Trust-Based Framework for Resilience to
Sensor-Targeted Attacks in Cyber-Physical Systems

Tracie Severson, Erick Rodriguez-Seda and Kiriakos Kiriakidis
Department of Weapons and Systems Engineering
United States Naval Academy
Email: {severson,rodrigue,kiriakid}@usna.edu

Brien Croteau, Deepak Krishnankutty, Ryan Robucci, Chintan Patel, and Nilanjan Banerjee,
Department of Computer Science and Electrical Engineering
University of Maryland, Baltimore County
Email: {croteau1,deepak1,robucci,patel2,nilan}@umbc.edu

Abstract—Networked control systems improve the efficiency of cyber-physical plants both functionally, by the availability of data generated even in far-flung locations, and operationally, by the adoption of standard protocols. A side-effect, however, is that now the safety and stability of a local process and, in turn, of the entire plant are more vulnerable to malicious agents. Leveraging the communication infrastructure, the authors here present the design of networked control systems with built-in resilience. Specifically, the paper addresses attacks known as false data injections that originate within compromised sensors. In the proposed framework for closed-loop control, the feedback signal is constructed by weighted consensus of estimates of the process state gathered from other interconnected processes. Observers are introduced to generate the state estimates from the local data. Side-channel monitors are attached to each primary sensor in order to assess proper code execution. These monitors provide estimates of the trust assigned to each observer output and, more importantly, independent of it serving as weights in the consensus algorithm. The authors tested the concept on a multi-sensor networked physical experiment with six primary sensors. The weighted consensus was demonstrated to yield a feedback signal within specified accuracy even if four of the six primary sensors were injecting false data.

I. INTRODUCTION

The integration of computing, communication, and control dictate the design of Cyber-Physical Systems (CPS) [1]. Enabled by advancements in all three areas, tight integration will achieve efficiencies by eliminating the need for multiple interfaces assessed in the engineering of control systems. Unfortunately, the flexibility provided by this integration also opens up several attack surfaces and security vulnerabilities. To address this problem, the authors envisage a framework for the next generation of Networked Control Systems that has built-in resilience to physical and cyber-attacks.

Fig. 1 demonstrates the concept and an overview of the approach. Here a physical plant is depicted as a circular embedding of interconnected and possibly geographically spread processes. Consider the group of processes \( P_i \), \( P_j \), and \( P_m \) with dedicated sensors \( S_i \), \( S_j \), and \( S_m \) that directly measure their respective states. At their geographic location, sensors \( S_i \), \( S_j \), and \( S_m \) are coupled to observers \( O_i \), \( O_j \), and \( O_m \), respectively. Using data relayed by the companion sensor, the function of each observer is to estimate (possibly from a far-flung location of the plant) the state of other processes connected to its own. Based on structural observability of the plant, the authors have developed a method for the realization of such estimators [2]. Data and estimates generated locally can be shared within the group of physically interconnected processes via the CPS communication network. The authors propose that the global values of the respective states of processes \( P_i \), \( P_j \), and \( P_m \) be determined by consensus. In addition to the observers, the concept requires the addition of side-channel monitors \( PM_i \), \( PM_j \), and \( PM_m \) in order

This work was supported by the U.S. Office of Naval Research under Award N00014-15-1-2179 and N0001417WX01442.
to analyze the power consumption profile of the sensors’ processing units [3]. By monitoring the code execution at the microprocessor level of sensors $S_i$, $S_j$, and $S_m$ using this side-channel, the trustworthiness of each sensor can be evaluated and their outputs weighted appropriately. These weights are taken into account along with the local estimates in a trust-based consensus algorithm that feeds back the control input $\hat{x}_c$. Using such a cross-layer approach (micro-level side-channel analysis and process-level consensus), the authors demonstrate that single-point failures can be avoided when a compromised sensor injects false data [3], [5].

A. Contributions

(i) This work proposes a framework that uses information gathered from side-channel power analysis of each sensor to weigh each sensor’s trustworthiness when estimating the true state of the plant. This approach provides two novel contributions. First, it hardens the system against an attack as both the primary and the side-channel sensors have to be compromised simultaneously, requiring two different attack modes. Second, it provides the system the ability to infer meaningful consensus even in the event that a majority of the primary sensors are compromised.

(ii) The authors have prototyped this trust-based model in the context of a multi-sensor networked testbench controlling a process around a set operating temperature. Using data collected from the testbed and power-supply data collected during hardware measurements, the authors demonstrate improved resilience and the ability to stay within a safe operating region even when four of its six primary sensors were compromised.

B. Related Work

This work contrasts with a large majority of published CPS research [8] that has concentrated on the detection of attacks or intrusions. While there are several papers that focus on countermeasures and responses to cyber-attacks on smart grids [7]–[9] and attacks of SCADA [10] and Industrial Control Systems (ICS) [11], in general not as much attention has been placed on the prevention and response to cyber-attacks on CPS systems.

This work also seeks to address the cyber-attack protection problem using the well-studied method of consensus in multi-agent systems [12]. Consensus refers to agents reaching agreement on a particular quantity of interest with every other agent in a consensus network. Consensus algorithms can also be applied to distributed estimation schemes, where multiple neighboring agents iteratively exchange their own estimates of the targets state in order to converge to a single state estimate over the entire network [13]. Most often it is assumed that all agents in a cooperative exercise are equally capable or trustworthy in seeking consensus on a particular decision.

State estimation, however, can be problematic when an agent has limited observability of the target [14]. Even in the case of full observability, a node may become compromised and report a corrupted value of that estimate. If the node is given an equal weighting, the performance of the consensus-based estimation framework may degrade and eventually drive the system unstable. To address this problem, trust-based consensus algorithms [15], place weights along the edges of the communication network between agents and force the estimate to converge towards a value that is closer to the agents with higher trust values.

Starting from a more general approach, [16] and [17] also address the problem of correctly estimating state information in the presence of adversarial attacks similar to those described in Section II-B. They both proved it is only possible to reconstruct the state of a Linear Time Invariant (LTI) system with optimal estimators if half or less of the sensors are attacked.

II. SECURITY MODEL

This section will summarize the threats a CPS system will be facing, the assumptions about the adversary carrying out the attack, and the security properties this method is striving to maintain. The format of this section is based on the one used in [18].

A. Threat Model

The adversary desires to – (i) drive the system to an arbitrary state to cause damage to the underlying physical plant, (ii) remain undetected until attack is launched; (iii) prevent valid sensor data from reaching the controller.

B. Adversary Model

Adversary can – (i) change code running on a sensor’s processing unit, (ii) replay past collected data, (iii) send data to the controller that is indistinguishable from that of uncompromised sensors, (iv) monitor and communicate to compromised sensors; (v) use multiple compromised sensors to collude. The adversary cannot – (i) compromise the micro-level side-channel measurement sensors; (ii) launch an attack that prevents the receipt of data from uncompromised sensors.

C. Security Properties

While all the usual security priorities are important, the most essential in a control system with distributed sensors are safety, integrity, and availability. Overall safety of the underlying plant and its operators is the most important goal. Data integrity is the assurance that the data being used to make control decisions reflects the actual state of the process being controlled. This work assures data integrity by preferentially using data from uncompromised sensors to form an estimate of the state at a critical node. Data availability is the guarantee that data from the sensors will continue to be made available to the controller in a timely manner.
III. BACKGROUND

A. Micro-level Measurements from Power Monitors

The data gathered from micro-level measurements utilizes side-channel leakage analysis to detect deceptive attacks against a sensor’s processing unit when its firmware is modified. Differential Power Analysis [19] is a well-studied topic in cryptanalysis usually focused on recovering secret information stored within a circuit or device by using statistical information from multiple measurements of a physically linked side-channel signal, such as the power consumption of the integrated circuit. The authors have previously [3], [5] proposed methods to use such measurements instead to detect the presence of an attack by determining when the code running inside a sensor’s processing unit has been modified. Fig. 2 gives an overview of the micro-level process to generate the metric used as a sensor quality indicator.

**Fig. 2.** Block Diagram of Micro-Level Measurement Process. Power-supply observations on sensor $S_i$, that form the reference waveform $b(n)$ are processed through a training phase to generate the golden waveform, $b_{gw}(n)$. As part of the testing phase, new waveforms $a(n)$ are collected from sensor $S_i$ and compared to $b_{gw}(n)$ using the correlation metric, $\rho_i$, defined in Eqn. 1 which is then passed to the consensus function as a quality metric.

“golden waveform” generated from the point-wise average over the training set of $M$ waveforms.

After deployment, new power supply traces, $a(n)$, are collected and compared to the golden waveform by first removing their direct current (dc) components (the mean over the entire sequence) $a_{ac}(n) = a(n) - a_{dc}(n)$, then calculating a correlation metric,

$$\rho_i = \frac{\sum_{n=1}^{N_{\text{max}}} a_{ac}(n) \cdot b_{ac}(n)}{\sqrt{\sum_{n=1}^{N_{\text{max}}} a_{ac}(n) a_{ac}(n) \cdot \sum_{n=1}^{N_{\text{max}}} b_{ac}(n) b_{ac}(n)}}. \quad (1)$$

The $\rho_i$ values, provide an early indicator to anomalous behavior exhibited by the associated sensor $S_i$ as shown in Fig. 2. Later, in Section IV, this information is utilized as the basis for a weighted estimator of compromised sensors in a networked CPS.

B. State Estimates from Observers

Without loss of generality, the plant depicted in Fig. 1 can be modeled as an LTI dynamic system with state matrix $A \in \mathbb{R}^{n \times n}$. The state of the plant is the vector that comprises the states, $x_j \in \mathbb{R}$, of the plant’s constituent processes, $P_j$, for $i = 1, \ldots, n$. The input matrix is diagonal $B \in \mathbb{R}^{n \times n}$ with 1 or 0 on the main diagonal depending on the availability of an input channel. Similarly, the output matrix is diagonal $C \in \mathbb{R}^{1 \times n}$ with 1 or 0 on the main diagonal depending on the availability of an output channel for sensing. The $i$-row of $C$ with 1 on the diagonal represents a process-specific output matrix $C_i \in \mathbb{R}^{1 \times n}$ corresponding to an output channel topped by the sensor $S_i$ that directly measures $x_i$.

The authors’ method produces an estimate, $\hat{x}_j \in \mathbb{R}$, of the state of process $P_i$ from measurement of the state of process $P_j$. The state $x_i$ must be reconstructable from data gathered by sensor $S_j$ and local knowledge of the model $(A, C_j)$. Since the model $(A, C_j)$ is not observable in general, the authors have developed an approach to identify an observable albeit reduced-order model $(\hat{A}_j, \hat{C}_j)$ for state reconstruction of $x_i$. The model $(\hat{A}_j, \hat{C}_j)$ is the basis for the design of the observer $O_j$ in Figs. 1 and 3. The authors’ approach in [20] is grounded in an early work on controllable subspaces [21]. The main result is that the graph of an unobservable model $(A, C_j)$ contains a maximal subgraph whose number of edges determines the dimension of the model’s observable subspace, which eventually yields $(\hat{A}_j, \hat{C}_j)$.

IV. TRUST-BASED FRAMEWORK

The observer estimates in Section III.B form the initial conditions for the trust-based consensus framework that mitigates the impact of multiple attacks of deception on a CPS. The output of the consensus framework is the consensus value that gets fed back into a closed-loop system, as shown in Fig. 3. Trusting each observer estimate equally works well when each sensor node properly communicates their observed estimate. In the presence of a false data attack, however, sensors may be reporting previously recorded estimates that appear normal but in fact deviate significantly
where the Perron matrix \( \mathbf{P} \) weighted consensus algorithm for agents with discrete time applying a weighted average consensus algorithm. A measure of trust amongst the sensor estimates before closed-loop systems, a method is presented for incorporating away from the actual value and may cause a system to final consensus value to be farther value of that estimate, an unweighted average consensus algorithm will force the final consensus value to be farther

\[ c_j(k) = \frac{1}{\Delta_j(k)} \sum_{m \in N_j} t_{jm}(k) a_{jm}(\hat{x}_m(k) - i \hat{x}_j(k)) \]  (3)

with \( \Delta_j(k) = \sum_{m \in N} t_{jm}(k) a_{jm} \) the weighted degree of all outgoing edges of \( j \). In matrix form the weighted consensus protocol is

\[ c(k) = -(I - D^{-1} \Gamma)x(k) \]  (4)

where \( \Gamma = T \circ A_G \) is the weighted adjacency matrix and the symbol \( \circ \) denotes the entrywise (or Hadamard) product. The matrix resulting from \( I - D^{-1} \Gamma \) is a normalized Laplacian matrix \( D^{-1}L = D^{-1}(D - \Gamma) \). If the normalized Laplacian is substituted in for \( P \) in Eqn. A then the collective consensus dynamics can be expressed as

\[ i \hat{x}(k+1) = [(1 - \epsilon)I + \epsilon D^{-1} \Gamma] i \hat{x}(k) \]  (5)

The proof of convergence for this protocol can be found in [13], and the final value is referred to as \( i \hat{x}_c \) in this work.

C. Computing Trust Values using Micro-level Measurements

A trust-based consensus protocol can mitigate the negative effects of sensors reporting erroneous data by weighting corrupted sensor opinions less than those that are not corrupted in the consensus algorithm. This section explains how the trust values for each sensor in the trust-matrix \( T \) are computed.

Section III-A describes how a large number of observations, \( N_{max} \), of each sensors processing unit are collected to form both a nominal (golden) waveform as well as new power supply traces. These new traces are then compared to the nominal waveform and a correlation metric \( r_t \) is calculated.

Since \( N_{max} \) is sufficiently large, the distribution of the correlation metric can be modeled as a normal distribution [21]. If a node gets compromised, the corrupted data will distort this normal distribution. A trust-based framework that monitors the nodes dynamic distribution will expose this difference and based on this, quantify a node’s trust level [22].

Let the ideal (or uncorrupted) nodes Probability Distribution Function (PDF) be defined as \( \rho \sim \mathcal{N}(\mu_\rho, \sigma^2_\rho) \). At each time step, let the actual PDF for each sensor \( S_i \) be defined as \( \rho_t \sim \mathcal{N}(\mu_{\rho_t}, \sigma^2_{\rho_t}) \). Using the ideal node frequency as a criterion, the difference between the ideal node PDF and actual node PDF [22] is measured. The smaller the difference between the two distributions, the more trustworthy a node is and vice versa. This difference is known in mathematical statistics as the Kullback-Leibler divergence and measures how one probability distribution diverges from a second expected probability distribution [23]. For two discrete probability distributions \( p \) and \( q \), the KL divergence of \( q \) from \( p \) is

\[ D_{KL}(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \]  (6)
where $x$ is one outcome on the sample space $X$. If both distributions are Gaussian, then Eqn. (6) can be written

$$D_{KL}(p||q) = \log \frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_q^2} - \frac{1}{2}$$

The KL-divergence value computed in Eqn. (6) provides a measure of how much one distribution differs from another (or how close the distributions match each other). Note that the KL-divergence value is not symmetric, $D_{KL}(p(x)||q(x)) \neq D_{KL}(q(x)||p(x))$, so it should not be considered a true distance metric. However, KL-divergence can be used to measure how close each sensor’s correlation value $\rho_j = q$ matches that of an uncorrupted or nominal sensor $\hat{\rho} = p$. It is expected that the KL-divergence values of uncorrupted nodes would be less than those of compromised nodes, thus providing quantifiable insight into a node’s trustworthiness. Because a larger KL-divergence value can be equated to a lower trust value, the trust value should be inversely proportional to the KL-divergence. Let $D_{KL}(j)$, represent the KL-divergence between sensor $S_j$ and a nominal (known uncorrupted) sensor. Then, the trust value of sensor $S_j$ is

$$t_j = \frac{1}{1 + \sqrt{D_{KL}(j)}}$$

where severe oscillations at the beginning of the trust computation are smoothed due to the decreased sensitivity of the square root function [72]. Note that all sensor nodes, including node $j$, will compute sensor $j$’s trust value based upon the same $\rho_j$ data from the micro-level measurements, thus $t_{ji} = t_j \forall i \in N$.

### D. Closing the Control Loop

The main control objective of any critical CPS is to maintain operational normalcy [23] regardless of disturbances, including cyber attacks. By operational normalcy it is meant that the regulated output $x_i$ should remain near its desired operating point $OP_i$ at all times. Mathematically, if one defines $\delta_i > 0$ as the safety-based bound on the maximum admissible deviation of $x_i$ from $OP_i$, then, the operational normalcy region for the $i$th process can be defined as

$$\mathcal{O}_i = \{x_i \in \mathbb{R} \mid OP_i - \delta_i \leq x_i \leq OP_i + \delta_i\}$$

To maintain operational normalcy, the authors propose to update the control action using the trust-based estimate $\hat{\delta}_c$ fed by the micro-level measurement as shown in Eqn. (8). The control action for the $i$th process becomes a function of $\hat{\delta}_c$, which takes into account not only information from the process’s sensor but from other possibly far-flung sensors. The expectation being that fusing information from different sensors according to their trustworthiness will lead to better resiliency and control performance under sensor-based attacks.

---

(a) Micro-level Analysis Setup  (b) Temperature Control Testbed

Fig. 5. In (a), a custom-designed board is shown with the FPGA under test and side-channel power supply current probes. In (b), the testbed consists of six LM34 temperature sensors mounted on an aluminum plate connected to six micro-controllers communicating via CAN bus. The heating element was located in the upper left corner of the plate.

V. EXPERIMENTAL DESIGN

A physical experiment was performed to verify the method proposed in Section IV. A physical testbed was employed to demonstrate responses to cyber-attacks when controlling the temperature of a sensitive process.

At the micro-level, the processing unit of a single sensor was replicated by an instantiation of an open source version of the TI MSP430 on an FPGA testbed (Fig. 5(a)), where two different code sequences were run and the power supply current traces were captured as described in Section III-A. In the baseline code sequence, the actual ADC output was used for all the measurements. For the replay attack, prior ADC values are sampled and stored in memory, and seven out of every ten measurements used the stored values instead of the correct ADC readings.

The temperature control testbed, shown in Fig. 5(b), was comprised of six TI LM34 temperature sensors spaced diagonally on a $8'' \times 8'' \times \frac{1}{2}''$ aluminum plate. Sensor $S_i$ lay directly on top of the thermoelectric heater, which simulated the correcting element or actuator. The remaining sensors $S_5, S_6, S_7, S_8$ were mounted at $1''$, $3''$, $5''$, $7''$ and $9''$ from the heater. Each sensor was connected to an LPC1768 ARM micro-controller (MBED) which communicated the temperature measurements along a CAN bus at $20$Hz. An additional MBED monitored the CAN bus and communicated these temperatures over a serial connection to a PC running MATLAB for data processing.

In this work, a simplified control scheme running at $0.4$Hz was employed that toggled power on and off of the heating element to maintain the temperature of the sixth node at $OP_6 = 90^\circ F$. When the feedback signal used to close the loop reported a temperature below $OP_6 = 90^\circ F$, the heater was turned on. When the signal reported otherwise, the heater was turned off. The maximum admissible deviation allowed from the operating point was defined as three standard deviations of the temperature sensor’s manufacturer accuracy [25], that is, $\mathcal{O}_6 = \{6^\circ F \in \mathbb{R} \mid 84^\circ F \leq 6^\circ F \leq 96^\circ F\}$.

The sensor-based replay attacks consisted of a mixture of old and new measurement data. At each control update instance, each sensor would report to its observer (on average) a total of ten temperature measurements. These set of samples for a single sensor would then be averaged and used as input
The micro-level measurements were delayed 30 seconds to account for the processing time required to measure the true values.

To illustrate the performance of the proposed control scheme, three different closed-loop experiments were conducted. The first experiment consisted in closing the temperature control loop using only the reading reported by sensor $S_6$, which was attacked after 120 s. The feedback signal being made up of compromised sensors reporting incorrectly low temperatures. Soon after the third sensor was attacked, the plate temperature exited its operational normalcy region of $90^\circ$ F $\pm 6^\circ$ F.

The third experiment consisted in closing the control loop using the trust-based weighted consensus from Section IV-C. The system was subjected to the same sequence of replay attacks conducted in the second experiment. As can be noticed from Fig. 8(a), after the start of each attack, the temperature at node six $x_6$ deviated slightly from the $OP_6$ but converged to $90^\circ$F shortly after. After each attack, the deviations were larger as the consensus algorithm fused data from more malicious sensors. However, in contrast to the use of a single measurement (see Fig. 7) and the general consensus case (refer to Fig. 6), the system was able to maintain operational normalcy despite the attack on over half of its primary sensors.

In order to show that the performance of the proposed resilient method did not depend on which sensors were under attack or in what order, a small Monte Carlo set of results is shown in Fig. 8(b). It shows the response of the proposed trust-based consensus to three different sequences (shown in Table I) of replay attacks. By design, the critical sensor $S_6$ was always compromised first, and the other nodes were selected using a random permutation generator.
cost and complexity of augmenting it with other sensors for temperature to stay within its safe operating range even in the sensors used the primary sensor’s power consumption as an form a trust-based weighting scheme. These side-channel algorithms to Close the Loop. (a) shows the closed-loop system response when maintaining a cyber-physical system within a safe operating region.

MATLAB function. In the three cases, the closed-loop actual temperature was stabilized within the operational normalcy region.

VII. CONCLUSION AND FUTURE WORK

In this paper, the authors considered the problem of maintaining a cyber-physical system within a safe operating envelope in the presence of multiple sensor cyber-attacks. A framework to incorporate information from additional side-channel power monitoring sensors was leveraged to form a trust-based weighting scheme. These side-channel sensors used the primary sensor’s power consumption as an additional natural process to determine trust. As demonstrated in a physical experiment, this framework allowed the critical temperature to stay within its safe operating range even in the presence of four of its six primary sensor modules becoming compromised. This work effectively addressed two important security properties, namely safety and data integrity.

In the future, the authors plan to incorporate other heterogeneous side-channel sensors along with the power supply sensors used in this work. Trade-offs between the system cost and complexity of augmenting it with other sensors for added resilience will be investigated.

REFERENCES


