^{AF}a_c(t) + \check{N}^{AF}\omega(t).$$

Moreover, by setting $L^{AF} = 0$, the dynamics of $e^{AF}(t)$ is governed by:

$$\dot{e}^{AF}(t) = F^{AF}e^{AF}(t) + (I - H^{AF}C)F_1f_1(t) + N\omega(t).$$

and consequently, the residual signal $res_{AF}(t) = Ce^{AF}(t)$ is only sensitive to the actuator fault $f_1(t)$. In addition, $\check{F}^{AF}$ should be Hurwitz in order to have a stable error dynamics $e^{AF}(t)$. This completes the proof of the Proposition 3. ∎

Proposition 4: The residual signal $res_{SF}(t) = y_p(t) - C\hat{x}^{SF}(t)$ is affected by the <u>pseudo actuator fault</u> $f_2(t)$ and is decoupled from $a_u(t)$, $a_y(t)$, and $f_1(t)$ in the sense of Definition 4, if $L^{SF} = 0$ and the following conditions for the augmented dynamic (14) hold:
1) $T^{SF} = I - H^{SF}C$;
2) $(I - H^{SF}C)F_1 = 0$;
3) $\check{F}^{SF}$ is Hurwitz.

Proof: Setting ℓ = SF, the proof follows along similar lines to that of Proposition 3 and is omitted for sake of brevity. ∎

As stated in [14], Conditions 2) and 3) in Proposition 1 are solvable if and only if $\text{rank}(CF_1) = \text{rank}(F_1)$; and $\text{rank}(CF_2) = \text{rank}(F_2)$. The next lemma provides sufficient conditions for isolability of sensors and actuator faults.

Theorem 1: The residuals $res_{AF}(t)$ and $res_{SF}(t)$ can be simultaneously generated to detect and isolate $f_1(t)$ and $f_2(t)$ if $F_1^\top F_2 = 0$.

Proof: In order to generate the residual signal $res_{AF}(t)$ Condition 2) in Proposition 3 should hold, which can be interpreted as requiring

$$\text{Im}(I - H^{AF}C) \subset \text{Ker}(F_2^\top). \tag{23}$$

and at the same time, the impact of $f_1(t)$ should show up in the dynamics of $e(t)$, that implies $(I - H^{AF}C)F_1 \neq 0$. The latter condition is equivalent to

$$\text{Im}(F_1^\top) \subset \text{Im}(I - H^{AF}C). \tag{24}$$

From (23) and (24), it can be inferred that $\text{Im}(F_1^\top) \subset \text{Ker}(F_2^\top)$, which implies that $F_1^\top F_2 = 0$. Note that the case of generating the residual signal $res_{SF}(t)$ provides one with the same result. This completes the proof of the Theorem 1. ∎

It follows from the definitions of F_1 and F_2 that the condition $F_1^\top F_2 = 0$ is always satisfied. Therefore, as long as Conditions (2) and (3) in Proposition 1 are solvable, the actuator faults and pseudo actuator faults can be detected and isolated.

Remark 5: To generate the residual signals $res_{AA}(t)$, $res_{SA}(t)$, $res_{AF}(t)$, and $res_{SF}(t)$ one needs to construct a bank of eight filters (four on each side) with the states $z_p^{AA}(t)$, $z_c^{AA}(t)$, $z_p^{SA}(t)$, $z_c^{SA}(t)$, $z_p^{AF}(t)$, $z_c^{AF}(t)$, $z_p^{SF}(t)$, and $z_c^{SF}(t)$ and four UIO-based detectors with the states $\hat{x}^{AA}(t)$, $\hat{x}^{SA}(t)$, $\hat{x}^{AF}(t)$, and $\hat{x}^{SF}(t)$ according to Propositions 1-4. In Propositions 1 and 2, the matrices K_p^{AA} and T_p^{SA} have been utilized to decouple sensor cyber attacks and actuator cyber attacks in sense of Definition 4 from the generated residual signals, respectively. Hence, one can conclude that there is no contradiction among the conditions to generate $res_{AA}(t)$ and $res_{SA}(t)$. Subsequently, from Theorem 1 it can be seen that no contradiction exists among the design conditions in the Propositions 3 and 4 to generate $res_{AF}(t)$ and $res_{SF}(t)$. Moreover, in Propositions 3 and 4, the matrix L^ℓ has been employed to decouple the cyber attack signals from $res_{AF}(t)$ and $res_{SF}(t)$, which indicates that there are no contradictions in the design conditions of Propositions 1 and 2.

IV. NUMERICAL CASE STUDIES

In this section, numerical case studies are provided to demonstrate and verify the capabilities and advantages of our proposed methodology as compared to the available results in the literature. For these case studies, a bank of filters and UIO-based detectors are designed to achieve detection and isolation of cyber attacks as well as faults by using the proposed methods in the Propositions 1-4. To simulate the covert and zero dynamics attacks the models in [2] and [1] are used, respectively.

Two types of cyber attacks are studied, namely covert attacks and zero dynamics attacks. Moreover, detection and isolation of <u>simultaneous</u> actuator and sensor bias faults with cyber attacks are also demonstrated and validated. A linear dynamical system with the following characteristic matrices and cyber attack and fault signatures is considered:

$$A^s = \begin{bmatrix} -1 & 0 & 1 & 0 \\ 0 & -3 & 0 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{bmatrix}, B^s = \begin{bmatrix} -2 & -1 \\ 0 & -2 \\ 0 & -3 \\ -4 & 0 \end{bmatrix},$$

$$C^s = \begin{bmatrix} 0.2 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 \end{bmatrix}, B_a^s = \begin{bmatrix} -2 & -1 \\ 0 & -2 \\ 0 & -3 \\ -4 & 0 \end{bmatrix},$$

$$L_1 = \begin{bmatrix} -2 \\ 0 \\ 0 \\ -4 \end{bmatrix}, L_2^a = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, D_{ac} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, N^a = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix},$$

$$A^{\mathrm{a}} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -3 \end{bmatrix}, \ C^{\mathrm{a}} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \ N^{\mathrm{s}} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix},$$

$$D_{\mathrm{a}} = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}, \tag{25}$$

where all the input and output channels are compromised by adversaries as they have access to two out of the four communication channels. The covariance matrices of $\omega^{\mathrm{s}}(t)$ and $\omega^{\mathrm{a}}(t)$ are specified as $Q = \mathrm{diag}(0.01, 0.01, 0.01, 0.01)$ and $R^{\mathrm{a}} = \mathrm{diag}(0.02, 0.02)$, respectively.

For the case studies, the design steps that are summarized in the Algorithms 1 and 2 in Appendix VI are utilized. A bank of plant side filters as given by (9), C&C side filters as presented by (8), and detectors as provided in (11) are designed such that the conditions of Propositions 1-4 are satisfied. Moreover, the residual signals $res_{\mathrm{AA}}(t)$, $res_{\mathrm{SA}}(t)$, $res_{\mathrm{AF}}(t)$, and $res_{\mathrm{SF}}(t)$ are generated according to Propositions 1-4, respectively.

Scenario 1 (Zero Dynamics Attacks): The system presented in (25) has a non-minimum phase zero at $s = 0.3028$, that is associated with the zero state direction $x_0^{\mathrm{s}} = [0, 0, -0.6514, 1]^{\top}$ and the zero input direction $u_0 = [-0.5757, 0.5]^{\top}$. To determine the threshold for the residual signals $res_{\mathrm{AA}}(t)$ and $res_{\mathrm{SA}}(t)$ of the actuator and sensor cyber attacks 100 Monte Carlo simulation runs are conducted according to Remark 1, and the threshold is determined as $\eta = 3.3$. The parameters of the filters and the UIO-based detector subject to actuator cyber attack are designed as follows:

$$F_{\mathrm{p}}^{\mathrm{AA}} = \begin{bmatrix} -3 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & -4 & 0 \\ 0 & 0 & 0 & -5 \end{bmatrix}, \ T_{\mathrm{p}}^{\mathrm{AA}} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 1 \\ 2 & 0 & 0 & -4 \\ 0 & 1 & 0 & 0 \end{bmatrix},$$

$$L_{\mathrm{p}}^{\mathrm{AA}} = \begin{bmatrix} 0 & 0 & 4 & -1 \\ 0 & 0 & 3 & -2 \\ 0 & 0 & 2 & -3 \\ 0 & 0 & 5 & -1 \end{bmatrix}, \ H^{\mathrm{AA}} = \begin{bmatrix} 5 & -5 \\ 0 & 0 \\ 0 & 0 \\ 10 & -10 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

$$K_1^{\mathrm{AA}} = \begin{bmatrix} 6 & -2 \\ -3 & 1 \\ 6 & 2 \\ 3 & 1 \\ 3 & 1 \\ 6 & 2 \\ 3 & 1 \end{bmatrix}, \ L^{\mathrm{AA}} = \begin{bmatrix} 0 & 0 & 4 & -1 \\ 0 & 0 & 3 & -2 \\ 0 & 0 & 2 & -3 \\ 0 & 0 & 5 & -1 \\ 0 & 0 & 3 & -2 \\ 0 & 0 & 2 & -3 \\ 0 & 0 & 5 & -1 \end{bmatrix}, \ K_{\mathrm{p}}^{\mathrm{AA}} = [0]_{4 \times 2},$$

As can be seen in Fig. 3, the residual signal $res_{\mathrm{AA}}(t) = y_{\mathrm{p}}(t) - C\hat{x}^{\mathrm{AA}}(t)$ that is designed to detect actuator cyber attacks has increased (due to a zero dynamics attack) while the other residuals are successfully below the threshold.

Scenario 2 (Covert Attacks): In this scenario, a covert attack scenario is considered. The adversary is capable of completely removing the impact of actuator cyber attack $a_{\mathrm{u}}(t) = [2, 1]^{\top}$ from the sensor measurements by using the sensor cyber attack $D_{\mathrm{a}} a_{\mathrm{y}}(t) = -Cx_{\mathrm{cov}}(t)$, where $\dot{x}_{\mathrm{cov}}(t) = Ax_{\mathrm{cov}}(t) + B_{\mathrm{a}} a_{\mathrm{u}}(t)$ and $x_{\mathrm{cov}}(0) = x(0)$. The impact of this cyber attack at $t = 10$ (s) can be seen on sensor measurements on the plant side as shown in Fig. 4. However, the received sensor measurements on the C&C side do not show any anomaly in outputs. The parameters of the detector are the same as in Scenario 1, but to detect sensor cyber attacks a set of filters are designed to satisfy the conditions that are provided in Proposition 2 to generate $res_{\mathrm{SA}}(t)$.
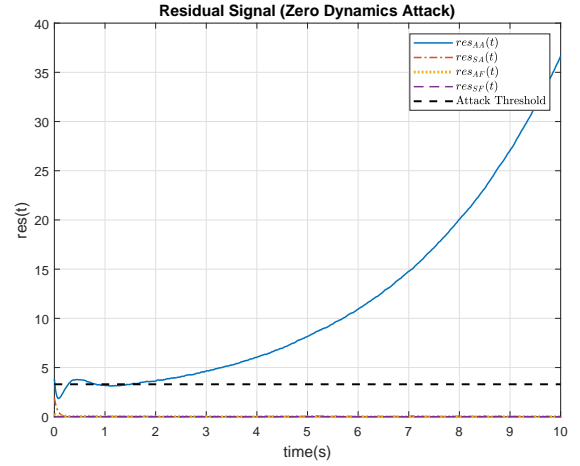

Fig. 3. Detection of a zero dynamics attack that is injected at $t = 0$ (s).
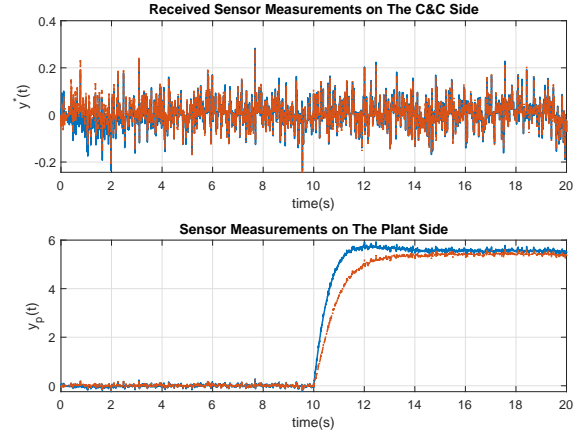

Fig. 4. Difference between output of the system on the plant side and the C&C side due to injection of covert attack at $t = 10$ (s).
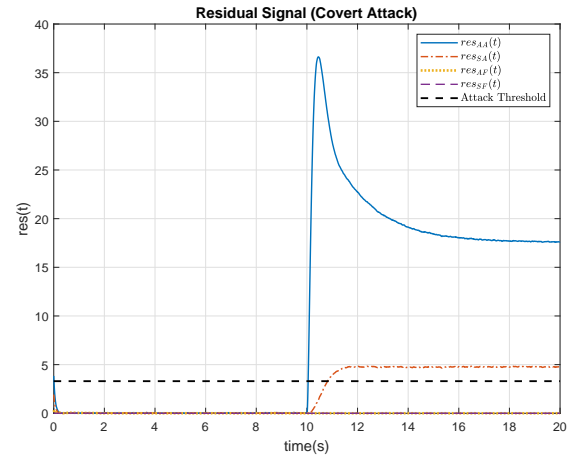

Fig. 5. Detection of actuator and sensor cyber attacks in case of covert attacks.

As shown in Fig. 5, the increase in actuator and sensor cyber attacks residuals, $res_{\mathrm{AA}}(t)$ and $res_{\mathrm{SA}}(t)$, respectively, that exceed the threshold indicate the occurrence of these cyber attacks.
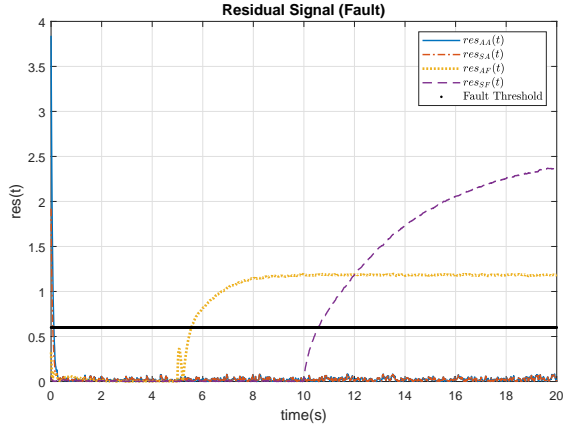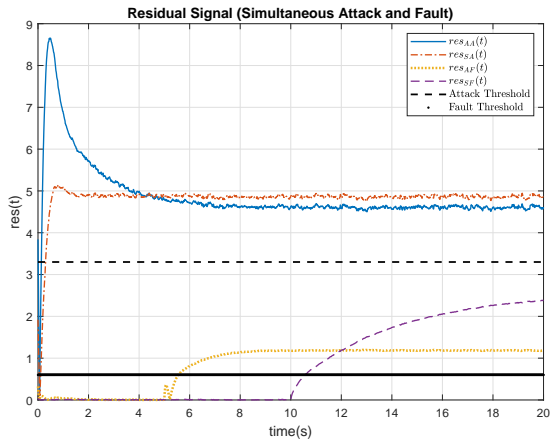
Fig. 6. Detection of actuator and sensor faults.

Fig. 8. Residual signals when Condition (9) of the Proposition 1 is not satisfied.

Fig. 7. Detection and isolation of different simultaneous cyber attacks and faults.

Fig. 9. False detection of cyber attack by using the proposed method in [24] while there is only a fault in the system (actuator fault is injected from $t = 5$ (s) onwards).

Scenario 3 (Faults): Using Proposition 3, the UIO-based detector and its corresponding residual signal $res_{AF}(t)$ that is sensitive to actuator faults are first designed. Then, based on conditions in Proposition 4 to detect sensor faults the matrices for the UIO-based detector and the residual signal $res_{SF}(t)$ are selected. The threshold for residuals that are used to detect actuator and sensor faults is computed according to the method provided in Remark 1 and is set to $\eta = 0.6$. In this scenario, the actuator fault, $f_1(t) = 40$, has occurred at $t = 5$ (s) and the pseudo actuator fault, $f_2(t) = 20$, also exists in the system from $t = 10$ (s) onwards. It can be observed from Fig. 6 that due to occurrence of faults the corresponding residuals have been increased.

Scenario 4 (Simultaneous Injection of Cyber Attack and Fault): In this scenario, the detection and isolation of simultaneous cyber attacks and faults is demonstrated. In this scenario, the system is under a covert attack at $t = 0$ (s) and an actuator fault and sensor faults occur at $t = 5$ (s) and $t = 10$ (s), respectively. As depicted in Fig. 7, these anomalies can be both detected and isolated successfully.

Scenario 5 (Condition (9) of the Proposition 1 is not Satisfied): In this scenario, we have intentionally designed our monitoring system in a manner such that Condition (9) of the Proposition 1 is not satisfied. Therefore, we can illustrate its importance in our proposed methodology. In Fig. 8, it can be seen that when the above condition is not satisfied the adversary is now capable of performing
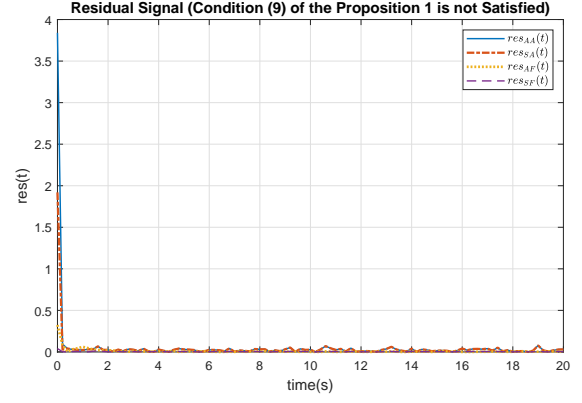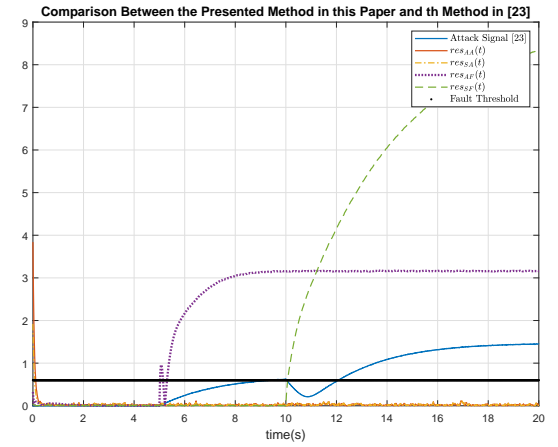
"undetectable controllable attack" (refer to Remark 3 and [9]) on $P_{\Sigma_u}(s)$ and completely eliminate or cancel out impacts of the actuator cyber attack on the residual.

Comparative Study with Results Available in the Literature: In order to provide a comparison with the existing results in the literature, the proposed approach in [24] is applied to our case studies. The following periodic modulation matrix was developed in [24]:

$$
S(k) = \begin{cases} S_1 & ; k = 1 \ (0 \leq t \leq t_1) \\ \vdots \\ S_T & ; k = T \ (t_{T-1} \leq t \leq t_T) \end{cases}
$$

where $S(k)$ is the modulation matrix on the input, $S_1, \ldots, S_T \in \mathbb{R}^{m \times m}$ are constant matrices, and $T = m$. The idea in [24] is to disrupt the knowledge of the adversary from the system by employing the modulation $S(k)$. Using the detection method in [24], it is shown in Fig. 9 that despite having no actuator and sensor cyber attacks, the attack residual signal increases which *misleadingly* indicates the existence of cyber attacks (false positive). However, in the same figure it is shown that by using our proposed method in Propositions 1-4 and generating $res_{AA}(t)$, $res_{SA}(t)$, $res_{AF}(t)$, and $res_{SF}(t)$, the occurrence of actuator fault in the system was correctly detected and isolated.

TABLE I

TPR MEASURE FOR ACTUATOR ATTACK DETECTION ACCORDING TO THE
PROPOSED METHODOLOGY CORRESPONDING TO PROPOSITION 1.

Types of Anomalies	TPR% (Proposition 1)
AA	96%
AA & SA	96%
AA & AF	95%
AA & SF	96%
AA & SA & AF	95%
AA & SA & SF	96%
AA & AF & SF	96%
AA & SA & AF & SF	95%

TABLE II

TPR MEASURE FOR SENSOR ATTACK DETECTION ACCORDING TO THE
PROPOSED METHODOLOGY CORRESPONDING TO PROPOSITION 2.

Types of Anomalies	TPR% (Proposition 2)
SA	99%
SA & AA	99%
SA & AF	99%
SA & SF	99%
SA & AA & AF	99%
SA & AA & SF	99%
SA & AF & SF	99%
SA & AA & AF & SF	90%

TABLE III

TPR MEASURE FOR ACTUATOR FAULT DETECTION ACCORDING TO THE
PROPOSED METHODOLOGY IN PROPOSITION 3.

Types of Anomalies	TPR% (Proposition 3)
AF	93%
AF & AA	93%
AF & SA	93%
AF & SF	93%
AF & AA & SA	93%
AF & AA & SF	93%
AF & SA & SF	92%
AF & AA & SA & SF	93%

TABLE IV

TPR MEASURE FOR SENSOR FAULT DETECTION ACCORDING TO THE
PROPOSED METHODOLOGY IN PROPOSITION 4.

Types of Anomalies	TPR% (Proposition 4)
SF	96%
SF & AA	96%
SF & SA	96%
SF & AF	96%
SF & AA & SA	96%
SF & AA & AF	96%
SF & SA & AF	96%
SF & AA & SA & AF	96%

A. Quantitative performance evaluation

Our proposed CAFDI methodologies under different noise levels are quantitatively evaluated through 100 different Monte Carlo simulation runs. A confusion matrix [33] is employed to evaluate the performance of our proposed methods. Given a classifier and its corresponding instances, four possible outcomes are specified as (1) TP (True Positive), if the instance is positive and is truly classified as positive, (2) FN (False Negative), if the instance is positive and incorrectly classified as negative, (3) TN (True Negative), if the instance is negative and correctly classified as negative, and (4) FP (False Positive), if the instance is negative and incorrectly classified as positive [33].

Based on the possible outcomes the metric true positive rate (TPR) which indicates the rate of correct detection is used as a performance measure in this paper. This performance measure can be computed by using the expression $TPR = TP/(TP + FN)$. In this subsection, "AA", "SA", "AF", and "SF" are used to denote actuator attack, sensor attack, actuator fault, and sensor fault, respectively. The TPR results for the proposed methods that are developed in Propositions 1-4 for injection of cyber attacks and faults are presented in Tables I-IV.

The rows in Table I indicate the TPR of actuator attack (AA) detection given different scenarios for simultaneous occurrence of anomalies in the system, such as occurrence of AA and SA, AA and AF, and AA and SF. Furthermore, the second column in this table shows the computed TPR for Proposition 1. In Table II the rows show the TPR of sensor attack (SA) detection in various scenarios for simultaneous occurrence of anomalies in the system. Moreover, the second column corresponds to the computed TPR of detection for SA where Proposition 2 is utilized. In Table III, the computed TPR of detection of actuator fault (AF) in presence of different anomalies are shown in the rows. Finally, the rows in Table IV indicate the TPR for sensor fault (SF) under simultaneous occurrences of anomalies in the system.

V. CONCLUSION

In this paper, the problem of simultaneous detection and isolation of machine induced faults and intelligent malicious adversarial cyber attacks has been studied. A methodology based on the cyber-physical systems (CPS) two side filters and a UIO-based detector has been proposed. In this method, a filter was designed on the plant side with its dynamics different from the C&C side filter so that even if the adversary estimates the parameters of the C&C side filter they cannot identify the parameters of the plant side filter. Moreover, this methodology inhibits adversaries from disguising their cyber attacks. Using the proposed strategy, one is capable of *simultaneously* detecting machine induced actuator and sensor faults as well as undetectable cyber attacks, such as covert and zero dynamics attacks, and detectable cyber attacks, such as the replay attack. In future work we will consider non-ideal communication networks. Furthermore, to make the cyber-physical systems model closer to the real-world applications, we will extend the results of this paper to a multi-agent based framework.

VI. APPENDIX

REFERENCES

[1] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "A secure control framework for resource-limited adversaries," *Automatica*, vol. 51, no. Supplement C, pp. 135 – 148, 2015.

[2] F. Pasqualetti, F. Drfler, 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.

[3] S. K. Khaitan and J. D. McCalley, "Design techniques and applications of cyberphysical systems: A survey," *IEEE Systems Journal*, vol. 9, no. 2, pp. 350–365, June 2015.

[4] Y. Li, P. Zhang, L. Zhang, and B. Wang, "Active synchronous detection of deception attacks in microgrid control systems," *IEEE Transactions on Smart Grid*, vol. 8, no. 1, pp. 373–375, Jan 2017.

[5] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept 2009, pp. 911–918.

[6] F. Miao, M. Pajic, and G. J. Pappas, "Stochastic game approach for replay attack detection," in *52nd IEEE Conference on Decision and Control*, Dec 2013, pp. 1854–1859.

[7] J. P. Hespanha, P. Naghshtabrizi, and Y. Xu, "A survey of recent results in networked control systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 138–162, Jan 2007.

[8] Z. Zhao, Y. Yang, Y. Li, and R. Liu, "Security analysis for cyber-physical systems under undetectable attacks: A geometric approach," *International Journal of Robust and Nonlinear Control*.

Algorithm 1 Pseudo code for cyber attack detection based on Propositions 1 and 2.

UIO-based detector design:
1) Find H^{AA} such that $(I - H^{AA}C)F_1 = 0$ and $(I - H^{AA}C)F_2 = 0$.
2) Compute K_1^{AA} such that $F^{AA} = A - H^{AA}CA - K_1C$ is Hurwitz.
3) Set $T^{AA} = I - H^{AA}C$.
4) Find L^{AA} such that $L^{AA}D_{ac} = 0$ and check if the Rosenbrock system matrix

$$\begin{bmatrix} sI - F^{AA} & L^{AA} \\ C & 0_{p \times n} \end{bmatrix}$$

is left-invertible, if not go to Step 1 where H^{AA}, K_1^{AA}, and L^{AA} are changed.

Design of filters and residual generation subject to actuator cyber attack detection (Proposition 1):
5) Find K_p^{AA} such that $K_p^{AA}D_a = 0$.
6) Compute L_p^{AA} such that $L_p^{AA}D_{ac} = 0$.
7) Find a diagonal matrix F_p^{AA} and the matrix T_p^{AA} such that the Rosenbrock system matrix

$$P_{\Sigma_u}(s) = \begin{bmatrix} sI - (F_p^{AA} + L_p^{AA}) & -T_p^{AA}B_a \\ L^{AA} & 0_{(n+p_f+p) \times m_a} \end{bmatrix}$$

does not have any non-minimum phase zero dynamics and rank $(L^{AA}T_p^{AA}B_a) = $ rank $(T_p^{AA}B_a)$.
8) Check if $(F_p^{AA} + L_p^{AA})$ is Hurwitz, if not go to Step (6).
9) Generate the residual signal $res_{AA}(t)$ and compute the threshold η_{AA} according to Remark 1.

Design of filters and residual generation subject to sensor cyber attack detection (Proposition 2):
10) Set $T^{SA} = T^{AA}$, $H^{SA} = H^{AA}$, $L^{SA} = L^{AA}$, and $F^{SA} = F^{AA}$.
11) Find T_p^{SA} such that $T_p^{SA}B_a = 0$.
12) Compute L_p^{SA} such that $L_p^{SA}D_{ac} = 0$.
13) Find a diagonal matrix F_p^{SA} and the matrix K_p^{SA} such that the Rosenbrock system matrix

$$P_{\Sigma_y}(s) = \begin{bmatrix} sI - (F_p^{SA} + L_p^{SA}) & K_p^{SA}D_a \\ L^{SA} & 0_{(n+p_f+p) \times p_a} \end{bmatrix}$$

does not have any non-minimum phase zero dynamics and rank $(L^{SA}K_p^{SA}D_a) = $ rank $(K_p^{SA}D_a)$.
14) Generate the residual signal $res_{SA}(t)$ and compute the threshold η according to Remark 1.

Algorithm 2 Pseudo code for fault detection based on Propositions 3 and 4.

UIO-based detector design and residual generation subject to actuator fault detection (Proposition 3):
1) Find H^{AF} such that $(I - H^{AF}C)F_2 = 0$.
2) Compute K_1^{AF} such that $F^{AF} = A - H^{AF}CA - K_1^{AF}C$ is Hurwitz.
3) Set $T^{AF} = I - H^{AF}C$.
4) Set $L^{AF} = 0$.
5) Generate the residual signal $res_{AF}(t)$ and compute the threshold η according to Remark 1.

UIO-based detector design and residual generation subject to sensor fault detection (Proposition 4):
6) Find H^{SF} such that $(I - H^{SF}C)F_1 = 0$.
7) Set $T^{SF} = I - H^{SF}C$ and $L^{SF} = L^{AF}$.
8) Generate the residual signal $res_{SF}(t)$ and compute the threshold η according to Remark 1.

[9] A. Baniamerian and K. Khorasani, "Security index of linear cyber-physical systems: A geometric perspective," in *2019 6th International Conference on Control, Decision and Information Technologies (CoDIT)*, April 2019, pp. 391–396.
[10] A. Baniamerian, K. Khorasani, and N. Meskin, "Determination of security index for linear cyber-physical systems subject to malicious cyber attacks," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 4507–4513.
[11] A. Teixeira, H. Sandberg, and K. H. Johansson, "Networked control systems under cyber attacks with applications to power networks," in *Proceedings of the 2010 American Control Conference*, June 2010, pp. 3690–3696.
[12] M. A. Massoumnia, "A geometric approach to the synthesis of failure detection filters," *IEEE Transactions on Automatic Control*, vol. 31, no. 9, pp. 839–846, September 1986.
[13] S. H. Zad and M.-A. Massoumnia, "Generic solvability of the failure detection and identification problem," *Automatica*, vol. 35, no. 5, pp. 887 – 893, 1999.
[14] J. Chen, R. J. Patton, and H.-Y. Zhang, "Design of unknown input observers and robust fault detection filters," *International Journal of control*, vol. 63, no. 1, pp. 85–105, 1996.
[15] J. Wünnenberg and P. Frank, "Sensor fault detection via robust observers," in *System fault diagnostics, reliability and related knowledge-based approaches*. Springer, 1987, pp. 147–160.
[16] N. Tudoroiu and K. Khorasani, "Fault detection and diagnosis for satellite's attitude control system (acs) using an interactive multiple model (imm) approach," in *Proceedings of 2005 IEEE Conference on Control Applications, 2005. CCA 2005.*, Aug 2005, pp. 1287–1292.
[17] B. Pourbabaee, N. Meskin, and K. Khorasani, "Sensor fault detection, isolation, and identification using multiple-model-based hybrid kalman filter for gas turbine engines," *IEEE Transactions on Control Systems Technology*, vol. 24, no. 4, pp. 1184–1200, July 2016.
[18] N. Meskin, E. Naderi, and K. Khorasani, "A multiple model-based approach for fault diagnosis of jet engines," *IEEE Transactions on Control Systems Technology*, vol. 21, no. 1, pp. 254–262, Jan 2013.
[19] N. Meskin and K. Khorasani, "Fault detection and isolation of discrete-time markovian jump linear systems with application to a network of multi-agent systems having imperfect communication channels," *Automatica*, vol. 45, no. 9, pp. 2032 – 2040, 2009.
[20] M. Davoodi, N. Meskin, and K. Khorasani, "Simultaneous fault detection and consensus control design for a network of multi-agent systems," *Automatica*, vol. 66, pp. 185 – 194, 2016.
[21] I. Shames, A. M. Teixeira, H. Sandberg, and K. H. Johansson, "Distributed fault detection for interconnected second-order systems," *Automatica*, vol. 47, no. 12, pp. 2757 – 2764, 2011.
[22] J. Gertler, "Fault detection and isolation using parity relations," *Control engineering practice*, vol. 5, no. 5, pp. 653–661, 1997.
[23] R. J. Patton and J. Chen, "A review of parity space approaches to fault diagnosis," *IFAC Proceedings Volumes*, vol. 24, no. 6, pp. 65–81, 1991.
[24] A. Hoehn and P. Zhang, "Detection of covert attacks and zero dynamics attacks in cyber-physical systems," in *2016 American Control Conference (ACC)*. IEEE, 2016, pp. 302–307.
[25] A. Barboni, H. Rezaee, F. Boem, and T. Parisini, "Distributed detection of covert attacks for interconnected systems," in *2019 18th European Control Conference (ECC)*. IEEE, 2019, pp. 2240–2245.
[26] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "Revealing stealthy attacks in control systems," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Oct 2012, pp. 1806–1813.
[27] F. Miao, Q. Zhu, M. Pajic, and G. J. Pappas, "Coding schemes for securing cyber-physical systems against stealthy data injection attacks," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 1, pp. 106–117, March 2017.
[28] S. Weerakkody and B. Sinopoli, "Detecting integrity attacks on control systems using a moving target approach," in *2015 54th IEEE Conference on Decision and Control (CDC)*. IEEE, 2015, pp. 5820–5826.
[29] C. Schellenberger and P. Zhang, "Detection of covert attacks on cyber-physical systems by extending the system dynamics with an auxiliary system," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, Dec 2017, pp. 1374–1379.
[30] H. L. Trentelman, A. A. Stoorvogel, and M. Hautus, *Control theory for linear systems*. Springer Science & Business Media, 2012.
[31] M.-A. Massoumnia, G. C. Verghese, and A. S. Willsky, "Failure detection and identification," *IEEE Transactions on Automatic Control*, vol. 34, no. 3, pp. 316–321, 1989.
[32] W. M. Wonham, "Linear multivariable control," in *Optimal control theory and its applications*. Springer, 1974, pp. 392–424.
[33] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861 – 874, 2006, rOC Analysis in Pattern Recognition.