2021 American Control Conference (ACC)
New Orleans, USA, May 25-28, 2021

Detection of Hidden Attacks on Cyber-Physical Systems from Serial Magnitude and Sign Randomness Inconsistencies

Paul J Bonczek and Nicola Bezzo

Abstract— **Stealthy false data injection attacks on cyber-physical systems (CPSs) introduce erroneous measurement information to on-board sensors with the purpose to degrade system performance. An intelligent attacker is able to leverage knowledge of the system model and noise characteristics to alter sensor measurements while remaining undetected. To achieve this objective, the stealthy attack sequence is designed such that the detector performs similarly in the attacked and attack-free cases. Consequently, an attacker that wants to remain hidden will leave behind traces of inconsistent behavior, contradicting the system model. To deal with this problem, we propose a runtime monitor to find these inconsistencies in sensor measurements by monitoring for *serial inconsistencies* of the detection test measure. Specifically, we employ the chi-square fault detection procedure to monitor the magnitude and signed sequence of its chi-square test measure. We validate our approach with simulations on an unmanned ground vehicle (UGV) under stealthy attacks and compare the detection performance with various state-of-the-art anomaly detectors.**

I. Introduction

Modern cyber-physical systems (CPSs) have been the targets of malicious cyber-attacks due to their growing unsupervised, autonomous capabilities and many entry points to implement an attack. Their expanded complexities supported by an increased number of sensors and computers allow for autonomous capabilities in navigation, warehouse logistics, surveillance, warfare, and industrial operations. With a growing number of vulnerable access points for attackers on increasingly impactful systems in our society, it is crucial to provide tighter security measures to ensure proper performance and safety.

An intelligent attacker is able to implement a malicious attack sequence to manipulate the system of interest, all while remaining undetected. The execution of such a stealthy attack allows the intelligent attacker to degrade system performance and potentially cause damage to the unknowingly compromised system. Previously demonstrated attacks of this nature include cases like: the GPS spoofing of a vessel [1], different sensor and communication attacks on vehicle technologies [2], and the infamous Stuxnet attack [3].

In order to repel these stealthy attacks, detection algorithms are designed to find compromised system's components to maintain safe operation. Intelligent attackers, in turn, resort to new methods to hide and deceive on-board control systems and their anomaly detection counterparts. While such attack vectors are less effective since the attacker needs to maintain a low profile, performance degradation can be still accomplished if the attack is able to remain undetected.

Paul J Bonczek and Nicola Bezzo are with the Charles L. Brown Department of Electrical and Computer Engineering, and Link Lab, University of Virginia, Charlottesville, VA 22904, USA. Email: {pjb4xn, nb6be}@virginia.edu

We note, however, that in general an attacker needs to create inconsistent behavior with respect to the known model in order to hijack a system. In this work, we consider the chi-square detection scheme [4] that generates a scalar quadratic *test measure* for attack detection. This test measure is extracted from the sum of squares of the residual vector — defined as the vector of differences between sensor measurements and the state prediction. To detect inconsistencies, we monitor the serial behavior of the test measure difference; specifically, we observe the characteristics of consecutive test measures throughout a sequence of measurement data and compare them to an expectation extracted from prior knowledge about the system model. Our proposed Serial Detector is then designed to generate an alarm rate at runtime for detection purposes to discover inconsistent magnitude and sign behavior due to deceptive sensor attacks.

The main objective of this work is to find intelligent sensor attack sequences that deliberately attempt to remain hidden from conventional detection techniques in noisy dynamical systems. The contribution of this paper is twofold: 1) we propose the Serial Detector to monitor inconsistent magnitude and sign behavior of the test measure difference within a system employing a chi-square fault detection procedure, and 2) we characterize a worst-case scenario that an attacker can exploit to remain undetected from our proposed detector.

A. Related Work

The field of CPS security has garnered much interest in the robotics, controls, and computer science communities to protect critical systems. In recent literature, several procedures that analyze components of the residual in the control system feedback, namely the χ^2 test measure [4], for attack detection have also been exploited. For example, the model-based Cumulative Sum (CUSUM) procedure proposed in [5] leverages the known noise characteristics of the system model and sequentially sums the test measure to detect changes within its distribution. In [6], authors included a coding matrix to the sensor outputs, unknown to attackers, to detect stealthy attacks through an iterative optimization algorithm for solving a transformation matrix. Other similar detection procedures, such as in [7], leverage watermarking of the control inputs to discover stealthy attacks.

Our recent works on attack detectors that monitor for non-random residual behavior have enabled the ability to find previously undetectable attacks when compared to conventional detection procedures. In [8], a windowed detector leveraging the Wilcoxon-Signed Rank [9] and Serial Independence Runs [10] tests was proposed to find non-random patterns over a sequence of sensor data. Similarly, in [11] we characterized the Cumulative Sign (CUSIGN) detector with the purpose

978-1-6654-4197-1/$31.00 ©2021 AACC

of finding non-random signed residual behavior by checking for changes in probability of the signed values. In this paper, we expand on these previous works by developing a runtime monitor for both non-random magnitude and sign behaviors, further strengthening detection capabilities.

The remainder of this work is organized as follows. In Section II we begin with system and estimation models along with the problem formulation, followed by the description of our Serial Detection framework in Section III. We provide an attack analysis in Section IV to expose a worst-case scenario for Serial Detection. Finally, in Section V we present numerical simulations to demonstrate the performance of our proposed detector and compare with three state-of-the-art algorithms, before discussing our conclusions in Section VI.

II. Preliminaries & Problem Formulation

In this work we consider discrete-time linear time-invariant (LTI) systems in the following form:

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{\nu}_k, \\ \boldsymbol{y}_k &= \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{\eta}_k, \end{aligned} \tag{1}$$

where the state vector $\boldsymbol{x}_k \in \mathbb{R}^n$, $k \in \mathbb{N}$ evolves due to the discrete-time state transition and input matrices $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n\times m}$, control input $\boldsymbol{u}_k \in \mathbb{R}^m$, and additive i.i.d. zero-mean Gaussian process uncertainty $\boldsymbol{\nu}_k = \mathcal{N}(0, \boldsymbol{Q}) \in \mathbb{R}^n$ described by the covariance matrix $\boldsymbol{Q} \in \mathbb{R}^{n\times n}, \boldsymbol{Q} \geq 0$. The output vector $\boldsymbol{y}_k \in \mathbb{R}^s$ represents the measured system states with additive i.i.d. zero-mean Gaussian measurement uncertainty $\boldsymbol{\eta}_k = \mathcal{N}(0, \boldsymbol{R}) \in \mathbb{R}^s$ with covariance matrix $\boldsymbol{R} \in \mathbb{R}^{s\times s}, \boldsymbol{R} \geq 0$ that provides measurements to s sensors.

We consider sensor measurements $\boldsymbol{y}_k$ that can be altered due to an additive attack vector $\boldsymbol{\xi}_k \in \mathbb{R}^s$, which results in an attacked output measurement vector described by

$$\tilde{\boldsymbol{y}}_k = \boldsymbol{y}_k + \boldsymbol{\xi}_k = \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{\eta}_k + \boldsymbol{\xi}_k \in \mathbb{R}^s. \tag{2}$$

During operations, a steady state Kalman Filter with gain matrix $\boldsymbol{L} \in \mathbb{R}^{n\times s}$ is implemented to provide a state estimate $\hat{\boldsymbol{x}}_k \in \mathbb{R}^n$ in the form:

$$\begin{aligned} \hat{\boldsymbol{x}}_{k+1} &= \boldsymbol{A}\hat{\boldsymbol{x}}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{L}(\tilde{\boldsymbol{y}}_k - \boldsymbol{C}\hat{\boldsymbol{x}}_k), \\ \boldsymbol{L} &= \boldsymbol{P}\boldsymbol{C}^\mathsf{T}(\boldsymbol{C}\boldsymbol{P}\boldsymbol{C}^\mathsf{T} + \boldsymbol{R})^{-1}. \end{aligned} \tag{3}$$

The gain matrix $\boldsymbol{L}$ results in a minimal steady state estimation error covariance matrix $\boldsymbol{P} = \mathbb{E}[\boldsymbol{e}_k\boldsymbol{e}_k^\mathsf{T}]$, where $\boldsymbol{e}_k = \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k$ is the estimation error. The measurement residual vector is defined as

$$\boldsymbol{r}_k = \tilde{\boldsymbol{y}}_k - \boldsymbol{C}\hat{\boldsymbol{x}}_k = \boldsymbol{C}\boldsymbol{e}_k + \boldsymbol{\eta}_k \in \mathbb{R}^s, \tag{4}$$

with an expected residual covariance matrix, in attack-free conditions (i.e. $\boldsymbol{\xi}_k = \boldsymbol{0}$), described by:

$$\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{r}_k\boldsymbol{r}_k^\mathsf{T}] = \boldsymbol{C}\boldsymbol{P}\boldsymbol{C}^\mathsf{T} + \boldsymbol{R} \in \mathbb{R}^{s\times s}. \tag{5}$$

For the measurement residual, we test two different hypotheses: The null hypothesis $\mathcal{H}_0$ for nominal scenario (attack-free) and the alternative hypothesis $\mathcal{H}_a$ where attacks are present. Formally, the hypotheses are written as

$$\mathcal{H}_0 : \begin{cases} \mathbb{E}[\boldsymbol{r}_k] = 0, \\ \mathbb{E}[\boldsymbol{r}_k\boldsymbol{r}_k^\mathsf{T}] = \boldsymbol{\Sigma}, \end{cases} \quad \mathcal{H}_a : \begin{cases} \mathbb{E}[\boldsymbol{r}_k] \neq 0, \text{and/or} \\ \mathbb{E}[\boldsymbol{r}_k\boldsymbol{r}_k^\mathsf{T}] \neq \boldsymbol{\Sigma}. \end{cases} \tag{6}$$

In this work, we consider a single detector to monitor the system for sensor attacks by way of the chi-square detector, which produces a scalar quadratic *test measure* z_k by

$$z_k = \boldsymbol{r}_k^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{r}_k \in \mathbb{R}_{\geq 0}. \tag{7}$$

In the absence of attacks, the measurement residual is an s-dimensional vector of normally distributed random variables $\boldsymbol{r}_k \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, satisfying the null hypothesis $\mathcal{H}_0$ in (6). The test measure z_k is then expected to be a random variable that follows a chi-square distribution with s degrees of freedom, i.e. $z_k \sim \chi^2(s)$, that follows:

$$\mathbb{E}[z_k] = s, \quad \text{Var}[z_k] = 2s. \tag{8}$$

A. Undetected Attacks

A successful attacker is capable of modeling an attack sequence to achieve a desirable effect while remaining undetectable to any on-board fault detection mechanisms. In order to accomplish such stealthy behavior, it is necessary to attain information about critical aspects of the system, such as: acquiring knowledge to the modeled dynamics, sensor measurements, state estimator, and detection procedure(s). To intentionally avoid detection, an intelligent attacker will carefully construct an attack sequence to evade raising any flags. Below we describe a sequence an attacker may take with respect to the Bad-Data detector [4] leveraging the chi-square test measure procedure. However, this may be extended to satisfy any detection procedure using a similar concept to avoid detection.

Zero-alarm attacks are sequences designed by an attacker that maintains the test measure from exceeding the defined threshold value ($z_k \leq \tau_z$). This class of attack does not trigger an alarm throughout the attack sequence, as the test measure never passes the threshold. In order to satisfy such requirements, an attacker can construct the attack vector by

$$\boldsymbol{\xi}_k = -\boldsymbol{C}\boldsymbol{e}_k - \boldsymbol{\eta}_k + \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\delta}_k, \tag{9}$$

where $\boldsymbol{\delta}_k \in \mathbb{R}^s$ is a vector that satisfies $\boldsymbol{\delta}_k^\mathsf{T}\boldsymbol{\delta}_k \leq \tau_z$. With this attack vector designed at a time k, the test measure z_k results in

$$\begin{aligned} z_k &= (\tilde{\boldsymbol{y}}_k - \boldsymbol{C}\hat{\boldsymbol{x}}_k)^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{y}}_k - \boldsymbol{C}\hat{\boldsymbol{x}}_k) \\ &= (\boldsymbol{C}\boldsymbol{e}_k + \boldsymbol{\eta}_k + \boldsymbol{\xi}_k)^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{C}\boldsymbol{e}_k + \boldsymbol{\eta}_k + \boldsymbol{\xi}_k) \leq \tau_z, \end{aligned} \tag{10}$$

that remains within the threshold value to not trigger an alarm. While generating an attack sequence that does not trigger alarms may seem like a favorable attack design, it is necessary to recall that alarms are triggered in a system operating in normal conditions without attacks. If alarms are no longer being triggered as designed for in an attack-free case, then these conditions may raise suspicions of a possible attack. To avoid these alarm rate discrepancies, an attacker would want to design an attack sequence that is undetectable to emulate normal (attack-free) conditions. This class of attack brings us to develop a sequence that exploits the system uncertainties to execute such a malicious attack.

Hidden attacks can be defined as designed attack sequences such that alarms are triggered at the same rate as the desired false alarm rate during nominal, attack-free operation. As shown in Fig. 1, during a hidden attack, a smart attacker can design a sequence where the test measure

z_k exceeds the threshold τ_z at the same rate as nominal conditions. To tune for a desired alarm rate α (in the attack-free scenario) for Bad-Data detection while leveraging the chi-square procedure, the specific threshold τ_z is found by

$$\tau_z = 2\gamma^{-1}\Big(1-\alpha, \frac{s}{2}\Big), \tag{11}$$

to achieve a desired alarm rate, where $\gamma^{-1}(\cdot,\cdot)$ is the *inverse regularized lower incomplete gamma function* [12]. The vector $\boldsymbol{\delta}_k$ from (9) is designed such that

$$\mathbb{P}(z_k > \tau_z) = \mathbb{P}(\boldsymbol{\delta}_k^\mathsf{T}\boldsymbol{\delta}_k > \tau_z) \approx \alpha, \tag{12}$$

to remain hidden from detection.

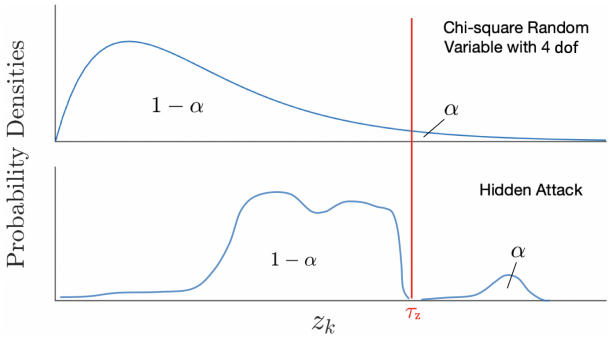

Fig. 1. (top) The chi-square distribution of z_k with $s = 4$ degrees of freedom (dof). A correctly chosen threshold value τ_z results in the test measure z_k exceeding the threshold at a desired rate of α in the attack-free case. Similarly, during a hidden attack (bottom), an attacker can design an attack sequence such that the alarm rate matches the desired alarm rate α, all while altering the distribution of z_k and remaining hidden from detection.

B. Problem Formulation

An attacker with the specific objective to hijack a system or to degrade system performance, will leave traces of inconsistent behavior. In this work, we focus on deceptive sensor attacks that purposely hide within the noise characteristics of the system model and evade detection of conventional fault detection procedures in order to remain undetected.

Definition 1: Sensor measurements are behaving consistently if:

- The test measure follows a chi-square distribution $z_k \sim \chi^2(s)$ that is determined by the s number of sensors.
- The signed test measure difference switches sign values at a proper (i.e., expected) rate.

Since we are considering sensor spoofing, an attack vector $\boldsymbol{\xi}_k$ containing malicious data can disrupt consistency, thereby causing the test measure to display non-random behavior. Formally, the problem that we are interested in solving is:

Problem 1: (Runtime Detection of Measurement Inconsistencies). Given the quadratic test measure z_k, computed from the residual $\boldsymbol{r}_k$ as defined in (4) and the inverse of the residual covariance matrix $\boldsymbol{\Sigma}$ in (5), find a policy to determine at runtime whether sensor measurements are consistent, i.e., if any condition in Definition 1 does not hold.

III. Serial Consistency of the Test Measure

The overall cyber-physical control system architecture including our detection procedure for serial consistencies is summarized in Fig. 2. The monitor is placed in the system feedback to observe the relationship between the sensor measurements and state prediction while leveraging the chi-square test measure. We focus our attention on stealthy sensor attack sequences where a malicious attacker may inject an attack signal to measurements at any point between the sensors and state estimator.

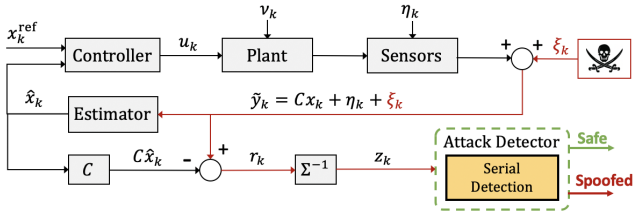

Fig. 2. The architecture of a CPS while experiencing sensor attacks with the Serial Detector placed in the system feedback.

A. Magnitude-based Detection

The design of the Serial Detector is to find inconsistent behavior of chi-square test measures within its expected distribution due to stealthy attacks to on-board sensor measurements. An attacker deliberately attempting to fool test measure-based detection algorithms may leave traces of inconsistencies within the serial sequence. We propose the Serial Detector that analyzes consecutive chi-square test measures at time instances k and $k-1$, called the *test measure difference*, that is described as:

$$\begin{aligned} d_k &= z_k - z_{k-1}, \\ &= \boldsymbol{r}_k^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{r}_k - \boldsymbol{r}_{k-1}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{r}_{k-1} \in \mathbb{R}. \end{aligned} \tag{13}$$

Proposition 1: A system that is free from sensor attacks, where we assume consecutive test measures are independent random variables that follow chi-square distributions $z_k, z_{k-1} \sim \chi^2(s)$ with s degrees of freedom, has the following expectations of the test measure difference d_k:

$$\begin{aligned} \mathbb{E}[d_k] &= \mathbb{E}[z_k] - \mathbb{E}[z_{k-1}] = 0, \\ \text{Var}[d_k] &= \text{Var}[z_k] + \text{Var}[z_{k-1}] = 4s. \end{aligned} \tag{14}$$

Given an attack-free system that follows the expectation (14) in Proposition 1, the test measure difference d_k follows

$$d_k \sim \mathcal{VG}\Big(\mathbb{E}[d_k], \sqrt{\text{Var}[d_k]}, 0, \frac{2}{s}\Big) \in \mathbb{R}, \tag{15}$$

where $\mathcal{VG}(\cdot,\cdot,\cdot,\cdot)$ denotes the *variance-gamma distribution* [13], which is a mixed distribution of the normal distribution and gamma distribution. As the chi-square distribution is a special case of the gamma distribution, the difference of two gamma random variables (i.e. chi-square random variables) results in the variance-gamma distribution [14]. The parameters within the variance-gamma distribution that describe the difference of two chi-square random variables, generalized in [15], are the location $c = \mathbb{E}[d_k]$, spread $\bar{\sigma} = \sqrt{\text{Var}[d_k]}$, asymmetry $\vartheta = 0$, and shape $\lambda = \frac{2}{s}$. The probability density function (PDF) of the variance-gamma distribution follows

$$\begin{aligned} f(x;c,\bar{\sigma},\vartheta,\lambda) = &\frac{2e^{(\vartheta(x-c)/\bar{\sigma}^2)}|x-c|^{\frac{1}{\lambda}-\frac{1}{2}}}{\bar{\sigma}\sqrt{2\pi}\lambda^{\frac{1}{\lambda}}\Gamma\big(\frac{1}{\lambda}\big)}\Bigg(\frac{1}{\sqrt{2\bar{\sigma}^2/\lambda+\vartheta^2}}\Bigg)^{\frac{1}{\lambda}-\frac{1}{2}} \\ &\times K_{\frac{1}{\lambda}-\frac{1}{2}}\Bigg(\frac{|x-c|\sqrt{2\bar{\sigma}^2/\lambda+\vartheta^2}}{\bar{\sigma}^2}\Bigg), \end{aligned} \tag{16}$$

where K_λ is the *modified Bessel function of the third kind* of order λ and $\Gamma(\cdot)$ is the *gamma function* [12]. During nominal conditions, the test measure difference d_k is a symmetric zero-mean distribution (i.e., parameters $c = \vartheta = 0$).

False Alarms: Similar to other detection algorithms in literature [4], [5], we leverage an alarm rate to diagnose the health of the system from sensor attacks. The magnitude-based detection scheme compares the test measure difference d_k to a threshold τ by:

$$\begin{cases} |d_k| > \tau_d & \longrightarrow \quad \textit{alarm: } \zeta_k^M = 1, \\ |d_k| \leq \tau_d & \longrightarrow \quad \textit{no alarm: } \zeta_k^M = 0, \end{cases} \tag{17}$$

where the chosen threshold τ_d is dependent on the expected test measure difference distribution described in (15). In Fig. 3 we show how the distribution of the test measure difference $d_k = z_k - z_{k-1}$ (the difference of two chi-square random variables) is affected by the number of sensors s.

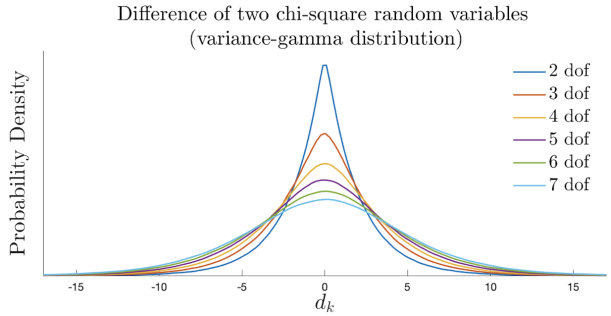

Fig. 3. The resulting distribution of the test measure difference $d_k = z_k - z_{k-1}$ follows a variance-gamma distribution in the attack-free scenario. Shown are the effects on the distribution of d_k for dof $= \{2, 3, 4, 5, 6, 7\}$.

Under nominal circumstances, i.e. in the absence of attacks, an alarm is triggered at a desired rate $\psi_{des}^M \in (0, 1)$ given the chosen threshold value. The following lemma provides a method to choose a threshold to satisfy a user-defined desired alarm rate.

Lemma 1: Assuming that the system is attack-free (i.e. $\boldsymbol{\xi}_k = \mathbf{0}$) and considering the procedure in (17) to trigger an alarm, a specific threshold value τ_d is chosen by

$$\tau_d = \{\tau_d \in \mathbb{R}_{>0} : \mathbb{P}(\zeta_k^M = 1) = \psi_{des}^M\}, \tag{18}$$

such that the result is a desired alarm rate ψ_{des}^M.

Proof: Let $F_{d_k}(x; c, \bar{\sigma}, \theta, \lambda)$ denote the cumulative distribution function (CDF) of the random variable d_k from the PDF in (16). We compute the inverse CDF for a given desired false alarm rate ψ_{des}^M to find the threshold value

$$\tau_d = F_{d_k}^{-1}\big(1 - \frac{\psi_{des}^M}{2}; c, \bar{\sigma}, \theta, \lambda\big) \in \mathbb{R}_{>0}, \tag{19}$$

such that $\mathbb{P}(\zeta_k^M = 1) = \mathbb{P}(|d_k| > \tau_d) = \psi_{des}^M$ to achieve a desired false alarm rate, thus concluding the proof. ■

Alarm Rate Estimation: We employ a runtime method of estimating the alarm rate such that we are able to eliminate the need to store a sequence of values. In this work, a Memoryless Runtime Estimator (MRE) [11] is leveraged to eliminate the need to use a windowed method to compute an alarm rate estimation $\hat{\psi}_k^M \in [0, 1]$. The MRE algorithm is updated by following

$$\hat{\psi}_k^M = \hat{\psi}_{k-1}^M + \frac{\zeta_k^M - \hat{\psi}_{k-1}^M}{\ell}, \tag{20}$$

where ℓ is a user-defined "pseudo-window" length. The resulting distribution while leveraging MRE can be approximated to a normal distribution for pseudo-window lengths $\ell \geq 10$ [11] consisting of a variance that follows that of a exponential moving average (EMA) [16].

Lemma 2: Given the test measure difference d_k defined in (13) for a system that is assumed to be attack-free and tuned for a desired false alarm rate ψ_{des}^M, the estimate alarm rate follows a Normal distribution described by

$$\hat{\psi}_k^M \sim \mathcal{N}\left(\psi_{des}^M, \frac{\psi_{des}^M(1 - \psi_{des}^M)}{2\ell - 1}\right). \tag{21}$$

Proof: We first characterize the magnitude-based detector tuned for a desired false alarm rate ψ_{des}^M as a Binomial distribution $\mathcal{B}(\cdot, \cdot)$ where ψ_{des}^M is a probability for a "success" during a specified number of "trials" (Refer to [12] for further explanations). By way of the binomial approximation for larger pseudo-window size $\ell \geq 10$, a normal distribution can be used to approximate the alarm rate while leveraging MRE (20) for estimation that results in

$$\mathbb{E}[\psi^M] = \psi_{des}^M, \quad \text{Var}[\psi^M] = \frac{\psi_{des}^M(1 - \psi_{des}^M)}{2\ell - 1}. \tag{22}$$

From (22) we obtain the distribution in (21) to characterize the estimated alarm rate for magnitude-based detection. ■

With the known expected false alarm rate distribution described in (21), we want to find bounds on the estimated alarm rate $\hat{\psi}_k^M$, $\forall k$ to determine if an attack has occurred. The following corollary provides detection bounds with a specific level of confidence $1 - \beta$, where $\beta \in [0, 1]$ is a user defined level of significance[1].

Corollary 1: Assuming a system with s sensors that employs the chi-square detection scheme that is monitoring the test measure difference (13) with a level of significance β while leveraging MRE (20), detection of sensor attacks occurs when $\Omega_- \leq \hat{\psi}_k^M \leq \Omega_+$ is not satisfied where

$$\Omega_\pm = \mathbb{E}[\psi^M] \pm Z\sqrt{\frac{\mathbb{E}[\psi^M](1 - \mathbb{E}[\psi^M])}{2\ell - 1}}. \tag{23}$$

Proof: We construct confidence intervals for a normally distributed variable of a specific confidence level, determined by z-score $Z = \left|\Phi^{-1}\left(\frac{\beta}{2}\right)\right|$, that provide detection bounds by

$$\begin{aligned} \mathbb{E}[\psi^M] - \left|\Phi^{-1}\left(\frac{\beta}{2}\right)\right|\sqrt{\frac{\mathbb{E}[\psi^M](1 - \mathbb{E}[\psi^M])}{2\ell - 1}} \leq \hat{\psi}_k^M \\ \leq \mathbb{E}[\psi^M] + \left|\Phi^{-1}\left(\frac{\beta}{2}\right)\right|\sqrt{\frac{\mathbb{E}[\psi^M](1 - \mathbb{E}[\psi^M])}{2\ell - 1}} \end{aligned} \tag{24}$$

which satisfy (23), concluding the proof. ■

Detection of sensor attacks occur when an estimated alarm rate $\hat{\psi}_k^M$ travels beyond the thresholds from $\Omega_\pm = [\Omega_-, \Omega_+]$.

[1]Reducing the value of β causes the detection bounds to move farther from the expected alarm rate, thus reducing the frequency of falsely "detecting" under nominal (i.e., no attack) conditions while consequently giving an attacker more freedom to design an attack without being detected, while the opposite is true when increasing β.

B. Signed Randomness

To further strengthen detection of inconsistencies within the test measure, we monitor the "runs" behavior of the test measure difference sequence. While a smart attacker may be able to fool the magnitude-based monitor as discussed in Section III-A, an attacker may leave traces of non-random behavior on the signed test measure difference. The test we use to monitor for signed randomness is influenced by the Serial Independence Runs (SIR) Test [10]. An example of the SIR test is shown in Fig. 4, where it monitors a sequence of data by first computing the difference between the current and previous data values and taking the sign of the difference to create a two-valued data sequence (i.e., positive and negative values). Then, the number of observed runs N_r, defined as consecutive values of the same sign, are counted over the sequence length. In Fig. 4 we see that over the sequence length $W = 14$ there are $N_r = 12$ runs, which are highlighted by the red and blue lines.

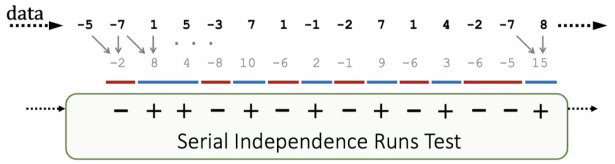

Fig. 4. A sequence of data, from left to right, converted to sequence of signed values while leveraging the Serial Independence Runs Test.

A drawback of the SIR test is the requirement to store the W length sequence of test measure differences d_k and then count the number of observed runs N_r over this sequence. Alternatively, we would like to use a window-less method to eliminate the need for storing an entire sequence to determine whether the signed test measure difference is behaving randomly. To this end, we propose to observe sign switches at runtime by triggering an alarm at a time k when the the present test measure difference is of the opposite sign from the previous test measure difference at time $k-1$.

We first compute the sign of the test measure difference by the following

$$\operatorname{sgn}(d_k) := \begin{cases} -1, & \text{if } d_k < 0, \\ 0, & \text{if } d_k = 0, \\ 1, & \text{if } d_k > 0, \end{cases} \tag{25}$$

and given that the distribution of the test measure difference d_k is symmetric (assuming nominal conditions) over the expected value $\mathbb{E}[d_k] = 0$, the probability of observing the signed values of the test measure difference are

$$\begin{aligned} \mathbb{P}\big(\operatorname{sgn}(d_k) = -1\big) &= 0.5, \\ \mathbb{P}\big(\operatorname{sgn}(d_k) = 0\big) &= 0, \\ \mathbb{P}\big(\operatorname{sgn}(d_k) = 1\big) &= 0.5. \end{aligned} \tag{26}$$

As sensor measurements are received at every kth time instance, the next test measure difference d_k is computed from (4), (7), and (13). A switch of the test measure difference sign signifies the end of a run and an alarm $\zeta_k^S \in \{0,1\}$ is triggered such that $\zeta_k^S = 1$ at a time instance k, otherwise $\zeta_k^S = 0$. The procedure to trigger a test measure difference alarm follows:

$$\zeta_k^S := \begin{cases} 1, & \text{if } \operatorname{sgn}(d_k) = -\operatorname{sgn}(d_{k-1}), \\ 0, & \text{otherwise.} \end{cases} \tag{27}$$

The alarm $\zeta_k^S \in \{0,1\}$ in (27) is then sent into the MRE to provide an updated runtime estimate of the test measure difference alarm rate $\hat{\psi}_k^S$ at time instance k.

Lemma 3: Given a system that is not experiencing attacks, the test measure difference alarm rate while leveraging MRE (20) for estimation is described as a Normally distributed random variable by the following

$$\hat{\psi}_k^S \sim \mathcal{N}\big(\mathbb{E}[\psi^S], \operatorname{Var}[\psi^S]\big). \tag{28}$$

Proof: We want to convert the distribution of expected runs $\mathbb{E}[N_R]$ of test measure differences over a window-based sequence of length W described in [10] by

$$N_R \sim \mathcal{N}\left(\frac{2W-1}{3}, \frac{16W-29}{90}\right), \tag{29}$$

to a runtime rate of expected test measure difference sign switching $\mathbb{E}[\psi^S]$. By first obtaining the asymptotic distribution and then transforming the expected observed runs to an expected rate of observed alarms $\mathbb{E}[\psi^S] = \frac{\mathbb{E}[N_R]}{W}$, we arrive to the expected sign switching alarm rate distribution

$$\mathbb{E}[\psi^S] = \frac{2}{3}, \;\; \operatorname{Var}[\psi^S] = \frac{16}{90(2\ell-1)}, \tag{30}$$

while leveraging MRE for window-less estimation. ■

The following corollary provides a proof for detection bounds of $\hat{\psi}_k^S$ to satisfy an expected alarm rate $\mathbb{E}[\psi^S]$.

Corollary 2: Given the test measure differences $d_k = z_k - z_{k-1}$, detection occurs by the test measure difference alarm rate when $\Psi_- \le \hat{\psi}_k^S \le \Psi_+$ is not satisfied where

$$\Psi_\pm = \pm\left|\Phi^{-1}\left(\frac{\beta}{2}\right)\right|\sqrt{\frac{16}{90(2\ell-1)}} + \frac{2}{3}. \tag{31}$$

Proof: For a desired level of significance β we find the bounds of $\hat{\psi}_k^S$ for an expected alarm rate $\mathbb{E}[\psi^S]$ are

$$-Z\sqrt{\frac{16}{90(2\ell-1)}} + \frac{2}{3} \le \hat{\psi}_k^S \le Z\sqrt{\frac{16}{90(2\ell-1)}} + \frac{2}{3}, \tag{32}$$

where z-score is $Z = \left|\Phi^{-1}\left(\frac{\beta}{2}\right)\right|$. From (32) we can finally obtain the detection bounds of $\Psi_\pm$ in (31) for alarm triggering at an expected alarm rate $\mathbb{E}[\psi^S]$. ■

We should note that while we describe a runtime method for detecting anomalous signed behavior within the serial sequence of a chi-square random variable z_k, this technique may be used on any randomly distribution variable. As this method is non-parametric, the signed behavior is independent from its underlying distribution [10], [17].

IV. Undetectable Attacks

This section analyzes the attack sequence that an attacker must make in order to remain undetected from our serial randomness-based detector. Continuing with assumptions previously made in Section II-A, we assume a worst-case scenario where a smart attacker has access to the system

model, noise characteristics, control inputs, and state estimator to fool our detection technique. In particular, we focus on the attack sequences of $\boldsymbol{\xi}_k$ that can disrupt nominal closed-loop system behavior while remaining hidden.

A. Magnitude-based Detection

We begin by considering an attack sequence that does not allow the magnitude-based alarm rate $\hat{\psi}_k^M$ to travel beyond detection bounds described in (23). If we recall the test measure difference d_k in (13), but written in terms of the sensor attack vector $\boldsymbol{\xi}_k$, we have

$$\begin{aligned} d_k &= \boldsymbol{r}_k^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{r}_k - \boldsymbol{r}_{k-1}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{r}_{k-1} \\ &= (\boldsymbol{C}\boldsymbol{e}_k + \boldsymbol{\eta}_k + \boldsymbol{\xi}_k)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{C}\boldsymbol{e}_k + \boldsymbol{\eta}_k + \boldsymbol{\xi}_k) \\ &\quad -(\boldsymbol{C}\boldsymbol{e}_{k-1} + \boldsymbol{\eta}_{k-1} + \boldsymbol{\xi}_{k-1})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{C}\boldsymbol{e}_{k-1} + \boldsymbol{\eta}_{k-1} + \boldsymbol{\xi}_{k-1}) \\ &= (\boldsymbol{C}\boldsymbol{e}_k + \boldsymbol{\eta}_k + \boldsymbol{\xi}_k)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{C}\boldsymbol{e}_k + \boldsymbol{\eta}_k + \boldsymbol{\xi}_k) - z_{k-1}. \end{aligned} \tag{33}$$

In order for an attacker to not trigger the alarm $\zeta_k^M = 1$ at time k, i.e. a zero-alarm attack, the sensor attack vector must maintain the test measure difference to satisfy $|d_k| \leq \tau_d$. For an attack vector sequence and the designed variance-gamma distribution threshold τ_d, we define a suitable vector

$$\boldsymbol{\delta}_k = \{\boldsymbol{\delta}_k \in \mathbb{R}^s : |\boldsymbol{\delta}_k^{\mathsf{T}}\boldsymbol{\delta}_k - z_{k-1}| \leq \tau_d\}, \tag{34}$$

that leads to the test measure difference d_k not triggering an alarm. Therefore, for any time k, the attack vector follows

$$\boldsymbol{\xi}_k = -\boldsymbol{C}\boldsymbol{e}_k - \boldsymbol{\eta}_k + \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\delta}_k, \tag{35}$$

where $\boldsymbol{\Sigma}^{\frac{1}{2}}$ is the symmetric square root of the residual covariance matrix (5), such that

$$|d_k| = |z_k - z_{k-1}| \leq \tau_d, \tag{36}$$

is satisfied. To remain hidden from detection, an attacker must trigger alarms at a rate which the system is expecting. For the case of hidden attacks to evade detection of our serial monitor for magnitude-based detection, a suitable vector $\boldsymbol{\delta}_k$ in (34) is constructed as

$$\mathbb{P}(|d_k| > \tau_d) = \mathbb{P}(|\boldsymbol{\delta}_k^{\mathsf{T}}\boldsymbol{\delta}_k - z_{k-1}| > \tau_d) \approx \psi_{des}^M, \tag{37}$$

to emulate the alarm rate that would be seen during nominal conditions. More specifically, an observed estimated alarm rate computed in (20) must remain within detection bounds found in (23) to remain undetected. To ensure detection does not occur for the magnitude-based alarm rate, an attacker must design the attack vector such that the alarm rate remains within detection bounds, $\hat{\psi}_k^M \in [\Omega_-, \Omega_+]$. To remain below the upper detection bound, the vector $\boldsymbol{\delta}_k$ follows

$$|\boldsymbol{\delta}_k^{\mathsf{T}}\boldsymbol{\delta}_k - z_{k-1}| \leq \tau_d \,:\, \left(\Omega_+ - \hat{\psi}_{k-1}^M - \frac{1-\hat{\psi}_{k-1}^M}{\ell}\right) < 0, \tag{38}$$

to guarantee $\hat{\psi}_k^M \leq \Omega_+$. Additionally, a requirement to remain above the lower detection bound adheres to

$$|\boldsymbol{\delta}_k^{\mathsf{T}}\boldsymbol{\delta}_k - z_{k-1}| > \tau_d \,:\, \left(\Omega_- - \hat{\psi}_{k-1}^M + \frac{\hat{\psi}_{k-1}^M}{\ell}\right) > 0. \tag{39}$$

B. Sign-based Detection

We continue with a scenario for an attacker to evade detection from the Serial Detector in which the attack design is required to satisfy signed randomness throughout the sequence. Similar to the magnitude-based detection in Section IV-A, the attack sequence must result in alarm rates that emulate attack-free conditions to remain hidden from detection. To achieve this, the sign-based alarm rate satisfies

$$\begin{aligned} &\mathbb{P}\big(\mathrm{sgn}(d_k) = -\mathrm{sgn}(d_{k-1})\big) = \\ &\mathbb{P}\big(\mathrm{sgn}(z_k - z_{k-1}) = -\mathrm{sgn}(z_{k-1} - z_{k-2})\big) \approx \mathbb{E}[\psi^S], \end{aligned} \tag{40}$$

in order to behave similarly to nominal conditions. In order to not cause a sign switching condition, i.e. signed-based alarm $\zeta_k^S = 0$, the sign of d_k must consist of the same sign as d_{k-1}. In terms of the vector $\boldsymbol{\delta}_k$ while leveraging (9), the following inequality

$$\begin{cases} \boldsymbol{\delta}_k^{\mathsf{T}}\boldsymbol{\delta}_k > z_{k-1}, & \text{if } d_{k-1} > 0, \\ \boldsymbol{\delta}_k^{\mathsf{T}}\boldsymbol{\delta}_k < z_{k-1}, & \text{if } d_{k-1} < 0, \end{cases} \tag{41}$$

must be satisfied to not cause a sign change, thus not triggering an alarm. If the signed component alarm rate ψ_k^S, $\forall k$ approaches the upper detection bound, the following equation guarantees $\hat{\psi}_k^S \leq \Psi_+$, where

$$\mathrm{sgn}(\boldsymbol{\delta}_k^{\mathsf{T}}\boldsymbol{\delta}_k - z_{k-1}) = \mathrm{sgn}(d_{k-1}) \,:\, \left(\Psi_+ - \hat{\psi}_{k-1}^S - \frac{1-\hat{\psi}_{k-1}^S}{\ell}\right) < 0, \tag{42}$$

thus maintaining the alarm rate within bounds. Similarly, the requirement to not cross below the lower bound adheres to

$$\mathrm{sgn}(\boldsymbol{\delta}_k^{\mathsf{T}}\boldsymbol{\delta}_k - z_{k-1}) = -\mathrm{sgn}(d_{k-1}) \,:\, \left(\Psi_- - \hat{\psi}_{k-1}^S + \frac{\hat{\psi}_{k-1}^S}{\ell}\right) > 0, \tag{43}$$

to remain undetectable from the Serial Detector.

V. Results

The proposed Serial Detector was validated in simulation and compared to state-of-the-art detection techniques: Bad-Data (BD) [4], Cumulative Sum (CUSUM) [5], and Cumulative Sign (CUSIGN) [11] detectors. The case study presented in this paper is an autonomous differential-drive UGV with the following linearized model [18]

$$\begin{aligned} \dot{v} &= \frac{1}{m}(F_l + F_r - B_r v), \\ \dot{\omega} &= \frac{1}{I_z}\Big(\frac{w}{2}(F_l - F_r) - B_l\omega\Big), \;\; \dot{\theta} = \omega, \end{aligned} \tag{44}$$

where v, θ, and ω denote velocity, vehicle heading angle, and angular velocity, forming the state vector $\boldsymbol{x} = [v, \theta, \omega]^{\mathsf{T}}$. F_l and F_r describe the left and right input forces from the wheels, w is the vehicle width, while B_r and B_l are resistances due to the wheels rolling and turning. Two sensors ($s = 2$) receive measurements of the states $x_1 = v$ and $x_2 = \theta$ with a sampling rate $t_s = 0.01$.

We perform two different attack sequences: *Bias Attack* where the attack sequence concentrates the test measure distribution such that the magnitude detectors (BD and CUSUM) trigger alarms at a desired rate while signed behavior monitored by CUSIGN remains consistent whereas a *Pattern Attack* creates patterned concentrations on the chi-square test measure difference d_k. In Fig. 5, the resulting

distributions for each case are shown that include: (a) the *No Attack* case where $z_k \sim \chi^2(s=2)$, (b) *Bias Attack*, and (c) *Pattern Attack*. Both the Bad-Data and CUSUM detectors are tuned for a desired alarm rate of $\alpha = 0.20$ (see [4], [5]) and the CUSIGN detector has an expected alarm rate of 0.0833 (see [11]). The magnitude component of our proposed Serial Detector is tuned for an expected alarm rate $\mathbb{E}[\psi^M] = 0.20$ and the expected alarm rate for the signed component is $\mathbb{E}[\psi^S] = \frac{2}{3}$. All detectors employ detection bounds that are 3 standard deviations from their expectation.

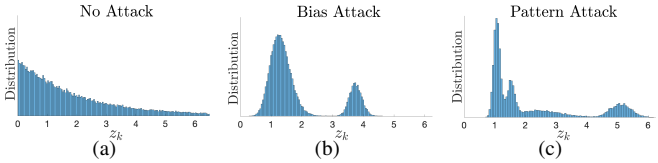

Fig. 5. The test measure z_k distributions when (a) *No attack*, (b) *Bias Attack*, and (c) *Pattern Attack* occur in the simulation case study.

Next, we include a simulation showing the detector alarm rates during *No Attack* at times $k < 20000$, *Bias Attack* at $20000 \leq k < 40000$ and *Pattern Attack* beginning at time step $k \geq 40000$. During the *Bias Attack* in Fig. 6, the attack fools the BD, CUSUM, and CUSIGN detectors, but the magnitude component of the serial monitor notices the change in the test measure sequence due to the attack. The sign component does not detect the attack, as a bias attack does not disrupt the change of signed behavior of the test measure difference d_k. The *Pattern Attack*, while preserving expected test measure difference magnitude behavior, interferes with the expected sign switching rate of the test measure difference. As expected, in the absence of sensor attacks where $k < 20000$, alarm rates for all detection procedures have distributions centered at their expectations. While these modeled attack sequences are primitively designed examples that can fool comparative detectors (e.g. BD, CUSUM, and CUSIGN detectors), the Serial Detector is able to exploit hidden behaviors to strengthen detection capabilities.

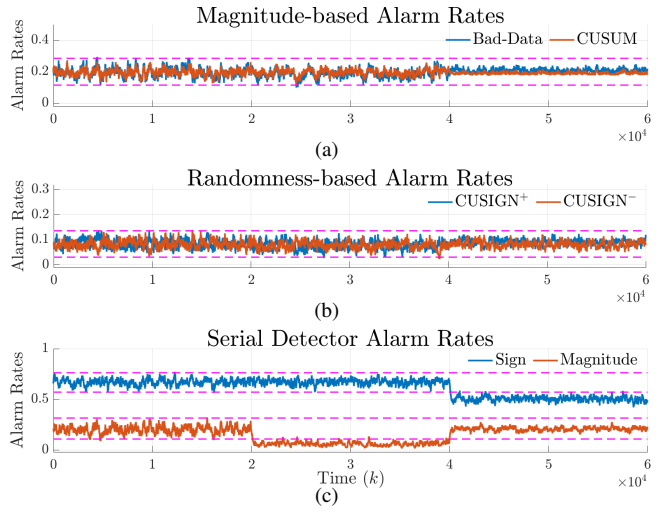

Fig. 6. Resulting alarm rates during the *No Attack*, *Pattern Attack*, and *Bias Attack*. During both the attack scenarios, the comparable detectors (BD, CUSUM, and CUSIGN) are fooled, while the magnitude and sign components of the Serial Detector discover the *Bias* and *Pattern* attacks. Dashed magenta lines represent 3σ detection bounds for each detector.

VI. Conclusions

In this paper we have proposed the Serial Detector to discover inconsistent test measure behavior due to hidden cyber-attacks while employing a chi-square detection procedure. Our detection approach monitors the magnitude and signed sequence of the test measure differences to detect inconsistent behavior. We characterized the expected alarm rates for both magnitude and sign, which are dependent on the system model. Furthermore, we provide bounds on detection while also providing an analysis of the detection bounds of our scheme. The proposed approach was validated through simulations on a UGV case study. While our proposed Serial Detector can not replace traditional test measure-based detection schemes, however, it can provide another layer of security to detect hidden attacks that are deceptive to these state-of-the-art detectors.

Acknowledgments

This work is based on research sponsored by ONR under agreement number N000141712012, and NSF under grant #1816591.

References

[1] J. Bhatti and T. E. Humphreys, "Hostile control of ships via false gps signals: Demonstration and detection," *Navigation*, vol. 64, no. 1, pp. 51–66, 2017.

[2] C. Valasek and C. Miller, "Remote Exploitation of an Unaltered Passenger Vehicle," p. 93, 2015.

[3] A. Nourian and S. Madnick, "A systems theoretic approach to the security threats in cyber physical systems applied to stuxnet," *IEEE Transactions on Dependable and Secure Computing*, 2018.

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

[5] C. Murguia and J. Ruths, "On model-based detectors for linear time-invariant stochastic systems under sensor attacks," *IET Control Theory Applications*, vol. 13, no. 8, pp. 1051–1061, 2019.

[6] F. Miao, Q. Zhu, M. Pajic, and G. J. Pappas, "Coding sensor outputs for injection attacks detection," in *53rd IEEE Conference on Decision and Control*, Dec 2014, pp. 5776–5781.

[7] Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 93–109, 2015.

[8] P. J. Bonczek, S. Gao, and N. Bezzo, "Model-based randomness monitor for stealthy sensor attacks," in *2020 American Control Conference (ACC)*, 2020, pp. 2036–2042.

[9] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[10] C. Cammarota, "The difference-sign runs length distribution in testing for serial independence," *Journal of Applied Statistics*, vol. 38, no. 5, pp. 1033–1043, 2011.

[11] P. J. Bonczek and N. Bezzo, "Memoryless cumulative sign detector for stealthy cps sensor attacks," in *21st International Federation of Automatic Control (IFAC) World Congress*, 2020.

[12] S. M. Ross, *Introduction to Probability Models, Ninth Edition*. Orlando, FL, USA: Academic Press, Inc., 2006.

[13] E. Seneta, "Fitting the variance-gamma model to financial data," *Journal of Applied Probability*, vol. 41, no. A, p. 177–187, 2004.

[14] B. Klar, "A note on gamma difference distributions," *Journal of Statistical Computation and Simulation*, vol. 85, no. 18, pp. 3708–3715, 2015.

[15] A. Ferrari, "A note on sum and difference of correlated chi-squared variables," 2019. [Online]. Available: https://arxiv.org/abs/1906.09982

[16] S. I. Gass and C. M. Harris, "Encyclopedia of operations research and management science," *Journal of the Operational Research Society*, vol. 48, no. 7, pp. 759–760, 1997.

[17] S. Siegel, *Nonparametric statistics for the behavioral sciences*. McGraw-Hill New York, 1956.

[18] J. J. Nutaro, *Building software for simulation: theory and algorithms, with applications in C++*. John Wiley & Sons, 2011.