Formal Synthesis of Monitoring and Detection Systems for Secure CPS Implementations

Ipsita Koley¹, Saurav Kumar Ghosh¹, Soumyajit Dey¹, Debdeep Mukhopadhyay¹,

Amogh Kashyap K N², Sachin Kumar Singh², Lavanya Lokesh², Jithin Nalu Purakkal², Nishant Sinha²

1 *Indian Institute of Technology,Kharagpur,* ²*Robert Bosch Engineering and Business Solutions Private Limited*

{ipsitakoley, soumyajit, debdeep}@iitkgp.ac.in, saurav.kumar.ghosh@cse.iitkgp.ernet.in

{Amogh.Kashyap, SachinKumar.Singh, Lokesh.Lavanya, Jithin.NaluPurakkal2,Sinha.Nishant}@in.bosch.com

Abstract—We consider the problem of securing a given control loop implementation of a cyber-physical system (CPS) in the presence of Man-in-the-Middle attacks on data exchange between plant and controller over a compromised network. To this end, there exists various detection schemes which provide mathematical guarantees against such attacks for the theoretical control model. However, such guarantees may not hold for the actual control software implementation. In this article, we propose a formal approach towards synthesizing attack detectors with varying thresholds which can prevent performance degrading stealthy attacks while minimizing false alarms.

Index Terms—Cyber Physical System, False data injection attack, Formal method, Residue based detector.

I. INTRODUCTION

Unattended communication among devices in distributed CPS implementations makes new pathways for malicious interference. Given that such systems often need to perform safety critical functionalities with real time deadlines within stringent power, energy requirements, the impact of attacks on safety-critical CPS may have catastrophic consequences. In the past decade, many such high profile attacks have been reported spanning a variety of application domains ($[1]-[4]$ $[1]-[4]$). It is infeasible to physically secure every packet transmission between CPS components due to limited communication bandwidth as well as lightweight nature of computing nodes. This rules out using heavyweight cryptographic encryption techniques (like RSA, AES) along with MACs for securing all intra-vehicular communication [\[5\]](#page-3-2). Hence, it makes sense to enhance the security of CPS implementations by using suitable lightweight monitoring primitives considering that an attacker has already breached into the CPS communication infrastructure.

In this work, we focus on residue-based monitoring and detection systems which compute the difference between plant output measurements received through a communication network and the estimates of the same based on earlier measurements and knowledge about system dynamics, raising an alarm if the difference (i.e. the *residue*) exceeds a predefined threshold. Since this type of anomaly detector uses the properties of the control system to detect an adversarial action, it does not impose any significant overhead to the system's resource consumption in terms of communication and computation. Although there exists significant literature on

residue-based detectors [\[6\]](#page-3-3)–[\[8\]](#page-3-4), none of these works discusses an effective methodology for synthesizing thresholds given a control system specification. Also existing works consider static thresholds only, i.e. the difference in measurement and estimate is compared with a constant pre-fixed threshold for all closed loop iterations of the system.

As a potential example of targeted performance degrading attack, consider the situation when the reference point of a controller changes due to occurrence of some event. For such systems, with a comparatively smaller fault injection at the later stage of dynamics (i.e. when nearing the reference), an attacker can prevent the system from reaching the close vicinity of the reference. This brings in interesting trade-offs from the detector design point of view. In a static thresholdbased detection scheme, if the threshold is decided based on the required attack amount at the later phase of settling time, it may be the case that any process or measurement noise induced by environmental disturbance in the system is considered an attack and a *false alarm* is generated. This implies the False Alarm Rate (FAR) will increase. If the threshold is decided based on the attacker's effort at the earlier phase of settling time, the attacker can easily bypass the detection scheme by injecting sufficiently small anomalies whenever the system is very close to the reference and deteriorate the system's performance. This motivates the case for a variable threshold based anomaly detection method which may ensure reduced FAR while identifying even small attack efforts that may lead to potential performance degradation.

(a) Effect of noise and attack

(b) Static vs dynamic threshold

Fig. 1: Trajectory tracking system

As a motivational example, we consider a trajectory tracking system (Fig. [1\)](#page-0-0) taken from [\[9\]](#page-3-5). A suitably crafted attack can steer the system towards instability as shown in the same figure. In Fig. [1b](#page-0-0), we consider three possible residue based detectors, with the smaller threshold th , the bigger threshold Th and the variable threshold curve v_{th} . Note that with th ,

The authors acknowledge generous support from Robert Bosch Engineering and Business Solutions Private Limited.

the detector considers even the harmless noise as an attack, as shown (Fig. [1b](#page-0-0)). On the other hand, with Th , the actual attack could easily bypass the detector. However using the variable threshold curve v_{th} (dotted red line in Fig. [1b](#page-0-0)), the attack does not remain stealthy while harmless noise is allowed to pass reducing the FAR.

In this article, we propose a formal approach for synthesizing residue based attack detectors with variable thresholds for CPS implementations that can prevent stealthy attacks. These detectors are also guaranteed to have smaller FAR w.r.t. provably safe static threshold based detector options.

II. SYSTEM MODEL AND PROBLEM FORMULATION

Consider a discrete linear time invariant (LTI) plant model S given as, $x_{k+1} = Ax_k + Bu_k + w_k$, $y_k = Cx_k + Du_k + v_k$, where $x_k \in \mathbb{R}^n$, $y_k \in \mathbb{R}^m$, $w_k \in \mathbb{R}^n \sim \mathcal{N}(0, Q)$ and $v_k \in \mathbb{R}^m \sim \mathcal{N}(0, R)$ represent system state variables, sensor measurements of plant, zero mean Gaussian process and measurement noise at k^{th} sampling instance respectively. Also A, B and C are transition matrix, input map and output map for the plant model respectively. To estimate the system states \hat{x}_k from the observed ones, a Kalman filter based observer is deployed, given by, $z_k = y_k - C\hat{x}_k$, $\hat{x}_{k+1} = A\hat{x}_k + Bu_k + Lz_k$, where residue $z_k = y_k - C\hat{x}_k$ is the difference between measured and estimated output at k^{th} time instance, and L is Kalman gain. The controller output u_k is computed as, $u_k = -K\hat{x}_k$. In this paper we contemplate false data injection based attack scenario in which the attacker falsifies the sensor measurements by injecting $a_k \in \mathbb{R}^m$ to sensor output y_k at k^{th} sampling instance. The resulting altered sensor measurements $\tilde{y}_k = y_k + a_k$ are fed to the estimator which in turn affects the control input calculation. Due to this, the closed loop dynamics deviates from the expected behavior. For this system description, we consider a threshold based detection scheme such that the detector will raise an alarm whenever $||z_k|| \geq Th[k]$ where $Th[k]$ is threshold at k^{th} sampling instance. We say an attack is stealthy if some given safety or performance criteria of the system is violated by the attacker while $||z_k||$ remains below $Th[k]$ for all k.

Formal Problem Statement: Consider the plant model S as discussed earlier, a controller implemented as a software program C running on an ECU, and a safety or performance criteria pfc. The objective of C is to satisfy pfc withing some $(j+T)$ th samples starting from any sampling instance j. An l length threshold specification Th is represented by a vector $\in \mathbb{R}^l$. Threshold is said to be *static* if $Th[i]$ is same for all i, else it is *variable*. We formally define the threshold synthesis problem as follows.

Given $\langle S, C, pfc \rangle$, what would be an optimal threshold spec*ification* Th *such that any stealthy attack is guaranteed to be detected as well as FAR is minimized ?*

III. THRESHOLD SYNTHESIS AND METHODOLOGY

As a first step in our approach, we propose Algorithm [1](#page-1-0) which formally checks the implementation $\mathcal C$ and identifies whether there exists any possible attack vector that can violate the target properties of the system. Given S, C, let x_{des} be the reference point for the system and the target property pfc is to reach $x_T \in \{x_{des} + \epsilon\}$ for some $\epsilon \in \mathbb{R}$ within a finite number of iterations, say T starting from any initial state $x_1 \in V \subseteq$ \mathbb{R}^n . An attacker would want to achieve $x_T \notin \{x_{des} + \epsilon\}$ after T such closed loop iterations. Some CPS implementations often incorporate certain monitoring constraints in addition to residue based attack detectors to check the sanity of the sensor measurements. Such constraints are captured using suitable predicates denoted as mdc. Algorithm [1](#page-1-0) takes as input mdc, pfc, a threshold vector Th and a finite duration T allotted for achieving pfc. The variable a_k signifying false data is assigned a value nondeterministically (Line 4) and is added with measurements subsequently. We say that an attack is *stealthy* but successful when predicates $||z_k|| \leq Th[k]$ and mdc are satisfied, but pfc is violated. This is modeled by the assertion A in Line 9. A is given as input to an SMT tool with the assert clause. It returns a successful attack vector A if the assertion A is satisfied (Line 11). Otherwise, it returns NULL (Line 13) which guarantees that no attack vector exists that remains stealthy over T iterations and violates the performance criteria pfc of the system.

Input Control property pfc , existing monitoring constraint mdc , computed threshold vector Th , attack duration T		
Output Attack vector A (if it exists, otherwise NULL)		
1: function $ATVECSYN(Th, pfc, mdc, T)$		
2: $x_1 \leftarrow \mathcal{V}$: $\hat{x}_1 \leftarrow 0$: $u_1 \leftarrow 0$:	\triangleright Initialization	
3: for $k = 1$ to T do		
4: $a_k \leftarrow nondeterministic_choice;$		
5: $u_k \leftarrow C x_k + D u_k + a_k$; $\hat{u}_k \leftarrow C \hat{x}_k + D u_k$;		
6: $z_k \leftarrow y_k - \hat{y}_k$;		
7: $x_{k+1} \leftarrow Ax_k + Bu_k$; $\hat{x}_{k+1} \leftarrow A\hat{x}_k + Bu_k + Lz_k$;		
8: $u_{k+1} \leftarrow -K\hat{x}_{k+1}$		
9: $\mathbb{A} \leftarrow$ assert($(\forall Th[p] \in Th, z_n < Th[p] \&\& mdc) \rightarrow pfc$);		
10: if A is <i>valid</i> then		
11: return $A \leftarrow [a_1 \cdots a_T]$;		
12: else		
13: return NULL;		

Algorithm 1: Attack vector synthesis

We now propose a methodology in Algorithm [2](#page-2-0) to synthesize a monotonically decreasing vector of thresholds to provably secure a given CPS against attacks. Given the state space of possible *l*-length variable threshold functions ($l \in \mathbb{N}$), we formulate heuristic approaches guided by our hypothesis of monotonically decreasing thresholds. To verify whether existing monitoring constraint (if any) suffices to detect any stealthy attack we generate an attack vector without any threshold based detector (Line 2−3) using Algorithm [1.](#page-1-0) If any attack vector is retrieved, we make a greedy choice and select the sampling instance i where maximum residue is generated due to this attack (Line 4) as a pivot point. A threshold at i is introduced to thwart the current attack (Line 5). With this new threshold we call Algorithm [1](#page-1-0) (Line 6) to check if any attack can bypass this detector. If found, we now search for new thresholds to be added to Th to stop this new attack in the following manner.

Case 1a [Line $9 - 11$]: For any of the existing thresholds $Th[p] \in Th$, we try to find out whether the current attack has produced any residue $||z_k|| \ge Th[p]$ before the p^{th} instance,

i.e. $k \leq p$. If any such z_k exists, we consider the maximum of them and include it to Th while ensuring monotonicity (Line $9 - 10$). If we get such a new threshold $Th[i]$ that keeps the monotonic decreasing order in Th intact, we stop searching (Line 11). Otherwise, we consider *Case* 1b.

Algorithm 2: Pivot Based Threshold Synthesis

Case 1b [Line 12 – 15]: For any of the thresholds $Th[p] \in$ Th , we try to find out whether the current attack has produced any residue $||z_k|| \ge Th[j]$ for all $j \in [k+1, T]$ where $k > p$. In that case, we consider the maximum of them (Line 12) and include it to Th while ensuring monotonicity (Line $13 - 14$). Otherwise, one or more existing thresholds in Th need to be reduced to detect the current attack (*Case* 1c).

Case 1c [Line 16 − 21]: We choose the candidate threshold $Th[i]$ from Th which can be reduced with minimum effort i.e. the minimum difference between the current threshold value $Th[i]$ and the residue $||z_i||$ generated by the attack (Line 17). For that i, we set $Th[i] = ||z_i||$ (Line 18) and adjust subsequent thresholds in order to ensure monotonicity (Line $19 - 21$).

Once a new threshold is introduced or existing thresholds are modified to detect the current attack, we call Algorithm [1](#page-1-0) (Line 6) with the modified Th . If it returns NULL, it is ensured that the latest Th is enough to thwart any stealthy attack. If not, we repeat the process with Case $1a$, $1b$ or $1c$ with the newly generated attack vector. While Algorithm [2](#page-2-0) can be used to synthesize monotonically decreasing thresholds, it can take a long time to converge. Hence we propose Algorithm [3](#page-2-1) which also starts with generating an attack vector without considering any threshold using Algorithm [1](#page-1-0) and finds the sampling instance *i* at which maximum residue is generated (Line $3-4$). Considering a staircase approximation of the target variable threshold vector, we maintain the vector Steps to keep track of the heights of the step edges of the staircase where $Steps[k]$ denotes height of the k^{th} step. In this algorithm, a step captures a subsequence of consecutive constant thresholds. First step of staircase is created by setting $\forall 1 \leq j \leq i$, $Th[j] = Steps[i]$ where $Steps[i] = ||z_i||$ (Line 5 – 6). With this new threshold vector Th , we call Algorithm [1](#page-1-0) to check if any attack can bypass this detector. If yes, we generate new threshold steps in the following ways.

Input Performance criteria pfc , existing monitoring constraint mdc , number of sam- pling instances T required by the controller to attain pfc		
Output Threshold Vector $Th = \{ Th[i]:$ threshold required at i^{th} sampling instance		
to thwart false data injection attack }		
1: function STEPWISETHRESHOLDSYN(pfc , mdc , T)		
2: $Th \leftarrow NULL; Steps \leftarrow NULL;$	\triangleright Initialization	
3: if $ATTVECSYN(Th, pfc, mdc, T)$ then		
4: $\exists i \in [1, T] \forall k \in [1, T] \; i \neq k \wedge z_i \geq z_k $		
5: $Steps[i] \leftarrow z_i ;$		
6: $\forall 1 \leq j \leq i$, $Th[j] \leftarrow Steps[i]; k \leftarrow i;$		
7: while ATTVECSYN(Th, pfc, mdc, T) $\wedge k \neq T$ do \triangleright Initial steps formation		
8: $\exists i \ i < T \land Th[i] \neq 0 \land Th[i+1] = 0;$		
9: $\exists k \ i < k < T \ \forall j \ i < j < T \ \land j \neq k \ \land z_j \leq z_k \leq Th[i];$		
10: $Steps[k] \leftarrow z_k ;$		
11: $\forall i \leq j \leq k \ Th[j] \leftarrow Steps[k];$		
12: while $A \leftarrow$ ATTVECSYN(<i>Th</i> , <i>pfc</i> , <i>mdc</i> , <i>T</i>) do \triangleright Reducing height of steps		
13: $k \leftarrow$ MINAREARECTANGLE (A, \mathcal{Steps}, T) ;		
14: $\exists p \; k \; q \; p \; q \; \mathbb{Y}$ $q \; k \; q \; q \; \mathbb{Y}$ \land $p \neq q \land$ $Steps[p] \; \leq \; z_k \land$		
$Steps[q] \leq z_k \wedge Steps[p] \geq Steps[q];$		
15: $Steps[k] \leftarrow Th[k]; Steps[p] \leftarrow z_k ;$		
16: $\forall i \ k < i \leq p$, $Th[i] \leftarrow Steps[p]$;		
17: return Th		
18: function MINAREARECTANGLE($A, Steps', T$)		
19: $MinArea \leftarrow \infty$; $MinAreaPosition \leftarrow NULL$;		
20: for $i = 1$ to T do		
21: $Area_i \leftarrow 0$		
22: while $\exists j \ i \leq j \leq T \ \forall k \ i \leq k \leq T \ \land j \neq k \land \text{Steps}'[j] >$		
$Steps'[k] \wedge Steps'[j] > z_i $ do;		
23: $Area_i \leftarrow Area_i + (Steps'[j] - z_i) \times (j - i);$		
24: $Steps'[j] \leftarrow NULL;$		
25: if $Area_i$ < $MinArea$ then		
26: $MinArea \leftarrow Area; MinAreaPosition \leftarrow i;$		
27: return MinAreaPosition:		

Algorithm 3: Step-wise Threshold Synthesis

Case 2a [Line $7 - 11$]: Let i be the last step with non-zero threshold (Line 8). To generate a new step after i , we find out the sampling instance k at which the maximum residue is generated by the current attack vector such that $k > i$. The record of new step edge $Steps[k] = ||z_k||$ is added to Steps vector (Line 10) and the new step is enforced by setting $\forall j \in (i, k], Th[j] = Steps[k]$ (Line 11). If the last step edge is at $i = T$ or no stealthy attack can be found that bypasses the current threshold steps, we proceed to *Case* 2b to build new steps by fine-graining the existing ones.

Case 2b [Line $12 - 17$]: In this case, we have two possibilities. If no attack vector exists (Line 12), then the algorithm terminates. If any attack is found that bypasses the current detector threshold Th , heights of the existing steps need to be reduced. Instead of diminishing the height of an entire step, we break a portion or the whole step whichever involves minimum effort i.e. the minimum area from under the threshold curve that can be removed to detect the current attack. The function MINAREARECTANGLE (Line $18 - 27$) computes such minimum area ensuring both staircase like structure and monotonic decreasing property.

IV. CASE STUDY AND OBSERVATIONS

We demonstrate the efficacy of our approach using a Vehicle Stability Controller (VSC) case study. The VSC system receives data from four wheel speed sensors (WSS), lateral acceleration (Ay), longitudinal acceleration (Ax), yaw rate sensor (Yrs) and steering angle sensor (SaS). Generated actuator command is sent to the hydraulic unit of a vehicle. Wheel speed sensors are hardwired between the wheels and the controller unit. However, data from Ay, Ax, Yrs, SaS, along with actuator signal, are transferred through CAN bus and is considered vulnerable to attack. In this work, we use VSC model of [\[10\]](#page-3-6). Sampling period is considered as $T_s = 40ms$. Relevant variables are taken from [\[11\]](#page-3-7). We consider an attack model where the attacker forges output of both Yrs and Ay sensors. However, most modern automobiles have monitoring systems already in place to detect any abnormal behavior of VSC. We consider one such monitoring system which performs the following checks for all measurements: 1) *Range and gradient based monitors* check if range and gradient of yaw rate γ and lateral acceleration a_y are within permissible limit; 2) *Relation based monitor* checks if difference between measured yaw rate γ from Yrs and estimated yaw rate γ_{est} from Ay is less than $allowedDiff$. An immediate violation of both the schemes does not raise an alarm. It waits for certain duration, called *dead zone*. Continuous violation during the dead zone causes the monitoring system to raise an alarm. The allowedDiff, range of γ , gradient of γ , range of a_y and gradient of a_y are considered 0.035 rad/s, 0.2 rad/s, 0.175 rad/s², 15 m/s² and 2 m/s³ respectively. The dead zone is considered to be 300 ms i.e. $\left\lfloor \frac{300}{T_s} \right\rfloor = 7$ samples. We define pfc of the system as: yaw rate must reach within 80% of desired value within 50 sampling instances.

Fig. 2: Attack demonstration on VSC

To verify whether this apparently efficient monitoring system can be bypassed by an attacker, we formulate an SMT problem in Algorithm [1.](#page-1-0) We model all monitors as predicate mdc in Algorithm [1.](#page-1-0) We include pfc and mdc in the assertion clause A (Line 9 of Algorithm [1\)](#page-1-0) and use the popular SMT solver $Z3$ [\[12\]](#page-3-8). The output array A , in Algorithm [1,](#page-1-0) if nonempty, reports attack vectors for the system. The effect of one such synthesized vector for the VSC system is demonstrated in Fig. [2a](#page-3-9). The attack bypasses the existing monitoring

schemes (Fig. [2b](#page-3-9),c). For mitigating these vulnerabilities, we synthesize suitable residue based detectors using our methods.

With pfc , mdc of VSC and T as input, we execute Algorithms [2](#page-2-0) and [3](#page-2-1) with a timeout of 12 hours for each SMT call. Based on the greedy choices made during simulation, Algorithm [2](#page-2-0) terminates in the 56^{th} round while Algorithm [3](#page-2-1) terminates much faster, in the

Fig. 3: Output of variable threshold synthesis Algorithms

 $37th$ round. The final threshold sets computed by both algorithms are presented in Fig. [3.](#page-3-10) For comparison purpose, we also synthesize a static threshold based detector for VSC. We generate 1000 random measurement noise vectors of bounded length with each value sampled from a suitably small range such that pfc is maintained. Among these, we discard the noise vectors that are detected by mdc. From the remaining, we compute false alarm rate of the three threshold based detectors as: a) 61.5% for Algorithm [2,](#page-2-0) b) 45.6% for Algorithm [3,](#page-2-1) and c) 98.9% for static threshold based detector. We can see that both our proposed algorithms outperform static threshold based detector in terms of FAR.

V. CONCLUSION

In the present work, we provide a synthesis mechanism for variable threshold based detectors in the context of securing CPS implementations. Our approach, based on formal techniques, can provide provable guarantees for an actual controller implementation instead of probabilistic guarantees as is standard for mathematical control models. In future, we would like to perform more exhaustive experimental as well as analytical evaluation of our proposed techniques.

REFERENCES

- [1] S. Karnouskos, "Stuxnet worm impact on industrial cyber-physical system security," in *IECON*. IEEE, 2011.
- [2] J. Slay *et al.*, "Lessons learned from the maroochy water breach," in *ICCIP*. Springer, 2007.
- [3] R. Khan *et al.*, "Threat analysis of blackenergy malware for synchrophasor based real-time control and monitoring in smart grid." *ICS-CSR*, vol. 16, pp. 1–11, 2016.
- [4] Y. Shoukry *et al.*, "Non-invasive spoofing attacks for anti-lock braking systems," in *CHES*. Springer, 2013.
- [5] I. Jovanov *et al.*, "Secure state estimation with cumulative message authentication," in *CDC*. IEEE, 2018.
- [6] Y. Liu, *et al.*, "False data injection attacks against state estimation in electric power grids," *ACM Transactions on Information and System Security*, vol. 14, no. 1, p. 13, 2011.
- [7] H. Sandberg *et al.*, "On security indices for state estimators in power networks," in *SCS, Stockholm*, 2010.
- [8] Y. Mo and B. Sinopoli, "False data injection attacks in cyber physical systems," in *SCS, Stockholm*, 2010.
- [9] A. J. Kerns *et al.*, "Unmanned aircraft capture and control via gps spoofing," *Journal of Field Robotics*, vol. 31, no. 4, pp. 617–636, 2014.
- [10] Y. Aoki *et al.*, "Experimental demonstration of body slip angle control based on a novel linear observer for electric vehicle," in *IECON*. IEEE, 2005.
- [11] S. Zheng *et al.*, "Controller design for vehicle stability enhancement," *Control Engineering Practice*, vol. 14, no. 12, pp. 1413–1421, 2006.
- [12] L. De Moura *et al.*, "Z3: An efficient smt solver," in *TACAS*. Springer, 2008.