Secure State Estimation with Cumulative Message Authentication

Ilija Jovanov and Miroslav Pajic

***Abstract*— With network-based attacks, such as Man-in-the-Middle (MitM) attacks, the attacker can inject false data to force a closed-loop system into any undesired state, unless even intermittently integrity of delivered sensor measurements is enforced. Yet, the use of standard cryptographic techniques that ensure data integrity, such as Message Authentication Codes (MACs), introduces significant communication and computation overhead. Thus, in this work we explore the use of cumulative MACs that significantly reduce network overhead. We consider systems with Kalman filter-based state estimators and sequential probability ratio test (SPRT) intrusion detectors. We show that strong estimation guarantees under MitM attacks can be obtained even with intermittent use of a *single* cumulative MAC that is added to appropriate sensor measurements transmitted over the network. We present a design-time methodology to evaluate the effects of any given cumulative integrity enforcement policy on reachable state-estimation errors for any type of stealthy attacks; this provides a base for design of cumulative enforcement policies with desired performance guarantees even in the presence of MitM attacks. Finally, we illustrate the effectiveness of our approach on an automated steering control.**

I. Introduction

Techniques for securing cyber-physical systems (CPS) have been mostly limited to either making the physical layer inaccessible, or to the use of cyber methods to secure communication over the underlying networks. The latter mainly include cryptographic mechanisms that ensure data confidentiality and/or integrity, such as message authentication codes (MACs) that are commonly employed in embedded control networks. On the other hand, the physical layer provides opportunities for the use of physics-based attack-detection and resilient control techniques; recent research efforts in this domain have focused on the utilization of a physical model of the system to improve security guarantees with a design of attack-resilient controllers or state estimators (e.g., [1], [16], [15]), as well as intrusion detectors (e.g., [10], [5], [6], [9]).

In this work, we focus on a standard Kalman-filter based control architecture that employs a residual-based intrusion detector; such detectors were introduced in the literature through χ^2 detectors (e.g., [10], [5], [3]), as well as sequential probability ratio test (SPRT) (e.g., [6], [4]) and CUSUM detectors (e.g., [2], [18]). Specifically, we consider the problem of stealthy false-data injection attacks on measurements provided by system sensors, commonly performed as network-based *Man-in-the-Middle* (MitM) attacks. For a large class of physical plants, it was shown that in such scenarios these attacks could introduce unbounded errors in state estimation while avoiding detection [10], [5]; this effectively allows a stealthy attacker to freely control the system.

This work was supported in part by Intel and the NSF CNS-1652544 and CNS-1505701 grants. It is also based on research sponsored by the ONR under agreements number N00014-17-1-2012 and N00014-17-1-2504.

I. Jovanov and M. Pajic are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA 27708. Email: {ilija.jovanov, miroslav.pajic}@duke.edu

We have recently shown that by combining such physics-based intrusion detectors with even intermittent use of cryptographic mechanisms that ensure data integrity, we can significantly limit the attacker's impact on the system [3], [2] – i.e., either the attacks are detected or the resulting estimation error is bounded. We consider systems that could utilize data authentication (e.g., MACs) but have limited resources, such as CAN-based automotive systems. In such scenarios, resource constraints are twofold – network bandwidth is usually tightly designed for the number of messages that are being transmitted [7], and computational power of CPUs used for MAC signing/checking is limited [8]. While intermittent integrity enforcements significantly reduce resources required to add security to existing systems, in most of these systems network bandwidth is the bottleneck [7], especially for certain classes of systems as the required number of authenticated packets depends on plant dynamics [2].

On the other hand, the use of *cumulative* (sometimes referred to as "delayed") message authentication schemes reduces the network load by transmitting a MAC for multiple data points in a single package [14]. Consequently, in this paper we extend our work from [3], [2], and provide theoretical basis for the intermittent use of cumulative data authentication that maintains the strong security guarantees. We also introduce a computationally efficient method to evaluate attacker's influence over time, which is then used to design and evaluate cumulative integrity enforcement policies that ensure low estimation errors even in the presence of attacks.

The paper is organized as follows. In Section II, we describe the system and attack models, and formalize the effects of cumulative data integrity enforcement policies on state estimation guarantees. Section III provides a design-time method to efficiently estimate worst-case bounds on performance degradation under stealthy attacks as well as evaluate effects of different cumulative integrity enforcement policies. Finally, Section IV illustrates the design and use of such authentication policies on an automotive case-study, before we provide concluding remarks in Section V.

Notation and Terminology: Support set of a vector $\mathbf{x}$, i.e., indices of all non-zero values of $\mathbf{x}$, is denoted by $supp(\mathbf{x})$, while sup stands for supremum. Moore-Penrose inverse of a matrix $\mathbf{A}$ is denoted by $\mathbf{A}^{\dagger}$. Zero vector of appropriate size is represented by $\mathbf{0}$, and empty set is denoted by $\emptyset$. We use $\mathbf{i}_j^T$ to denote the row vector with all elements equal to 0 except of j-th element that is equal to 1. For a set of all sensors $\mathcal{S} = \{s_1, ..., s_p\}$, projection matrix $\mathbf{P}_{\mathcal{K}}$ of set $\mathcal{K} \subseteq \mathcal{S}$ is $\mathbf{P}_{\mathcal{K}} = \left[\mathbf{i}_{k_1} | \ldots | \mathbf{i}_{k_{|\mathcal{K}|}}\right]^T$, where $k_1 < \cdots < k_{|\mathcal{K}|}$ – i.e., $\mathbf{P}_{\mathcal{K}}\mathbf{y}$ retains elements of vector $\mathbf{y}$ with indices in $\mathcal{K}$.

Finally, we denote probability of an event as $P(\cdot)$, while $E[\mathbf{x}]$ is the expected value of a random vector $\mathbf{x}$. Also, $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{A})$ denotes that vector $\mathbf{x}$ is a random variable with the zero-mean normal distribution with variance matrix $\mathbf{A}$.

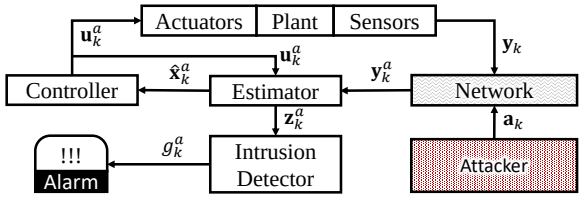

Fig. 1. System architecture – sensor measurements are communicated over a potentially compromised network to the estimator. We assume that the attacker can access and modify the sensor data transmitted over the network.

II. System Model

We consider a common networked control architecture (Fig. 1) where plant sensor measurements are transmitted over a potentially compromised network. The plant is modeled as an observable linear-time invariant system of the form

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k + \mathbf{w}_k, \quad \mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{v}_k, \tag{1}$$

where $\mathbf{x}_k \in \mathbb{R}^n$ and $\mathbf{y}_k \in \mathbb{R}^p$ denote the plant state vector and sensor measurements obtained from the set of sensors $\mathcal{S} = \{s_1, \ldots, s_p\}$ at time k, respectively. Also, we assume that the initial state $\mathbf{x}_0$, modeling noise $\mathbf{w}_k$, and sensor noise $\mathbf{v}_k$ are Gaussian random variables. Finally, by $\mathbf{R}$ we denote the covariance matrix of measurement noise vector $\mathbf{v}_k$.

Furthermore, a Kalman filter is used for state estimation. We assume the system has been running for some time before the first attack occurs; thus, the Kalman filer has reached a steady state and can be considered as a fixed gain estimator. Hence, the filter's state estimate, denoted by $\hat{\mathbf{x}}_k$, evolves as

$$\begin{aligned} \hat{\mathbf{x}}_k &= \mathbf{A}\hat{\mathbf{x}}_{k-1} + \mathbf{B}\mathbf{u}_{k-1} + \mathbf{K}\mathbf{z}_k \\ \mathbf{K} &= \mathbf{\Sigma}\mathbf{C}^T(\mathbf{C}\mathbf{\Sigma}\mathbf{C}^T + \mathbf{R})^{-1} \\ \mathbf{z}_k &= \mathbf{y}_k - \mathbf{C}(\mathbf{A}\hat{\mathbf{x}}_{k-1} + \mathbf{B}\mathbf{u}_{k-1}), \quad \mathbf{e}_k = \mathbf{x}_k - \hat{\mathbf{x}}_k. \end{aligned} \tag{2}$$

Here, matrix $\mathbf{K}$ represents the Kalman gain and $\mathbf{e}_k$ denotes the state estimation error. In addition, residue $\mathbf{z}_k$ captures the discrepancy between received sensor measurements and the values estimated by the model. As such, $\mathbf{z}_k$ is commonly used for intrusion and anomaly detection (e.g., [2], [3], [10], [12]). Specifically, a detection function g_k is defined and alarm is triggered when $g_k > h$, where h is some predefined detection threshold. Note that even without attacks there is a probability that an alarm will be triggered due to noise. We denote such probability (i.e., of false alarm) at time k as

$$\beta_k = P(g_k > threshold). \tag{3}$$

These detectors are commonly designed to operate with a low false-positive rate, and thus when the system is not under attack, the value of β_k should be low. While this is a general setup to capture intrusion detectors, more details about the considered intrusion detector is provided in Section II-B.

A. Attack Model

We assume that the network has been compromised; the attacker could launch MitM attacks that may change the delivered messages from a subset of sensors $\mathcal{K} \subseteq \mathcal{S}$ when no security mechanisms are employed to protect integrity of the transmitted packets. We model this influence as additive disturbance $\mathbf{a}_k$ on sensors from $\mathcal{K}$, and thus in the presence of MitM attacks the system evolves as

$$\begin{aligned} \mathbf{x}_{k+1}^a &= \mathbf{A}\mathbf{x}_k^a + \mathbf{B}\mathbf{u}_k^a + \mathbf{w}_k \\ \mathbf{y}_k^a &= \mathbf{C}\mathbf{x}_k^a + \mathbf{a}_k + \mathbf{v}_k, \end{aligned} \tag{4}$$

where $\mathbf{a}_k$ is a sparse vector with support in $\mathcal{K}$ (i.e., $supp(\mathbf{a}_k) = \mathcal{K}$). Here, we assume that when no data integrity mechanisms are used, such as MACs, the nonzero injected signals can have any values. On the other hand, for all time steps k when MACs are used, $\mathbf{a}_k$ has to be equal to zero. Finally, note that in (4) and throughout the paper, we use superscript "a" to denote the evolution of a compromised system – i.e., the corresponding system values in the presence of attacks. For instance, by $\hat{\mathbf{x}}_k^a$ and $\mathbf{z}_k^a$ we denote the Kalman filter estimate and residue from (2) in the presence of attacks, and thus the state estimation error is $\mathbf{e}_k^a = \mathbf{x}_k^a - \hat{\mathbf{x}}_k^a$.

To formally capture attack impact, we focus on the difference between the compromised and non-compromised system's evolution. Specifically, we consider the difference in estimation error and filter residue due to attacks, defined as

$$\begin{aligned} \Delta\mathbf{e}_k &= \mathbf{e}_k^a - \mathbf{e}_k \\ \Delta\mathbf{z}_k &= \mathbf{z}_k^a - \mathbf{z}_k. \end{aligned} \tag{5}$$

By adding the definitions of $\mathbf{e}_k, \mathbf{e}_k^a, \mathbf{z}_k, \mathbf{z}_k^a$ from (2) into (5), it follows that

$$\begin{aligned} \Delta\mathbf{e}_{k+1} &= (\mathbf{A} - \mathbf{KCA})\Delta\mathbf{e}_k - \mathbf{K}\mathbf{a}_{k+1} \\ \Delta\mathbf{z}_{k+1} &= \mathbf{CA}\Delta\mathbf{e}_k + \mathbf{a}_{k+1}. \end{aligned} \tag{6}$$

We show later in paper (see (30)), that when attacks start after the system has reached the steady state it holds that $\Delta\mathbf{e}_k = E[\mathbf{e}_k^a]$. Thus, the no-noise dynamical system from (6) captures the evolution of the expected state estimation error due to MitM attacks. This allows us to formalize the attacker's capabilities and goals. Specifically, while designing attack vectors $\mathbf{a}_k$, we assume a very powerful attack model where:

1) The attacker has the full system knowledge. This includes knowledge of the system dynamics and employed controllers (i.e., estimator and intrusion detector), as well as the real-time information about the values of all signal values in all the sensors and controller.
2) The attacker possesses unlimited computational power. This allows the attacker to plan and inject attacks in real-time without consideration for computational complexity of the problem.

The attacker's main goal is to maximize the change in the state estimation due to false-data injection attacks. Thus, since $\Delta\mathbf{e}_k$ captures the expected discrepancy between the estimated and real states of the plant, the attacker's goal can be captured as maximization of $\Delta\mathbf{e}_k$ until the error $\Delta\mathbf{e}_k \in \mathcal{R}_{risk}$ is reached; here, $\mathcal{R}_{risk} \subseteq \mathbb{R}^n$ denotes a set of estimation errors that may permanently damage the system.

Additional goal is to make the attack undetectable by the employed intrusion detector – i.e., to inject a *stealthy* MitM attack that maintains the probability of detection β_k^a low, very close to the false positive alarm probability β_k in the case without attacks. Thus, the *stealthiness constraint* is specified

$$\beta_k^a \leq \beta_k + \varepsilon, \tag{7}$$

where ε is a small positive constant relative to β_k that captures the largest inconspicuous deviation of alarm probability due to attacks. It is important to note that ε thus implicitly denotes the amount of risk that the attacker is willing to take. In practical use, this will allow for trade-off analysis from the

attacker's perspective, between tolerable risks and the impact he can have on the system, as we illustrated in [3], [2] for systems employing standard MACs to ensure data integrity.

B. Intrusion Detector

In this paper, we assume that a commonly utilized sequential probability ratio test (SPRT) is used for intrusion detection, as in [6], [4]; however, our analysis can be easily extended to other detectors such as the CUSUM detector using the techniques from [2]. Hence, we define

$$g_k = g_{k-1} + \Lambda_k;\ \Lambda_k = \log\left(\frac{f_a(\mathbf{z}_k)}{f(\mathbf{z}_k)}\right), \tag{8}$$

where $f_a(\mathbf{z}_k)$ and $f(\mathbf{z}_k)$ denote the probability density functions for $\mathbf{z}_k^a$ and $\mathbf{z}_k$ respectively. This detector attempts to distinguish between two hypothesis – $\mathcal{H}_0 : \mathbf{z}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$, and $\mathcal{H}_1 : \mathbf{z}_k \not\sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$. To formally capture the behavior of the detector, we first address the following two challenges.

First, SPRT-based detectors in general employ two decision thresholds – a lower and an upper threshold. When detection function g_k is below the lower decision threshold, it is assumed that the system is not compromised, while when g_k is between the thresholds it is undecided whether the system has been compromised; finally, when g_k is above the upper threshold h, an alarm is triggered. In addition, after a decision is reached, g_k resets to zero to avoid infinite summation. Without affecting generality of our results, to simplify our analysis we assume that for stealthy attacks this reset never occurs. Note that our analysis is focused on providing guarantees even for worst-case scenarios with maximal deviation in state estimation errors $\Delta\mathbf{e}_k$. Hence, this is a valid assumption as maximization of $\Delta\mathbf{e}_k$ would result in 'largest' attack vectors $\mathbf{a}_k$ that do not trigger the stealthiness condition (7); since from (4) and (2), increase in the size of $\mathbf{a}_k$ would in general result in larger g_k, optimal attacks would effectively result in increased (as much as possible without violating (7)) values for g_k.

The second challenge is the SPRT assumption that $f_a(\mathbf{z}_k)$ is known. For realistic attack scenarios, we cannot impose such unrealistic constraints, and thus the attacker is modeled as a non-deterministic rather than probabilistic signal. This in turn limits our capability to exactly compute Λ_k. We resolve this challenge by introducing an approximation similar to the one from [6], [4] – i.e., we specify that

$$\Lambda_k = \frac{1}{2}\mathbf{z}_k^T\mathbf{Q}^{-1}\mathbf{z}_k + \log c\sqrt{(2\pi)^p det(\mathbf{Q})}, \tag{9}$$

where $c = e^{\frac{-p}{2}}/\sqrt{(2\pi)^p det(\mathbf{Q})}$. The above equation uses the idea similar to χ^2 detectors [11], [13], [10] and allows us to argue about the attack through statistic properties of $\mathbf{z}_k$ while retaining zero-centered g_k. Therefore, from (8) and (9) along with the assumption of non-resetting SPRT, we obtain the following form of the SPRT detection function

$$\begin{aligned} g_k &= \sum_{\mathcal{T}=1}^{k}\left(\frac{1}{2}\mathbf{z}_{\mathcal{T}}^T\mathbf{Q}^{-1}\mathbf{z}_{\mathcal{T}} + \log c\sqrt{(2\pi)^p det(\mathbf{Q})}\right) = \\ &= \frac{1}{2}\sum_{\mathcal{T}=1}^{k}\left(\mathbf{z}_{\mathcal{T}}^T\mathbf{Q}^{-1}\mathbf{z}_{\mathcal{T}}\right) + k\log c\sqrt{(2\pi)^p det(\mathbf{Q})}, \end{aligned} \tag{10}$$

or rather

$$g_k = \frac{1}{2}\sum_{i=1}^{k}(\mathbf{z}_i^T\mathbf{Q}^{-1}\mathbf{z}_i) - \frac{kp}{2}, \tag{11}$$

As described before, the probability of false positives detected by SPRT can now be defined as $\beta_k = P(g_k > h)$.

Finally, we can now define the set of stealthy attacks.

Definition 1. *[2] Let us denote $\mathbf{a}_{1..k} = [\mathbf{a}_1^T \dots \mathbf{a}_k^T]^T$. Then, the set of all stealthy attacks up to time k is defined as*

$$\mathcal{A}_k = \{\mathbf{a}_{1..k} | \beta_{k'}^a \leq \beta_{k'} + \varepsilon,\ \forall k', 1 \leq k' \leq k\}. \tag{12}$$

C. Cumulative Message Authentication

When standard MACs are used, at every time-instance when integrity of a sensor measurement is enforced, the corresponding MAC for the measurement *only* is computed and attached to the transmitted packet. Thus, we define standard intermittent integrity enforcement policies as follows.

Definition 2. *[3], [2] A standard intermittent data authentication policy (μ_s, f_s, L_s), where $\mu_s = \{t_j\}_{j=0}^{\infty}$ with $t_0 \geq f_s$, and for all $j > 0$, $t_{j-1} < t_j$, such that $L_s = \sup_{j>0}(t_j - t_{j-1})$, ensures that*

$$\mathbf{a}_{t_j} = \mathbf{a}_{t_j-1} = ... = \mathbf{a}_{t_j-f_s+1} = \mathbf{0}, \forall j \geq 0.$$

Intuitively, a standard intermittent data authentication policy ensures that an attack does not influence sensor values (i.e., $\mathbf{a}_k = \mathbf{0}$) within pre-specified time windows of f sample points when data integrity is enforced; the end-points of these windows are captured in the sequence μ_s and all the windows (i.e., their beginnings or ends) are separated by at most L samples. In [2], we showed that for a very general class of intrusion detectors, when a standard intermittent data authentication policy is used with $f = \min(\psi, q_{un})$, where ψ denotes the observability index of the $(\mathbf{A}, \mathbf{C})$ pair and q_{un} denotes the number of unstable eigenvalues of $\mathbf{A}$, the attacker cannot introduce unbounded state estimation errors while remaining stealthy. This also allows us to commonly use f that is directly determined from the plant's dynamics.

On the other hand, with *cumulative integrity enforcement policies*, blocks of f_c measurements are used to compute a single MAC (thus, the term *cumulative MACs*), which is then attached to communication packets containing the last measurements from these blocks. Hence, when a cumulative MAC, which is computed over f_c last time-points, is received at time t_j, the controller will be able to detect whether false-data has been injected in the last f_C transmissions. To formalize this notion, we introduce the following definition.

Definition 3. *An intermittent cumulative data authentication policy (μ_c, f_c, L_c) at time (i.e., sample point) k, such that $\mu_c = \{t_j\}_{j=0}^{\infty}$ where $t_0 \geq f_c$, and for all $j > 0$, $t_{j-1} < t_j$, with $L_c = \sup_{j>0}(t_j - t_{j-1})$, ensures that*

$$\mathbf{a}_t = \mathbf{a}_{t-1} = ... = \mathbf{a}_{t-f_c+1} = \mathbf{0}, \forall t \in \mathcal{D}_k, \tag{13}$$

where $\mathcal{D}_k = \{t_j \mid t_j \in \mu_c \text{ and } t_j \leq k\}$.

Unlike when standard MACs are used as in Definition 2, a cumulative data authentication policy as defined above, denies the attack impact retroactively, only after a cumulative MAC, which is computed over a block of f_c consecutive

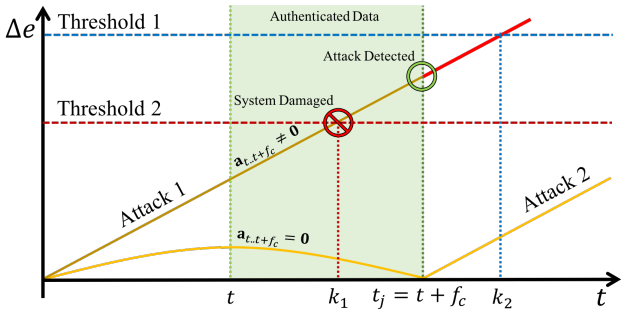

Fig. 2. Stealthy attacks depending on the threshold of the $\mathcal{R}_{risk}$ region. Cumulative MAC is assumed to arrive at $t_j = t + f_c$. When Threshold 1 describes $\mathcal{R}_{risk}$ region, "Attack 1" is detected after cumulative MAC is received at t_j without reaching $\mathcal{R}_{risk}$, and thus attacker has to perform "Attack 2" to remain stealthy. On the other hand, when $\mathcal{R}_{risk}$ boundary is Threshold 2, the attack reaches $\mathcal{R}_{risk}$ states before authentication, and successfully damages the system with "Attack 1" before it is detected.

measurements, is received. Thus, if the attacker cannot reach a desired error by the time a cumulative MAC arrives, he should not insert false data during the time points used to compute the MAC; otherwise, modified data will not pass the check when the MAC arrives and the attack will be detected.

To illustrate this, consider Fig. 2 that shows two attacks – one that attempts to stay stealthy only prior to time $t_j = t+f_c$ ("Attack 1") and the other that remains stealthy after ("Attack 2"). Until time t_j, authenticity of sensor data between t and t_j cannot be valideted, and thus false-data vectors do not have to be zero, as captured in (13). In this case, if $\mathcal{R}_{risk}$ threshold is $Threshold$ 2, Attack 1 will force the system into $\mathcal{R}_{risk}$, and thus the attacker does not need to be concerned with authentication checking at time t_j (i.e., that he would be detected then). However, if $\mathcal{R}_{risk}$ threshold is $Threshold$ 1, Attack 1 would be unable to reach $\mathcal{R}_{risk}$ by t_j, and thus needs to continue after. Yet, with the arrival of cumulative MAC at t_j the attack will be detected. Therefore, in this case Attack 2 should be used, since this strategy remains stealthy even with the newly considered integrity enforcement policy.

To capture constraints cumulative authentication imposes on the attacker, let us define the attack vector's support set for a standard policy (μ, f, L) according to Definition 2 as

$$\tilde{\mathcal{K}}_j = \begin{cases} \emptyset, & j - i \in \mu, \text{ for some } i,\ 0 \leq i < f, \\ \mathcal{K}, & \text{otherwise} \end{cases} \tag{14}$$

From Definition 3, it follows that an intermittent cumulative data authentication policy (μ_c, f_c, L_c) at time k changes the support set of a stealthy attack $\mathbf{a}_{1..k}$ as

$$supp(\mathbf{a}_j) = \begin{cases} \tilde{\mathcal{K}}_j, & \text{for some } j, j \leq k - f, \\ \mathcal{K}, & \text{otherwise} \end{cases} \tag{15}$$

Finally, for a stealthy attack $\mathbf{a}_{1..k}$ from Definition 1, we denote its support set as $supp(\mathbf{a}_{1..k}) = \mathcal{Q}_k \subseteq \{1, \ldots, pk\}$.

III. Reachability Analysis for Systems with Intermittent Cumulative Integrity Enforcement

In this section, we introduce a method to compute the set of all reachable estimation errors caused by a stealthy attack that are obtainable with probability at least η. Parameter η exists to limit values of $\mathbf{e}_k$ to realistically obtainable values, as $\mathbf{e}_k$ is a Gaussian random variable and thus takes the values from an unbounded set. Set of reachable estimation errors allows us to estimate the effects of a stealthy attack on the system, as it captures deviation from the steady state of the system. We refer to this set as the k-reachable region of the state estimation error, and define it as follows.

Definition 4. *The k-reachable region $\mathcal{R}_k$ of the state estimation error under the attack (i.e., $\mathbf{e}_k^a$) is the set*

$$\mathcal{R}_k = \left\{ \mathbf{e} \in \mathbb{R}^n \;\middle|\; \begin{array}{l} \mathbf{e}\mathbf{e}^T \preccurlyeq E[\mathbf{e}_k^a]E[\mathbf{e}_k^a]^T + \gamma\Sigma, \\ \mathbf{e}_k^a = \mathbf{e}_k^a(\mathbf{a}_{1..k}),\ \mathbf{a}_{1..k} \in \mathcal{A}_k \end{array} \right\}, \tag{16}$$

where $\Sigma = Cov(\mathbf{e}_k^a)$, and for a cumulative distribution function $F_e(\cdot)$ of $\mathbf{e}_k$, we can relate γ to η as

$$\eta = F_e(\gamma Cov(\mathbf{e}_k)) - F_e(-\gamma Cov(\mathbf{e}_k)).$$

Note that $Cov(\mathbf{e}_k^a) = Cov(\mathbf{e}_k)$, with $\mathbf{e}_k$ being Gaussian (the full proof is available in [2]). To find an analytic solution for these regions, we use the following lemma.

Lemma 1. *[2] For a system with a detection function*

$$g_k^a = \sum_{i=1}^{k} c_i \mathbf{z}_i^T \mathbf{Q}^{-1} \mathbf{z}_i,$$

the set of all stealthy attacks at time k, $\mathcal{A}_k$, can be overapproximated by the set

$$\overline{\mathcal{A}}_k = \{\mathbf{a}_{1..k} | \|\Delta \mathbf{z}_{k'}\|_{\mathbf{Q}^{-1}} \leq \overline{\alpha},\ \forall k', 1 \leq k' \leq k\} \tag{17}$$

(i.e., $\overline{\mathcal{A}}_k \supseteq \mathcal{A}_k$), where c_i are non-negative constants, $\overline{\alpha} = \alpha_{\chi^2}(\varepsilon, p, h/c_{max})$ with $c_{max} = \max(c_1, ..., c_k)$, and α_{χ^2} defines an upper bound of $\|\mathbf{z}_k\|_{\mathbf{Q}^{-1}}$ when χ^2 detector is utilized instead of the SPRT.

The above lemma allows us to obtain an analytic solution for k-reachable regions under a specific cumulative data authentication policy, as specified in the following theorem.

Theorem 1. *The k-reachable region $\mathcal{R}_k$ under an intermittent cumulative data authentication policy (μ_c, f_c, L_c) can be overestimated as*

$$\overline{\mathcal{R}_k} = \left\{ \mathbf{e}_k^a | \mathbf{e}_k^a \mathbf{e}_k^{aT} \preccurlyeq \alpha_{\chi^2}^2 \mathbf{M}_k P_{\mathcal{Q}_k}^{\dagger} \mathbf{\Theta}_k^{-1} (\mathbf{M}_k P_{\mathcal{Q}_k}^{\dagger})^T + \gamma \mathbf{\Sigma} \right\}, \tag{18}$$

where $\alpha_{\chi^2}^2 = \alpha_{\chi^2}^2(\varepsilon, kp, 2h + kp)$ is an upper bound on $\|\mathbf{z}_k^a\|_{\mathbf{Q}^{-1}}$. Furthermore,

$$\mathbf{\Theta}_k = \sum_{\tau=1}^{k} \mathbf{H}_\tau^T \mathbf{Q}^{-1} \mathbf{H}_\tau \tag{19}$$

$$\mathbf{H}_\tau = \left[\mathbf{N}_\tau \mathbf{P}_{\mathcal{Q}_\tau}^{\dagger} \;\; \mathbf{0}_{p \times (|\mathcal{Q}_k| - |\mathcal{Q}_\tau|)} \right] \tag{20}$$

$$\mathbf{N}_k = \left[\; -\mathbf{C}\mathbf{A}\mathbf{M}_{k-1} \;\middle|\; \mathbf{I} \; \right] \tag{21}$$

$$\mathbf{M}_k = -\left[\; (\mathbf{A} - \mathbf{K}\mathbf{C}\mathbf{A})^{k-1}\mathbf{K} \;\middle|\; \ldots \;\middle|\; \mathbf{K} \; \right]. \tag{22}$$

Similar results can be found in [2] and [6]. However, there are two essential differences that separate these theorems. In [2], data is authenticated instantaneously, causing the attacker to plan ahead and providing limits even before authentication occurs. On the other hand, in [6], authors do not consider data authentication, and consider $E[g_k^a] > h$ rather than $P(g_k^a) > h$ as the stealthiness constraint.

Before we provide the theorem's proof, we introduce the following lemma.

Lemma 2. *For any $k \geq 1$, the matrix $\mathbf{\Theta}_k$ is positive definite.*

The proof is based on the idea that ${\mathbf{P}^{\dagger}_{\mathcal{Q}_1}}^T \mathbf{P}^{\dagger}_{\mathcal{Q}_1} = \mathbf{I}_{|\mathcal{Q}_1|}$ and that elements of the sum from (19) are orthogonal. We omit the proof due to space constraints since it follows the approach used to show a similar result in [2].

Proof of Theorem 1. From (11), it follows that

$$\beta_k = P(g_k > h) = P(\sum_{i=1}^{k} (\mathbf{z}_i^T \mathbf{Q}^{-1} \mathbf{z}_i) > 2h + pk).$$

Using Lemma 1, the stealthiness condition can be overapproximated by

$$\|\Delta \mathbf{z}_k\|_{\mathbf{Q}^{-1}} \leq \alpha_{\chi^2}. \tag{23}$$

We represent (6) in their non-recursive form, and substitute them in (23) to obtain that

$$\alpha^2_{\chi^2}(\varepsilon, kp, 2h + kp) - (\mathbf{P}_{\mathcal{Q}_t} \mathbf{a}_{1..k})^T \mathbf{\Theta}_k \mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k} \geq 0 \tag{24}$$

needs to be satisfied for the attacker to remain stealthy. From Lemma 2, $\mathbf{\Theta}_k$ is positive definite, and thus we can form Schur complement – i.e., the following holds

$$\begin{bmatrix} {\mathbf{\Theta}_k}^{-1} & \mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k} \\ (\mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k})^T & \alpha^2_{\chi^2}(\varepsilon, kp, 2h + kp) \end{bmatrix} \succcurlyeq 0. \tag{25}$$

In order to generalize error computation for any stealthy attack $\mathbf{a}_{1..k}$, let us introduce the matrix

$$\mathbf{G} = \begin{bmatrix} -\mathbf{M}_k \mathbf{P}^{\dagger}_{\mathcal{Q}_k} & \mathbf{0}_{n \times 1} \\ \mathbf{0}_{1 \times |\mathcal{Q}_k|} & 1 \end{bmatrix}. \tag{26}$$

Now, we can consider a quadratic representation

$$\mathbf{G} \begin{bmatrix} {\mathbf{\Theta}_k}^{-1} & \mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k} \\ (\mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k})^T & \alpha^2_{\chi^2}(\varepsilon, kp, 2h + kp) \end{bmatrix} \mathbf{G}^T \succcurlyeq 0, \tag{27}$$

from which it follows that

$$\begin{bmatrix} \mathbf{M}_k \mathbf{P}^{\dagger}_{\mathcal{Q}_k} {\mathbf{\Theta}_k}^{-1} (\mathbf{M}_k \mathbf{P}^{\dagger}_{\mathcal{Q}_k})^T & -\mathbf{M}_k \mathbf{P}^{\dagger}_{\mathcal{Q}_k} \mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k} \\ -(\mathbf{P}_{\mathcal{Q}_k} \mathbf{a}_{1..k})^T (\mathbf{M}_k \mathbf{P}^{\dagger}_{\mathcal{Q}_k})^T & \alpha^2_{\chi^2}(\varepsilon, kp, 2h + kp) \end{bmatrix} \succcurlyeq 0. \tag{28}$$

Once again, by employing Schur complement and the non-recursive form of (6), we obtain

$$\mathbf{M}_k P^{\dagger}_{\mathcal{Q}_k} \mathbf{\Theta}_k^{-1} (\mathbf{M}_k P^{\dagger}_{\mathcal{Q}_k})^T - \frac{1}{\alpha^2_{\chi^2}} \Delta \mathbf{e}_k \Delta {\mathbf{e}_k}^T \succcurlyeq 0. \tag{29}$$

Since $\Delta \mathbf{e}_k$ is a deterministic signal from (6) (i.e., the system modeled by (6) is noiseless), we have that

$$\Delta \mathbf{e}_k = E(\Delta \mathbf{e}_k) = E(\mathbf{e}^a_k) - E(\mathbf{e}_k) = E(\mathbf{e}^a_k). \tag{30}$$

Thus, the first condition in $\mathcal{R}_k$ can be overapproximated by

$$E[\mathbf{e}^a_k] E[\mathbf{e}^a_k]^T + \gamma \Sigma \preccurlyeq \alpha^2_{\chi^2} \mathbf{M}_k P^{\dagger}_{\mathcal{Q}_k} \mathbf{\Theta}_k^{-1} (\mathbf{M}_k P^{\dagger}_{\mathcal{Q}_k})^T + \gamma \mathbf{\Sigma}.$$

Our initial assumption was (23), which generates the attack set $\overline{\mathcal{A}_k}$ from Lemma 1 that overestimates $\mathcal{A}_k$. Thus, the second condition in $\mathcal{R}_k$ that $\mathbf{a}_{1..k} \in \mathcal{A}_k \subseteq \overline{\mathcal{A}_k}$ is satisfied, which concludes the proof. □

Algorithm 1 Deriving periodic integrity enforcement policies with cumulative f steps authentication.

Inputs: System model, set $\mathcal{R}_{risk}$, policy parameter f_c, ε.

```
Enforcement distance L = 0
repeat
    L = L + 1
    Form policy (μ, f, L) as in Theorem 2 with t_0 = L
    Assign k = 0, union of reachable regions R_∪ = ∅, and R̄_0 = 0
    repeat
        R_∪ = R_∪ ∪ R̄_k, k = k + 1
        Compute N_k and M_k from (21) and (22)
        Compute Θ_k from (19)
        Compute α(ε, kp, 2h + kp), and R̄_k from (31)
    until R̄_k ⊆ R_∪
until R̄_k ∩ R_risk ≠ ∅
Accept policy (μ, f, L − 1)
```

If we denote $\mathbf{Y} = \alpha^2_{\chi^2} \mathbf{M}_k P^{\dagger}_{\mathcal{Q}_k} \mathbf{\Theta}_k^{-1} (\mathbf{M}_k P^{\dagger}_{\mathcal{Q}_k})^T + \gamma \mathbf{\Sigma}$, then by using Schur complement one more time, the overapproximation of the k-reachable region $\mathcal{R}_k$ can be specified as

$$\overline{\mathcal{R}_k} = \left\{ \mathbf{e}^a_k | {\mathbf{e}^a_k}^T \mathbf{Y}_k^{-1} \mathbf{e}^a_k \leq 1 \right\}, \tag{31}$$

which is an ellipsoid that can be easily computed.

Finally, we deem cumulative data authentication policy (μ_c, f_c, L_c) to be successful when

$$\overline{\mathcal{R}_k} \cap \mathcal{R}_{risk} = \emptyset, \qquad \forall k \in \mathbb{N}.$$

A policy that satisfies this condition always exists if the system is safe without attacks – e.g., when $f_c = L_c = 1$, meaning that all attacks are prevented by authenticating every message without delay (using standard MACs). Furthermore, we obtain the following result using an approach similar to the one from [2] where standard MACs are intermittently employed; the idea is to map the problem into the problem of bounding the reachable set when standard intermittent integrity enforcement policy $(\mu_c, f_c, L_c + f_c)$ is used – we omit full proof due to space constraints.

Theorem 2. *Consider an LTI system from* (1) *with a intermittent cumulative data integrity policy (μ_c, f_c, L_c), where*

$$f_c = \min(\psi, q_{un}), \tag{32}$$

L_c is finite, ψ is the observability index of the $(\mathbf{A}, \mathbf{C})$ pair, and q_{un} denotes the number of unstable eigenvalues of $\mathbf{A}$. Then for all $k \in \mathbb{N}$, $\mathcal{R}_k$ is bounded.

Although in general an unbounded number of steps needs to be explored, as $\mathcal{R}_k$ will converge to its bounds arbitrarily slow, in practical tests we observed that after some system-dependent time point k_t, $\overline{\mathcal{R}_{k_t+1}} \subseteq \cup_{k=1}^{k_t} \overline{\mathcal{R}_k}$, which terminates the search. Therefore, we can design safe cumulative integrity enforcement policies using Algorithm 1.

IV. Case Study

We demonstrate our method on a steering control case-study [17]. We consider a vehicle that weighs $m = 1573kg$, and measures $l_f = 1.1m$ in front and $l_r = 1.58m$ behind its

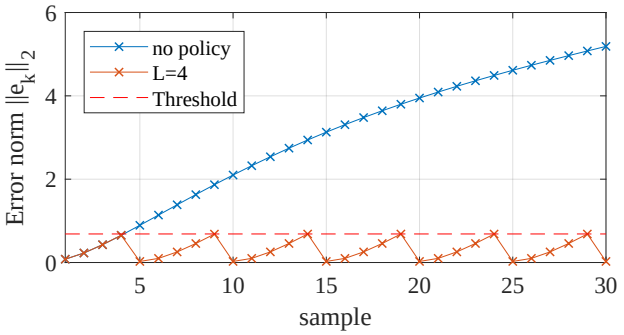

Fig. 3. Estimation error $\Delta \mathbf{e}_k$ due to false-data injection attacks for the system without authentication and with a periodic cumulative authentication policy $f_c = 4$, $L_c = 5$; the latter keeps the error below the 'safe' threshold.

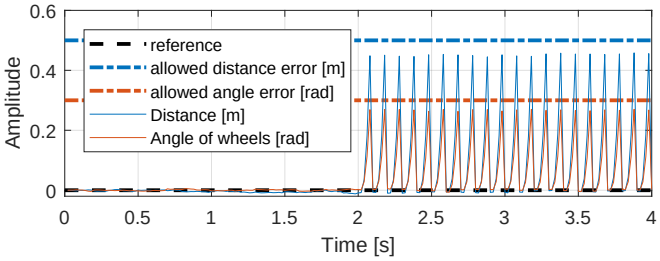

Fig. 4. Simulation of the steering control system in the presence of a stealthy attack on system sensors, with the attack starting at 2 s.

center of gravity. Yaw moment of inertia is $I_z = 2873kgm^2$, and front and rear cornering coefficients are both $C_{af} = C_{ar} = 80k$. At the time of the attack, vehicle is assumed to be moving at a constant speed of $v_x = 35m/s$. We also assume that the false positive probability is 5%, while the alarm-probability increase acceptable to the attacker is $\varepsilon = 0.01\%$. The model of the vehicle is provided as in (4), where

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -2\frac{C_{af}C_{ar}}{mv_x} & 2\frac{C_{af}C_{ar}}{m} & 2\frac{-C_{af}l_f+C_{ar}l_r}{mv_x} \\ 0 & 0 & 0 & 1 \\ 0 & -2\frac{C_{af}l_f-C_{ar}l_r}{I_z v_x} & 2\frac{C_{af}l_f-C_{ar}l_r}{I_z} & -2\frac{C_{af}l_f^2+C_{ar}l_r^2}{I_z v_x} \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 0 & 2C_{af}/m & 0 & 2\frac{C_{af}l_f}{I_z} \end{bmatrix}^T; \quad \mathbf{C} = \mathbf{I}_4$$

To define $\mathcal{R}_{risk}$ set, we assume that the attack is successful when $\|\mathbf{e}_k\|_2 \geq e_{max}$. Specifically, to impose constraints for the error in lateral position, lateral speed, steering angle and angular velocity of axle, we set $e_{max} = 0.6856$. Using the conditions from [10], [5], [2], it follows that the system is perfectly attackable (i.e., $\lim_{k\to\infty} \mathcal{R}_k$ is unbounded) when no data integrity enforcement policy is used, and thus a stealthy attacker will be able to reach $\mathcal{R}_{risk}$.

To protect from the attack, we choose a periodic cumulative integrity enforcement policy – i.e., MACs are sent at equidistant intervals L_c. We evaluated different values of L_c up to $L_c = 5$ for which the attacker almost reaches $\mathcal{R}_{risk}$. Plot of the maximal error norm evolution for the case without authentication versus the case with periodic cumulative authentication with $f_c = 4$ is shown in Fig. 3. In addition, Fig. 4 shows a $4\ s$ simulation of the estimation error if the system is compromised at $2s$, for the derived periodic cumulative authentication policy under the computed worst-case attack.

V. Conclusion

We have considered the problem of securing networked control systems against stealthy false-date injection attacks on sensor measurements. We have proposed the use of both cumulative and intermittent integrity enforcements to limit attack influence without imposing high communication (i.e., bandwidth) costs. We have shown that arbitrarily tight estimation requirements can be ensured even in the presence of attacks, with the use of appropriate integrity enforcement policies. We have introduced a formal model of stealthy attacks, and provided an efficient method to compute impact of stealthy attacks on estimation errors for systems that employ SPRT detectors. We have used this computation to guide the design of periodic cumulative authentication policies, which ensure that the system would never reach an unsafe state even under attack. Finally, we have illustrated the effectiveness of our approach on a secure vehicle steering control study.

References

[1] H. Fawzi, P. Tabuada, and S. Diggavi. Secure estimation and control for cyber-physical systems under adversarial attacks. *IEEE Transactions on Automatic Control*, 59(6):1454–1467, 2014.

[2] I. Jovanov and M. Pajic. Relaxing integrity requirements for resilient control systems. *CoRR*, abs/1707.02950, 2017.

[3] I. Jovanov and M. Pajic. Sporadic data integrity for secure state estimation. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 163–169, Dec 2017.

[4] C. Kwon and I. Hwang. Reachability analysis for safety assurance of cyber-physical systems against cyber attacks. *IEEE Transactions on Automatic Control*, 63(7):2272 – 2279, July 2018.

[5] C. Kwon, W. Liu, and I. Hwang. Analysis and design of stealthy cyber attacks on unmanned aerial systems. *Journal of Aerospace Information Systems*, 1(8), 2014.

[6] C. Kwon, S. Yantek, and I. Hwang. Real-time safety assessment of unmanned aircraft systems against stealthy cyber attacks. *Journal of Aerospace Information Systems*, pages 1–19, 2015.

[7] V. Lesi, I. Jovanov, and M. Pajic. Network scheduling for secure cyber-physical systems. In *2017 IEEE Real-Time Systems Symposium (RTSS)*, pages 45–55, Dec 2017.

[8] V. Lesi, I. Jovanov, and M. Pajic. Security-aware scheduling of embedded control tasks. *ACM Trans. Embed. Comput. Syst.*, 16(5s):188:1–188:21, Sept. 2017.

[9] 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*, 4(1):106–117, March 2017.

[10] Y. Mo, E. Garone, A. Casavola, and B. Sinopoli. False data injection attacks against state estimation in wireless sensor networks. In *49th IEEE Conf. on Decision and Control (CDC)*, pages 5967–5972, 2010.

[11] Y. Mo and B. Sinopoli. False data injection attacks in control systems. In *First Workshop on Secure Control Systems, CPS Week*, 2010.

[12] Y. Mo and B. Sinopoli. On the performance degradation of cyber-physical systems under stealthy integrity attacks. *IEEE Transactions on Automatic Control*, 61(9):2618–2624, 2016.

[13] Y. Mo, S. Weerakkody, and B. Sinopoli. Physical authentication of control systems: designing watermarked control inputs to detect counterfeit sensor outputs. *Control Systems*, 35(1):93–109, 2015.

[14] D. K. Nilsson, U. E. Larson, and E. Jonsson. Efficient in-vehicle delayed data authentication based on compound message authentication codes. In *Vehicular Technology Conference, 2008. VTC 2008-Fall. IEEE 68th*, pages 1–5. IEEE, 2008.

[15] M. Pajic, I. Lee, and G. J. Pappas. Attack-resilient state estimation for noisy dynamical systems. *IEEE Transactions on Control of Network Systems*, 4(1):82–92, March 2017.

[16] M. Pajic, J. Weimer, N. Bezzo, O. Sokolsky, G. J. Pappas, and I. Lee. Design and implementation of attack-resilient cyberphysical systems: With a focus on attack-resilient state estimators. *IEEE Control Systems*, 37(2):66–81, April 2017.

[17] R. Rajamani. *Vehicle dynamics and control*. Springer Science & Business Media, 2011.

[18] D. Umsonst, H. Sandberg, and A. A. Cárdenas. Security analysis of control system anomaly detectors. In *American Control Conference (ACC), 2017*, pages 5500–5506. IEEE, 2017.