A Secure Optimum Distributed Detection Scheme in Under-Attack Wireless Sensor Networks

Edmond Nurellari, Des McLernon, *Member, IEEE*, and Mounir Ghogho, *Senior Member, IEEE*

***Abstract*—We address the problem of centralized detection of a binary event in the presence of β fraction falsifiable sensor nodes (SNs) (i.e., controlled by an attacker) for a bandwidth-constrained $under-attack$ spatially uncorrelated distributed wireless sensor network (WSN). The SNs send their one-bit test statistics over orthogonal channels to the fusion center (FC), which linearly combines them to reach to a final decision. Adopting the modified deflection coefficient as an alternative function to be optimized, we first derive in a closed-form the FC optimal weights combining. But as these optimal weights require $a-priori$ knowledge that cannot be attained in practice, this optimal weighted linear FC rule is not implementable. We also derive in a closed-form the expressions for the attacker "flipping probability" (defined in paper) and the minimum fraction of compromised SNs that makes the FC incapable of detecting. Next, based on the insights gained from these expressions, we propose a novel and non-complex reliability-based strategy to identify the compromised SNs and then adapt the weights combining proportional to their assigned reliability metric. In this way, the FC identifies the compromised SNs and decreases their weights in order to reduce their contributions towards its final decision. Finally, simulation results illustrate that the proposed strategy significantly outperforms (in terms of FC's detection capability) the existing compromised SNs identification and mitigation schemes.**

***Index Terms*—Distributed detection, optimum fusion rule, falsified sensor nodes (SNs) observations, wireless sensor networks (WSNs).**

I. INTRODUCTION

CENTRALIZED detection of a binary event is one of the most important applications of wireless sensor networks (WSNs) [1], [2]. Deployed over a field, multiple coordinated SNs report their processed observations to a fusion center (FC). Then, upon receiving all the contributions from each SN, the FC optimally combines them to declare a global decision. Unfortunately, these tiny devices suffer from constrained bandwidth and limited available on-board power. Furthermore, the geographically distributed nature of such a system makes them quite vulnerable to a different type of attack. Hence, incorporating security into WSNs has been a challenging task. Like all other networks [3], WSNs are also vulnerable to various security issues. Furthermore, the local SNs decision process (i.e., local detection performance) itself is subject to various security threats. The detection performance strongly depends on the reliability of these SNs in the network. While fusing the data received by the spatially deployed SNs allows making a reliable FC decision with respect to the status of the phenomena, it is possible that one or more SNs (compromised by an attacker) deliberately falsify their local observations to degrade the FC detection performance. However, there are a number of different strategies as to how the test statistics received from each SN will be efficiently used in order to arrive at a reliable FC final decision. We will first give a brief review on the related work before introducing our proposed approach.

The framework of distributed detection under $attack-free$ WSNs has been extensively studied in [4]-[14], to name but just a few. While references [4]-[8] consider distributed detection by assuming unlimited bandwidth/resources in WSNs, the authors of [9]-[14] relax this assumption by considering distributed detection over bandwidth-constrained/energy-constrained WSNs. But these approaches are vulnerable to security attacks as some of the SNs reporting to the FC may be compromised. As a result, the FC is not robust against such attacks and its detection performance will be degraded.

Now, security vulnerabilities can be exploited by different types of attacks that can be launched in a WSN, for example, jamming, spoofing, wiretap disruption attacks, etc [15]. Apart from these well-known traditional security threats, several recent studies consider the sensor node data falsification (SNDF) attack (known as a Byzantine attack, eg., [16], [17]). The Byzantine attack was first proposed by [18] and later widely used in the context of distributed detection ([16], [19], [20] and see references therein). In this work, we also consider the SNDF attack in which the compromised SNs send wrong local decision reports to the FC either to degrade the FC detection performance or to achieve their selfish greedy objectives.

The reported work on distributed detection over $attack-free$ WSNs is relatively high but there is limited consideration for $under-attack$ WSNs, see for example, [16]-[21] and references therein. In [21], a probabilistic test statistic falsification (TSF) attack is proposed and theoretical performance evaluation (in terms of destructiveness and stealthiness) is obtained. The authors of [22], in the context of smart grids, propose heuristic centralized algorithms to derive various strategies (attacker versus defender dynamics). Then, a distributed algorithm is proposed that guarantees convergence to the centralized solution taken at the FC. Reference [23], in the

This work was partly supported by the Engineering and Physical Sciences Research Council through the Project Shaking Tunnel Vision under Grant EP/N03435X/1.

E. Nurellari is with School of Engineering, University of Lincoln, Brayford Pool, Lincoln, LN6 7TS, UK., and also with the School of Electronic and Electrical Engineering, University of Leeds, LS2 9JT, Leeds, U.K. (e-mail: elen@leeds.ac.uk).

D. McLernon is with the School of Electronic and Electrical Engineering, University of Leeds, LS2 9JT, Leeds, U.K. (e-mail: d.c.mclernon@leeds.ac.uk).

M. Ghogho is with the School of Electronic and Electrical Engineering, University of Leeds, Leeds LS2 9JT, U.K., and also with the International University of Rabat, Rabat 11 100, Morocco (e-mail: m.ghogho@leeds.ac.uk).

context of cognitive radio (CR), proposed a prefiltering scheme of sensing data and a trust factor is assigned to each user to detect the malicious CR ones. The authors of [24], in the context of target localization, also consider binary Byzantine attacks where the SNs transmit to the FC their binary decisions and they propose two techniques to mitigate the compromised SNs negative impact on the FC decision. To mitigate the Byzantine effect on the data fusion problem in cooperative spectrum sensing, a weighted sequential probability ratio test was proposed in [25]. However, these schemes require $a-priori$ information and/or due to the high computational complexity are not always feasible in the context of WSNs. In [26], a reputation-based scheme is proposed for identifying the compromised SNs by accumulating the deviations between each SN's decision and the FC's decision over a time window duration. Then, the identified compromised SNs are totally excluded from the data fusion process. Different from [26], the authors in [27] use the FC's decision as an evaluation basis to assign to each SN a reputation measure, classifying each SN as either reliable, partially reliable or malicious. In this way, the SNs classified as malicious will be excluded from the fusion process (i.e., assigned zero weight), and the one decided on as reliable will be assigned a unity weight and the partially reliable ones are assigned a 0.5 weight. However, identifying and then totally excluding the compromised SNs contributions from the FC decision process may not be the best strategy. For instance, we might end up excluding SNs contributing towards the FC global decision that might have high local signal-to-noise-ratios (SNRs). Recently, the authors in [19], [28] both consider a decentralized network in the presence of compromised SNs while in this paper we consider a centralized scheme. The authors in [19] propose a synchronous distributed weighted average consensus algorithm that is claimed to be robust to Byzantine attacks while reference [28] considers the detection and mitigation of data injection attacks in a randomized average consensus.

So, this work investigates the detection performance of an $under-attack$ WSN. To reduce the transmission and processing burden of the SNs, each SN generates the 1-bit local test statistic by performing energy detection [29] and reports this test statistic to the FC. As in [26], we relax the assumption of perfect knowledge of the true hypothesis [16] and we assume that the compromised SNs (controlled by the attacker) do not know the true state of the target. For the FC, we assume that it is not compromised and receives the test statistic from both types of SNs (i.e., compromised and honest). The transmission (SNs to FC) links are assumed error free (see eg., [16], [26]).

A. Contributions $\&$ Oragnization

Our main contributions are as follows:

(i) First, we develop an efficient FC linear weight combing framework for an $under{-}attack$ WSN. To further reduce the optimization complexity and to get an insight into the problem, we adopt the modified deflection coefficient (MDC) [7] as an alternative function to be optimized. Based on this (i.e., the MDC), we provide an optimization problem to be solved from both the FC's and the attacker's perspective. From the FC's perspective, we derive analytically (in a closed form) the optimal weight combiner for each SN. We show that these weights are a function of the local SNs probability of false alarm and probability of detection metrics as well as the SNs local test statistics "flipping probability" (to be defined later). Unfortunately, for the compromised SNs this $a\ priori$ knowledge cannot be obtained in practice (we propose a solution to this (see later (ii))). Then (from the attacker's perspective), we derive analytically (for a fixed number of compromised SNs) the optimum attacker local test statistics flipping probability and the minimum fraction of the compromised SNs that makes the FC incapable of detecting.

(ii) Next, based on this framework (i.e., FC linear weight combing strategy), we also propose a new non-complex and efficient (based on a reliability metric) FC detection scheme to identify the compromised SNs. Our approach is different from the existing approaches [16], [26], [27] in two important aspects: 1) We introduce a new reliability metric at the FC to identify the compromised SNs. First, we count the inconsistency between the FC's decision (where all the SNs contributions are considered) and the i^{th} local SN's decision over a time window. Similarly, we then count the inconsistency between the FC's decision (where the i^{th} SN contribution is not considered) and the i^{th} local SN's decision. Finally, the proposed reputation metric is evaluated as the difference between these two parameters; 2) Then, based on this reputation metric, we propose a novel FC weight computation strategy that ensures the following: a) for the identified compromised SNs, their weights are likely to be decreased proportionally to this metric (where the existing schemes assign a zero weight). b) In this way (based on this new reputation metric), the FC decides how much a SN should contribute to its final decision. We will show that this strategy outperforms the existing schemes where the identified compromised SNs are totally excluded from the FC final decision contribution (i.e., a zero weight is assigned).

Now, the summary of the paper is as follows. In Section II we describe the system model (SN sensing and local decision) and describe the compromised SNs attack model. Section III introduces the simplified linear weighted fusion rule and analyzes the optimization problem from both the FC's and the attacker's perspective. In Section IV we present our proposed compromised SNs identification metric and weight combining computation strategy. Finally, section V presents simulation results and in Section VI we give our conclusions.

II. Problem Formulation

Consider the problem of detecting the presence of an unknown but deterministic signal $s(n)$ by an $under-attack$ WSN consisting of M geographically distributed SNs and a FC (see Fig. 1). The honest SNs are represented with a black color and the compromised SNs (i.e., the ones controlled by the attacker) with a red color. The attacker's aim is to successfully manipulate the FC global decision making process while the FC would like to detect reliably (i.e., with very high probability). Next, we explain in more detail the sensing, the local decision and the compromised SNs attack model.

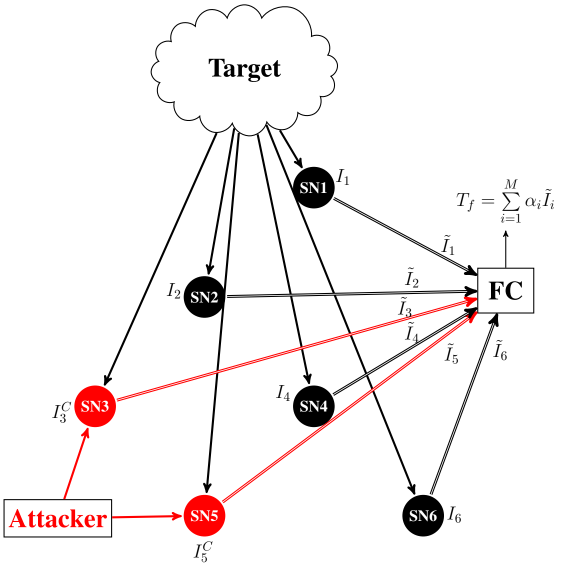


Fig. 1. Under attack schematic communication architecture between peripheral SNs and the fusion center (FC). Each of the i^{th} honest/compromised SNs represented with a black/red color generates a local (binary) indicator variable (I_i/I_i^C) by observing the target and performing the test in (5) with local detection threshold Λ/Λ_C. While the i^{th} $(i = \{1, 2, 4, 6\})$ honest SN indicator (test statistic) remains unchanged (i.e., $\tilde{I}_i = I_i$), the j^{th} $(j = \{3, 5\})$ compromised SN falsifies its indicator (test statistic) as in (8) before transmitting to the FC. Here i/j are the honest/compromised SN index.

A. Sensing

The measured signal at SN i is either:

$$\mathcal{H}_0 : y_i(n) = w_i(n) \quad (1)$$

$$\mathcal{H}_1 : y_i(n) = s_i(n) + w_i(n) \quad (2)$$

and energy estimation is performed at the i^{th} SN to give

$$T_i = \sum_{n=1}^{N} (y_i(n))^2, \ i = 1, 2, \ldots, M \quad (3)$$

which for large N has an approximately Gaussian distribution [29]. Furthermore, the noise samples are assumed to be identically and independently distributed (i.i.d.) across time and space. It is not difficult to show that

$$\begin{aligned} &\mathbb{E}\{T_i|\mathcal{H}_0\} = N\sigma_i^2, \qquad \mathrm{Var}\{T_i|\mathcal{H}_0\} = 2N\sigma_i^4 \\ &\mathbb{E}\{T_i|\mathcal{H}_1\} = N\sigma_i^2(1+\xi_i), \mathrm{Var}\{T_i|\mathcal{H}_1\} = 2N\sigma_i^4(1+2\xi_i) \quad (4) \end{aligned}$$

where $\xi_i = \sum_{n=1}^{N} s_i^2(n)/N\sigma_i^2$.

B. Local Decision

Based on its local energy estimation (3), the i^{th} SN generates a binary indicator random variable I_i as follows:

$$\left.\begin{array}{l} \text{if } T_i < \Lambda, I_i = 0 \implies \text{decide } \mathcal{H}_0 \\ \text{if } T_i \geq \Lambda, I_i = 1 \implies \text{decide } \mathcal{H}_1 \end{array}\right\} \quad (5)$$

where Λ is a local detection threshold that is the same for all the M SNs. The i^{th} SN local probability of false alarm (p_{fa}^i) and the local probability of detection (p_d^i) can be expressed as:

$$\begin{aligned} p_{fa}^i &= \Pr(T_i \geq \Lambda|\mathcal{H}_0) = Q\left(\frac{\Lambda - \mathbb{E}\{T_i|\mathcal{H}_0\}}{\sqrt{\mathrm{Var}\{T_i|\mathcal{H}_0\}}}\right) \\ p_d^i &= \Pr(T_i \geq \Lambda|\mathcal{H}_1) = Q\left(\frac{\Lambda - \mathbb{E}\{T_i|\mathcal{H}_1\}}{\sqrt{\mathrm{Var}\{T_i|\mathcal{H}_1\}}}\right) \quad (6) \end{aligned}$$

where $Q(.)$ is the Q-function. While the i^{th} honest SN transmits its actual one-bit test statistic (i.e., I_i in (5)) to the FC, the compromised SNs falsify them before transmitting to the FC. Next we introduce the attacker model.

C. Compromised SNs Attack

Different attack strategies could be adopted by the compromised SNs. In this work, the data falsification attack model widely used in [16], [20], [26] is considered. There are β fraction of SNs controlled and compromised by the attacker (the attacker controls the local detection threshold, the flipping probability, and the fraction β, all to be defined later). As before, (i.e., in the case of $attack - free$) each of the i^{th} compromised SNs perform the local test in (5) but now with a local detection threshold (Λ_C) controlled by the attacker and assumed to be the same for all the β fraction compromised SNs. That is:

$$\left.\begin{array}{l} \text{if } T_i < \Lambda_C, I_i^C = 0 \implies \text{decide } \mathcal{H}_0 \\ \text{if } T_i \geq \Lambda_C, I_i^C = 1 \implies \text{decide } \mathcal{H}_1. \end{array}\right\} \quad (7)$$

Now, the probability of false alarm[1] $(p_{fa}^{i,C})$ and the probability of detection $(p_d^{i,C})$ at the i^{th} compromised SN are respectively given as in (6) with $\Lambda = \Lambda_C$, while for the honest SNs it remains as in (6). After performing the test in (7), the compromised SNs further manipulate their binary indicator variables prior to FC transmission so as to yield the maximum possible FC degradation. Let P_C^{flip} be the probability that each compromised SN intentionally reports the opposite information to its actual local decision (i.e., flips the indicator random variable in (7) prior to FC transmission with probability P_C^{flip}). It is assumed that all the compromised SNs have the same probability of attack in a particular sensing period (see section IV for details). The remaining (1-β fraction) SNs are "honest" and report to the FC accordingly. Now, the i^{th} local binary indicator test statistic for the compromised SNs can be expressed as:

$$\tilde{I}_i = \begin{cases} 1 - I_i^C, & \text{with probability } P_C^{flip} \\ I_i^C, & \text{with probability } (1 - P_C^{flip}) \end{cases} \quad (8)$$

while for the honest SNs this relation is simply $\tilde{I}_i = I_i$. Similarly, the i^{th} compromised SN local probability of false alarm and the probability of detection can be shown to be

[1] Here the superscripts "i, C" refer to the i^{th} compromised SN.

$$\mathbf{R} = \big(1-\beta\big)\mathrm{diag}\begin{bmatrix} p_d^1\big(1-p_d^1\big) + \frac{\beta}{1-\beta}\Big(P_C^{flip} + p_d^{1,C}\big(1-2P_C^{flip}\big)\Big)\Big(1-P_C^{flip} + p_d^{1,C}\big(2P_C^{flip}-1\big)\Big) \\ p_d^2\big(1-p_d^2\big) + \frac{\beta}{1-\beta}\Big(P_C^{flip} + p_d^{2,C}\big(1-2P_C^{flip}\big)\Big)\Big(1-P_C^{flip} + p_d^{2,C}\big(2P_C^{flip}-1\big)\Big) \\ \vdots \\ p_d^M\big(1-p_d^M\big) + \frac{\beta}{1-\beta}\Big(P_C^{flip} + p_d^{M,C}\big(1-2P_C^{flip}\big)\Big)\Big(1-P_C^{flip} + p_d^{M,C}\big(2P_C^{flip}-1\big)\Big) \end{bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_M \end{bmatrix} \quad (18)$$

respectively:

$$\begin{aligned} \tilde{p}_{fa}^i &= P_C^{flip}\left(1-p_{fa}^{i,C}\right) + \left(1-P_C^{flip}\right)p_{fa}^{i,C} \\ \tilde{p}_d^i &= P_C^{flip}\left(1-p_d^{i,C}\right) + \left(1-P_C^{flip}\right)p_d^{i,C} \end{aligned} \quad (9)$$

while for the honest SNs clearly $\tilde{p}_{fa}^i = p_{fa}^i$ and $\tilde{p}_d^i = p_d^i$. Next, we introduce a simplified (optimum) linear fusion rule at the FC.

III. Simplified Fusion Rule-The Linear Approach

Now, the i^{th} SN transmits to the FC the one-bit local test statistic ($\tilde{I}_i$). The communication channels between SNs and the FC are assumed to be error-free in this paper. Upon receiving all the contributions from all the SNs (i.e., compromised and honest), the FC linearly combines them:

$$T_f = \sum_{i=1}^{M} \alpha_i \tilde{I}_i \quad (10)$$

where $\{\alpha_i\}_{i=1}^M$ are the optimum weights that we will derive later in section III-A. The FC then makes the final decision:

$$\left.\begin{array}{l} \text{if } T_f < \Lambda_f, \text{ decide } \mathcal{H}_0 \\ \text{if } T_f \geq \Lambda_f, \text{ decide } \mathcal{H}_1 \end{array}\right\} \quad (11)$$

where Λ_f is the FC detection threshold. Let

$$\begin{aligned} P_d &= \Pr\big(T_f \geq \Lambda_f|\mathcal{H}_1\big) \\ P_{fa} &= \Pr\big(T_f \geq \Lambda_f|\mathcal{H}_0\big) \end{aligned} \quad (14)$$

where P_d and P_{fa} are the system probability of detection and probability of false alarm respectively. For large M, T_f can be approximated by a Gaussian distribution and the P_d for a fixed P_{fa} is given as [30]:

$$P_d = Q\left(\frac{Q^{-1}\left(P_{fa}\right)\sqrt{\mathrm{Var}\left\{T_f|\mathcal{H}_0\right\}} - \mathbb{E}\left\{T_f|\mathcal{H}_1\right\} + \mathbb{E}\left\{T_f|\mathcal{H}_0\right\}}{\sqrt{\mathrm{Var}\left\{T_f|\mathcal{H}_1\right\}}}\right) \quad (15)$$

with appropriate quantities given in (12)-(13).

A. Weight Combining Optimisation

In this section, we would like to find the optimum weighting vector ($\boldsymbol{\alpha}_{opt}$) that maximizes (15). However, maximizing (15) w.r.t. $\boldsymbol{\alpha}$ is difficult and no closed form solution can be found. So we will approximate the optimal solution by adopting the modified deflection coefficient[2] (MDC) [7] as an alternative function to be maximized. This is given as:

$$\tilde{d}^2\left(\boldsymbol{\alpha}\right) = \left(\frac{\mathbb{E}\left\{T_f|\mathcal{H}_1\right\} - \mathbb{E}\left\{T_f|\mathcal{H}_0\right\}}{\sqrt{\mathrm{Var}\left\{T_f|\mathcal{H}_1\right\}}}\right)^2 = \frac{\left(\boldsymbol{b}^T\boldsymbol{\alpha}\right)^2}{\boldsymbol{\alpha}^T\mathbf{R}\boldsymbol{\alpha}} \quad (16)$$

where

$$\boldsymbol{b} = \begin{bmatrix} \big(1-\beta\big)\big(p_d^1 - p_{fa}^1\big) - \beta\big(p_d^{1,C} - p_{fa}^{1,C}\big)\big(2P_C^{flip}-1\big) \\ \big(1-\beta\big)\big(p_d^2 - p_{fa}^2\big) - \beta\big(p_d^{2,C} - p_{fa}^{2,C}\big)\big(2P_C^{flip}-1\big) \\ \vdots \\ \big(1-\beta\big)\big(p_d^M - p_{fa}^M\big) - \beta\big(p_d^{M,C} - p_{fa}^{M,C}\big)\big(2P_C^{flip}-1\big) \end{bmatrix} \quad (17)$$

and $\mathbf{R}$ and $\boldsymbol{\alpha}$ are given in (18). Now, our optimization problem is:

$$\boldsymbol{\alpha}_{opt} = \arg \max_{\boldsymbol{\alpha}}\Big(\tilde{d}^2\left(\boldsymbol{\alpha}\right)\Big). \quad (19)$$

Further, via the transformation $\boldsymbol{\psi} = \mathbf{R}^{1/2}\boldsymbol{\alpha}$, the deflection coefficient (16) becomes:

$$\tilde{d}^2\left(\boldsymbol{\psi}\right) = \frac{\boldsymbol{\psi}^T\mathbf{D}\boldsymbol{\psi}}{||\boldsymbol{\psi}||^2}, \ \mathbf{D} = \mathbf{R}^{-T/2}\boldsymbol{b}\boldsymbol{b}^T\mathbf{R}^{-1/2}. \quad (20)$$

So $\boldsymbol{\alpha}_{opt} = \mathbf{R}^{-1/2}\boldsymbol{\psi}_{opt} = k\mathbf{R}^{-1}\boldsymbol{b}$, where $\boldsymbol{\psi}_{opt} = k\mathbf{R}^{-1/2}\boldsymbol{b}$ is the eigenvector corresponding to the maximum eigenvalue of $\mathbf{D}$. Now, the optimum weight combining in (10) can be easily shown to be (21). Clearly, the optimum weights depend on the local probability of false alarm and the probability of detection metrics as well as on the β (fraction of compromised SNs) and the probability of flipping the local decisions by the attacker. For the SNs that are honest (i.e., controlled by the FC) these local probabilities are known (since the FC can set the local detection threshold itself). However, for the compromised SNs these local probabilities are not available at the FC (since the attacker takes control of the local detection threshold). To make the matters worse, the FC knows just the fraction of compromised SNs (i.e., β) but it cannot identify who they are. As a result, the FC cannot implement the optimum weight combining fusion rule (10).

Later, in section IV, we propose a simple but yet effective approach to possibly identify these compromised SNs and compute the optimum weights at the FC, based on their assigned reliability. Next, we derive the optimum attacker flipping probability (P_C^{flip}) which makes the FC incapable of detecting.

[2] In order to get insight into the system design parameters of the detection scheme, in this paper we adopt the MDC. This is due to its simplicity and close relationship with the detection performance. In general, P_d is a monotonically increasing function of the MDC and yields a good approximation in characterizing the detection performance.

$$\begin{aligned}\mathbb{E}\left\{T_f|\mathcal{H}_1\right\} &= (1-\beta)\sum_{i=1}^{M}\alpha_i p_d^i + \beta\left[P_C^{flip}\Big(\sum_{i=1}^{M}\alpha_i\big(1-p_d^{i,C}\big)\Big) + \big(1-P_C^{flip}\big)\Big(\sum_{i=1}^{M}\alpha_i p_d^{i,C}\Big)\right]\\ \mathbb{E}\left\{T_f|\mathcal{H}_0\right\} &= (1-\beta)\sum_{i=1}^{M}\alpha_i p_{fa}^i + \beta\left[P_C^{flip}\Big(\sum_{i=1}^{M}\alpha_i\big(1-p_{fa}^{i,C}\big)\Big) + \big(1-P_C^{flip}\big)\Big(\sum_{i=1}^{M}\alpha_i p_{fa}^{i,C}\Big)\right]. \end{aligned} \tag{12}$$

$$\begin{aligned}\mathrm{Var}\left\{T_f|\mathcal{H}_1\right\} = (1-\beta)\sum_{i=1}^{M}\alpha_i^2 p_d^i\Big(1-p_d^i\Big) + \beta\Bigg[\Bigg(P_C^{flip}\Big(\sum_{i=1}^{M}\alpha_i^2\big(1-p_d^{i,C}\big)\Big) + \big(1-P_C^{flip}\big)\Big(\sum_{i=1}^{M}\alpha_i^2 p_d^{i,C}\Big)\Bigg)\\ \Bigg(1-P_C^{flip}\Big(\sum_{i=1}^{M}\alpha_i^2\big(1-p_d^{i,C}\big)\Big) - \big(1-P_C^{flip}\big)\Big(\sum_{i=1}^{M}\alpha_i^2 p_d^{i,C}\Big)\Bigg)\Bigg]\end{aligned}$$

$$\begin{aligned}\mathrm{Var}\left\{T_f|\mathcal{H}_0\right\} = (1-\beta)\sum_{i=1}^{M}\alpha_i^2 p_{fa}^i\Big(1-p_{fa}^i\Big) + \beta\Bigg[\Bigg(P_C^{flip}\Big(\sum_{i=1}^{M}\alpha_i^2\big(1-p_{fa}^{i,C}\big)\Big) + \big(1-P_C^{flip}\big)\Big(\sum_{i=1}^{M}\alpha_i^2 p_{fa}^{i,C}\Big)\Bigg)\\ \Bigg(1-P_C^{flip}\Big(\sum_{i=1}^{M}\alpha_i^2\big(1-p_{fa}^{i,C}\big)\Big) - \big(1-P_C^{flip}\big)\Big(\sum_{i=1}^{M}\alpha_i^2 p_{fa}^{i,C}\Big)\Bigg)\Bigg].\end{aligned} \tag{13}$$

$$\alpha_{opt}^i = \frac{\big(1-\beta\big)\big(p_d^i - p_{fa}^i\big) + \beta\big(p_{fa}^{i,C} - p_d^{i,C}\big)\big(2P_C^{flip}-1\big)}{\big(1-\beta\big)\big(p_d^i\big(1-p_d^i\big)\big) + \beta\Big(P_C^{flip} + p_d^{i,C}\big(1-2P_C^{flip}\big)\Big)\Big(1-P_C^{flip}+p_d^{i,C}\big(2P_C^{flip}-1\big)\Big)}. \tag{21}$$

B. Attacker Flipping Probability Optimisation

So what is the optimum P_C^{flip} that the attacker needs to adopt for the compromised SNs in order to cause the maximum possible degradation to the FC (i.e., to possibly make the FC incapable of detecting)? Again, we use the modified deflection coefficient as an alternative function to be optimized and assume that the FC does not act strategically against the attacker strategy.

Lemma 1: The optimum flipping probability $\big(P_{C,opt}^{flip}\big)$ which minimizes the modified deflection coefficient is:

$$P_{C,opt}^{flip} = \frac{\beta-1}{2\beta}\left(\frac{\sum_{i=1}^{M}\alpha_i\big(p_d^i - p_{fa}^i\big)}{\sum_{i=1}^{M}\alpha_i\big(p_{fa}^{i,C} - p_d^{i,C}\big)}\right) + \frac{1}{2}. \tag{22}$$

Proof. Since the modified deflection coefficient is always non-negative, then its minimum is always greater than or equal to zero. So, the condition to make the minimum of the modified deflection coefficient zero is:

$$\begin{aligned}\boldsymbol{b}^T\boldsymbol{\alpha} = \big(1-\beta\big)\sum_{i=1}^{M}\alpha_i\big(p_d^i - p_{fa}^i\big) + \beta P_C^{flip}\sum_{i=1}^{M}\alpha_i\big(p_{fa}^{i,C} - p_d^{i,C}\big)\\ + \beta\big(1-P_C^{flip}\big)\sum_{i=1}^{M}\alpha_i\big(p_d^{i,C} - p_{fa}^{i,C}\big) = 0.\end{aligned} \tag{23}$$

Further simplification of the above and re-arrangement of the terms yields:

$$\begin{aligned}\beta\Big(\sum_{i=1}^{M}\alpha_i\big(p_{fa}^{i,C} - p_d^{i,C}\big)\Big)\Big(2P_C^{flip}-1\Big) = (\beta-1)\sum_{i=1}^{M}\alpha_i\big(p_d^i - p_{fa}^i\big)\\ \implies P_{C,opt}^{flip} = \frac{\beta-1}{2\beta}\left(\frac{\sum_{i=1}^{M}\alpha_i\big(p_d^i - p_{fa}^i\big)}{\sum_{i=1}^{M}\alpha_i\big(p_{fa}^{i,C} - p_d^{i,C}\big)}\right) + \frac{1}{2}.\end{aligned} \tag{24}$$

This concludes the proof. ■

In the special case when the attacker does not change the local detection threshold in (7) (i.e., $p_d^i = p_d^{i,C}$ and $p_{fa}^i = p_{fa}^{i,C}$), then the optimum probability of flipping the local decisions can be shown to be:

$$P_{C,opt}^{flip} = \begin{cases} \frac{1}{2} - \frac{\beta-1}{2\beta} = \frac{1}{2\beta}, & \text{for } 0.5 \leq \beta \leq 1 \\ \text{not applicable}, & \text{for } \beta = 0 \\ \text{not defined}, & \text{otherwise.} \end{cases} \tag{25}$$

Interestingly, in this case the optimum probability of flipping the local SNs decision is inversely proportional to the fraction of the compromised SNs (β). As expected, when β increases, the optimum probability of flipping the local decision in order to make the MDC zero decreases and vice-versa. Furthermore, when the half of the network is compromised (i.e., $\beta = 0.5$), the attacker can make the modified deflection coefficient zero with $P_{C,opt}^{flip} = 1$ (i.e., the local SNs should always flip their local decisions).

C. Minimum Fraction of Compromised SNs

Now, we are interesting in the minimum fraction of the compromised SNs that is needed to cause the maximum possible degradation to the FC. We state the result in the next Lemma.

Lemma 2: The minimum fraction of the compromised SNs needed to make the FC incapable of detecting or to make the modified deflection coefficient zero is $\beta_{min} \geq \frac{1}{2}$.

Proof. As we previously stated, the modified deflection coefficient is always non-negative and the minimum occurs at zero. From (23), the condition to make the modified deflection coefficient zero is:

$$\boldsymbol{b}^T\boldsymbol{\alpha} = \big(1-\beta\big)\sum_{i=1}^{M}\alpha_i\big(p_d^i - p_{fa}^i\big) + \beta P_C^{flip}\sum_{i=1}^{M}\alpha_i\big(p_{fa}^{i,C} - p_d^{i,C}\big) + \beta\big(1 - P_C^{flip}\big)\sum_{i=1}^{M}\alpha_i\big(p_d^{i,C} - p_{fa}^{i,C}\big) = 0. \quad (26)$$

After simplifying the above equation, the condition on β needed to make the FC incapable of detecting becomes:

$$\beta = \left(1 - \underbrace{\left(\frac{\Big(\sum_{i=1}^{M}\alpha_i\big(p_d^{i,C} - p_{fa}^{i,C}\big)\Big)\Big(1 - 2P_C^{flip}\Big)}{\sum_{i=1}^{M}\alpha_i\big(p_d^i - p_{fa}^i\big)}\right)}_{\text{(A)}}\right)^{-1}. \quad (27)$$

Now, the minimum of β (β_{min}) can be achieved when term (A) of (27) is minimum. We also know that for any real scalar a and b the following holds:

$$\min\big(\frac{a}{b}\big) \geq \frac{\min\big(a\big)}{\max\big(b\big)}. \quad (28)$$

Using (27) and (28), we now derive a lower bound on the minimum β. Clearly, we require that both the numerator and the denominator of term (A) take the minimum and the maximum values respectively. Now, the minimum of the numerator (i.e., $\min\bigg(\big(\sum_{i=1}^{M}\alpha_i\big(p_d^{i,C} - p_{fa}^{i,C}\big)\big)\big(1 - 2P_C^{flip}\big)\bigg)$) can be achieved if both $p_d^{i,C} = P_C^{flip} = 0$ and $p_{fa}^{i,C} = 1$ or alternatively when both $p_d^{i,C} = P_C^{flip} = 1$ and $p_{fa}^{i,C} = 0$. Similarly, the maximum of the denominator of term (A) (i.e., $\max\bigg(\sum_{i=1}^{M}\alpha_i\big(p_d^i - p_{fa}^i\big)\bigg)$) can be achieved when both $p_d^i = 1$ and $p_{fa}^i = 0$. Finally, using the above analysis we can easily show that:

$$\beta_{min} \geq \frac{1}{2}. \quad (29)$$

This concludes the proof. ■

In the special case when the attacker does not change the local detection threshold in (7) (i.e., $p_d^i = p_d^{i,C}$ and $p_{fa}^i = p_{fa}^{i,C}$), the minimum fraction of compromised SNs required to make the modified deflection coefficient zero (i.e., make the FC incapable of detecting) can be shown to be: $\beta_{min} = \frac{1}{2}$ and this can be achieved with $P_C^{flip} = 1$ (see (25)).

IV. Compromised SNs Identification and weight combining computation

In this section, we propose a scheme to identify the compromised SNs and compute the weight combining in (10) based on each SN assigned reliability. As in [26] and [27], we divide the local sensing process into time windows consisting of K sensing periods[3].

A. Compromised SNs Identification

At the fusion center, the received observations corresponding to the i^{th} SNs can be expressed as $\tilde{\boldsymbol{I}}_i = [\tilde{I}_i(1), \tilde{I}_i(2), \cdots, \tilde{I}_i(K)], \forall i = 1, 2, \cdots, M$. At the l^{th} sensing period, upon receiving the contributions from all the SNs (i.e., compromised and honest) the FC linearly combines them to yield:

$$T_f(l) = \sum_{j=1}^{M}\alpha_j^{AF}\tilde{I}_j(l), \quad l = 1, 2, \cdots, K$$
$$T_f^i(l) = \sum_{j=1, i\neq j}^{M}\alpha_j^{AF}\tilde{I}_j(l), \; l = 1, 2, \cdots, K, \; i = 1, 2, \cdots, M \quad (30)$$

where $T_f^i(l)$ is the final test statistic at the l^{th} sensing period without the contribution of the i^{th} SN; and $\{\alpha_j^{AF}\}_{j=1}^{M}$ are the optimum weights under an attack-free scenario and can be easily derived from (21) by substituting ($\beta = 0$, $P_C^{flip} = 0$, $p_{fa}^{i,C} = p_{fa}^i$ and $p_d^{i,C} = p_d^i$, $\forall i$). These can be shown to be:

$$\alpha_j^{AF} = \frac{p_d^j - p_{fa}^j}{p_d^j\big(1 - p_d^j\big)}. \quad (31)$$

Based on the test statistics (30), the FC then generates at the l^{th} sensing period two different indicator random variables as follows:

$$I_f(l) = \begin{cases} 0 \text{ if } T_f(l) < \Lambda_f \\ 1 \text{ if } T_f(l) \geq \Lambda_f \end{cases} \qquad I_f^i(l) = \begin{cases} 0 \text{ if } T_f^i(l) < \Lambda_f \\ 1 \text{ if } T_f^i(l) \geq \Lambda_f. \end{cases} \quad (32)$$

Now that the FC has evaluated these two indicator random variables (i.e., $I_f(l)$ and $I_f^i(l)$), it then compares them to the i^{th} SN local indicator variable $\tilde{I}_i(l)$ to yield:

$$d_i(l) = \begin{cases} 1 \text{ if } I_f(l) \neq \tilde{I}_i(l) \\ 0 \quad\quad \text{otherwise} \end{cases} \qquad \hat{d}_i(l) = \begin{cases} 1 \text{ if } I_f^i(l) \neq \tilde{I}_i(l) \\ 0 \quad\quad \text{otherwise} \end{cases} \quad (33)$$

where $d_i(l)$ represents the inconsistency between the FC's decision (all the SNs contributions are counted) and the i^{th} SN local decision. Similarly, $\hat{d}_i(l)$ represents the same but now the i^{th} SN is not considered at the FC decision. Note that all of the above steps are performed during the same time

[3] Each SN samples N times (see (3)) in each sensing interval and then performs the energy detection as in (5).

window K. After observing the reports for up to K sensing periods, the FC evaluates a reliability metric for the i^{th} SN:

$$r_i = \frac{1}{K}\left|\sum_{l=1}^{K}\left(d_i(l) - \hat{d}_i(l)\right)\right|, \quad i = 1, 2, \cdots, M. \tag{34}$$

It is worth mentioning that the r_i's for the compromised SNs are expected to be larger than those for the honest ones (see simulations results section later). Finally, the FC performs the reliability test:

$$\left.\begin{array}{ll} \text{if } r_i < \delta, & \text{decide reliable} \\ \text{if } r_i \geq \delta, & \text{decide not reliable} \end{array}\right\} \tag{35}$$

where δ is the reliability detection threshold. Now, the probability that a compromised SN has been $truly$ detected and the probability that an honest SN has been $falsely$ detected at the i^{th} SN are respectively:

$$\begin{aligned} P_d^{i,true} &= \Pr\big(r_i \geq \delta | Compromised\big) \\ P_d^{i,false} &= \Pr\big(r_i \geq \delta | Honest\big) \end{aligned} \tag{36}$$

where the superscript "$i, true$" and "$i, false$" represents the $true$ and $false$ detection at the i^{th} SN respectively. Obviously, the compromised SNs detection performance depends on the choice of the reliability detection threshold (δ). If we choose a large δ, $P_d^{i,false}$ is expected to be low. However, this also will result in a low $P_d^{i,true}$. On the other hand, choosing a lower δ it will increase the $P_d^{i,true}$ value but also an increase in $P_d^{i,false}$ will be noticed. Clearly, the reliability detection threshold imposes a trade-off between these two metrics. Note that in practice we wish to keep $P_d^{i,false}$ close to zero and $P_d^{i,true}$ close to one. Based on this reliability test (i.e., the test in (35)), next we will evaluate the weight combining in (10) such that the probability of detection in (15) is further improved.

B. Proposed Weight Combining Computation

In this section, we propose a weight combining computation based on the reliability test (35). Existing schemes use reliability-based metrics to possibly identify the compromised SNs and then totally exclude them from contributing to the FC process and decision. However, identifying and then excluding them from the detection process is not the optimum solution. For instance, we might end up removing (from contributing towards the global decision) compromised SNs that hold useful information in general (for example those SNs with high local SNRs). Different from the existing approaches, here we propose to update the weight combining (i.e., (31)) of each SN based on the $correctness$ of information reported to the FC. That is:

$$\alpha_i^{AF} = \begin{cases} \alpha_i^{AF} & \text{if } r_i < \delta \\ \alpha_i^{AF} - \mu r_i & \text{if } r_i \geq \delta \end{cases} \tag{37}$$

where $\mu \in [0, \infty]$ is the weight penalty that is the same for all the M SNs. For those SNs that are identified as being compromised by the attacker, the FC is likely to decrease their weights. For example, those SNs that are identified as influential and unreliable (i.e., r_i turn out to be relatively large) the FC decreases the current weights the most. However, for those SNs that are identified as compromised but not so influential to the FC decision process (i.e., r_i is relatively small) the FC decreases the weights proportional to r_i. With regard to SNs identified as honest, the FC keeps their weights unchanged. In this way, the FC decides through the weight combiner how much a local report should contribute to the FC final decision. This is a reasonable approach since if the report from a SN tends to be incorrect, it should be counted less in the final decision.

Next, in the simulation results, we will show that the reliability detection threshold (δ) and the weight penalty (μ) are crucial for the system detection performance. We will also show via simulations that there is an optimum δ and μ such that the system detection performance is maximized.

V. Simulations Results

Here we will evaluate numerically the performance of our proposed strategy and compare it to the $attack-free$ scheme [12] and the strategy in [26]. A WSN with a total of $M = 40$ SNs is considered (where a β fraction of these SNs are compromised by the attacker). For $\beta = 0.5$, $\beta = 0.25$, and $\beta = 0.1$, (SN21-SN40), (SN31-SN40), and (SN37-SN40) are respectively compromised. We let all the $\sigma_i^2 = 0.1$, such that $\xi_a = 10\log_{10}\left(\frac{1}{M}\sum_{i=1}^{M}\xi_i\right)$ = -10.5 dB with an arbitrarily chosen $\boldsymbol{s}(n) = [s_1(n), s_2(n), \cdots, s_M(n)]$ =[0.1, 0.175, 0.065, 0.027, 0.024, 0.026, 0.06, 0.09, 0.153, 0.11, 0.22, 0.12, 0.1, 0.024, 0.019, 0.05, 0.12, 0.1, 0.023, 0.021, 0.1, 0.175, 0.18, 0.027, 0.024, 0.026, 0.06, 0.09, 0.1, 0.065, 0.1, 0.175, 0.027, 0.024, 0.18, 0.026, 0.2, 0.09, 0.1, 0.18]^T, and where $\xi_i = \sum_{n=1}^{N} s_i^2(n)/N\sigma_i^2$. We will also refer to "equal weight" combining in (10) (i.e., $\alpha_i = 1, \forall i$) and use this as a benchmark. Finally, we use 10^5 Monte-Carlo simulations and choose a fixed (equal) local SNs threshold (Λ) in (5) and local SNs threshold (Λ_C) in (7) (i.e., more specifically, $\Lambda = \Lambda_C = 2.6$) such that $\bar{P}_d^{false} \leq 0.6$ (see Fig. 5-Fig. 7).

A. Impact of the time window length (K) on the malicious SN detection accuracy and on the system detection performance

In this section, we investigate the impact that the time window length (K) has on the compromised SNs identification accuracy of the proposed scheme. More precisely, we are interested in examining the two metrics, $P_d^{i,true}$ and $P_d^{i,false}$ (see (36)). Next, we examine the impact that this time window length (K) has on the system detection performance. More precisely, we will examine the two metrics P_d and P_{fa} (see (14)). Note that K affects these two metrics through the reliability metric r_i (see Fig. 2) in (34) which consequently affects the FC weight combining (37) that finally decides on the FC final test statistic (T_f) (see (10)).

In Fig. 2 we plot the reliability metric (r_i) against the FC detection threshold (Λ_f) for the compromised and the honest SNs. As expected, for the compromised but $influential$ SNs (i.e., SNs with the high local SNRs), the corresponding reliability metrics will be higher. In contrast, for the compromised

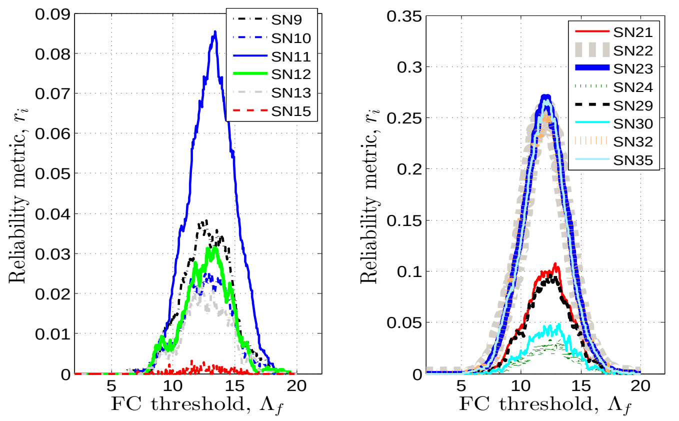


Fig. 2. The reliability metric (r_i) versus the FC detection threshold (Λ_f) parametrized on the SNs with $M = 40$, $N = 20$, $\beta = 0.5$, $P_C^{flip} = 1$ and $K = 150$.

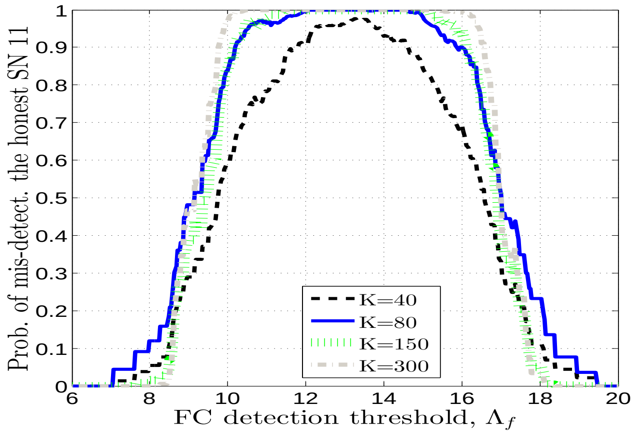


Fig. 4. Probability that the (honest) SN 11 has been falsely detected ($P_d^{11,false}$) versus the FC detection threshold (Λ_f), parametrized on K, with $M = 40$, $N = 20$, $\beta = 0.5$, $P_C^{flip} = 1$ and $\delta = 0.009$.

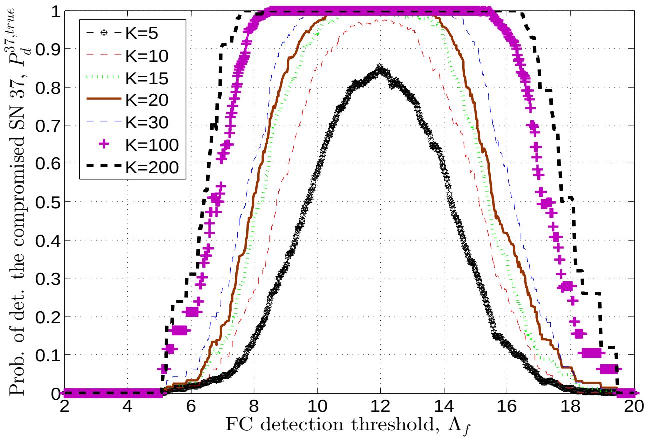


Fig. 3. Probability that the (compromised) SN 37 has been truly detected ($P_d^{37,true}$) versus the FC detection threshold (Λ_f), parametrized on K, with $M = 40$, $N = 20$, $\beta = 0.5$, $P_C^{flip} = 1$ and $\delta = 0.009$.

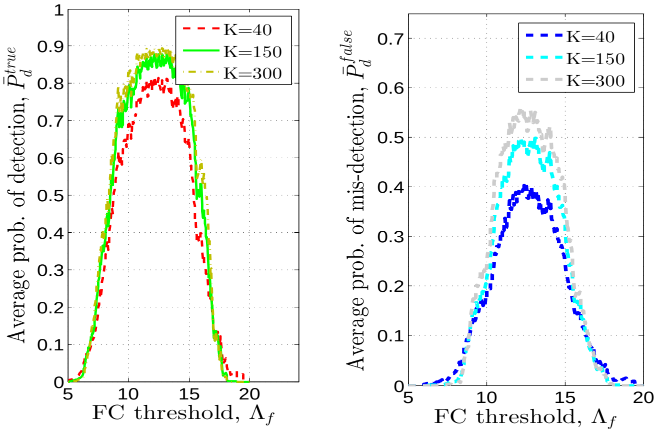


Fig. 5. Average probabilities: (left) of compromised SNs detection; (right) of honest SNs mis-detection versus the FC detection threshold (Λ_f), parametrized on K, with $M = 40$, $N = 20$, $\beta = 0.5$, $P_C^{flip} = 1$ and $\delta = 0.009$.

or honest SNs but $less\ influential$ (i.e., SNs with low SNRs), the corresponding reliability metrics with be lower.

In Fig. 3 we plot the probability of compromised SN's detection[4] (i.e., $truly$ detecting probability) ($P_d^{i,true}$) versus Λ_f, parametrized for different time window lengths (K). Clearly, as K increases, the detection accuracy (of the (compromised) SN 37) $P_d^{37,true}$ improves. In Fig. 4, we now plot the probability of honest SN's $mis-detection$[4] (i.e., $falsely$ detecting probability) ($P_d^{i,false}$) (see (36)) versus (like before) Λ_f for different time window lengths (K). Similarly (as in Fig. 3), we observe that the $mis-detection$ performance (of the (honest) SN 11) $P_d^{11,false}$ increases with K. Now, from Fig. 3 and Fig. 4 we conclude that increasing the time window length K not only improves the detection accuracy of the compromised SNs but at the same time increases (the undesired) $mis-detection$ probability of the honest SNs. This leads to a trade-off (while selecting the K parameter) between the compromised SNs detection accuracy and the honest SNs $mis-detection$ performance. Note that in practice we wish to keep $P_d^{i,true}$ high and $P_d^{i,false}$ low.

To give more generality to the results, in Fig. 5 we plot the average[5] performances (where the average is taken over the number of compromised/honest SNs). (left) We observe that while increasing K (more specifically from $K = 40$ to $K = 150$) we see an improvement in the average detection accuracy of compromised SNs. For larger K (e.g., $K = 300$) this improvement is negligible; (right) The same trend is observed for the average $mis-detection$ performance of the honest SNs.

In Fig. 6 we plot $\bar{P}_d^{i,true}$ and $\bar{P}_d^{false}$ versus the time window length (K) for a different FC detection thresholds (Λ_f). We can observe that the average compromised SNs detection

[4] SN 37 (Fig. 3) and SN 11 (Fig. 4) were chosen for comparison purposes as they possess the best and the worst performances among F and $(M-F)$ SNs for each case respectively. Here F and $(M-F)$ represents the number of compromised and honest SNs' respectively.

[5] The average performances are defined respectively as: $\bar{P}_d^{true} = \frac{1}{F}\sum_{i\in J} P_d^{i,true}$ and $\bar{P}_d^{false} = \frac{1}{M-F}\sum_{i\in \hat{J}} P_d^{i,false}$, where J ($\hat{J}$) represents the compromised (honest) SNs set with cardinality F ($[M-F]$) respectively.

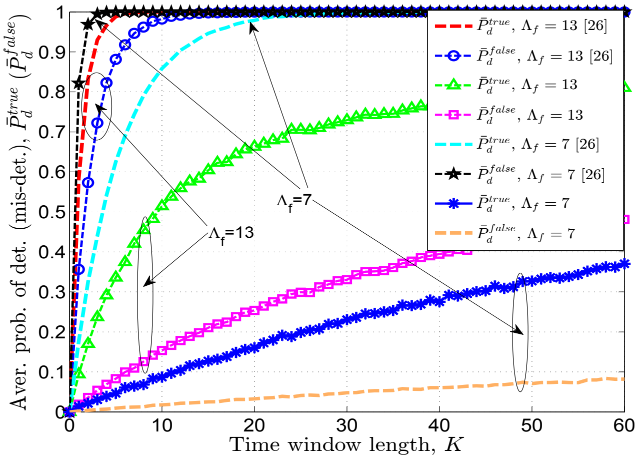


Fig. 6. Average compromised SNs detection probability against honest SNs mis-detection probability versus the time window length (K), parametrized on Λ_f, with $M = 40$, $N = 20$, $\beta = 0.5$, $P_C^{flip} = 1$ and $\delta = 0.009$.

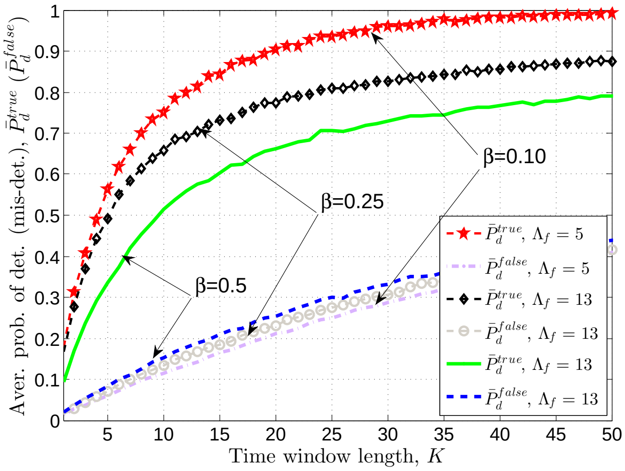


Fig. 7. Average compromised SNs detection probability and honest SNs mis-detection probability versus the time window length (K), parametrized on β with $M = 40$, $N = 20$, $P_C^{flip} = 1$ and $\delta = 0.009$.

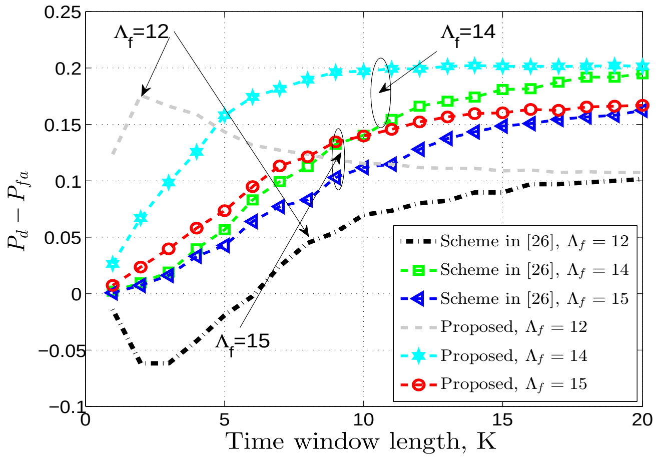


Fig. 8. The $P_d - P_{fa}$ metric versus the time window length (K), parametrized on the FC detection threshold (Λ_f), with $M = 40$, $N = 20$, $\beta = 0.25$, $P_C^{flip} = 1$, $\delta = 0.95$ and $\mu = 0.5$.

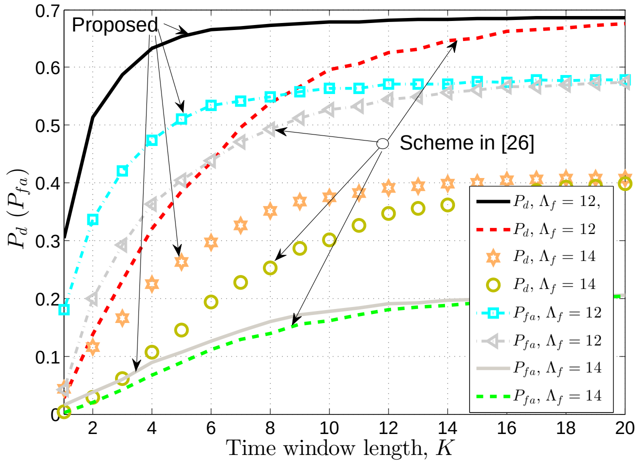


Fig. 9. Probability of detection (false alarm) P_d (P_{fa}) versus the time window length (K), parametrized on the FC detection threshold (Λ_f), with $M = 40$, $N = 20$, $\beta = 0.25$, $P_C^{flip} = 1$, $\delta = 0.95$ and $\mu = 0.5$.

performance ($\bar{P}_d^{i,true}$) improves with the time window length (K) for both schemes (i.e., the proposed one in this paper and the scheme proposed in [26]). Similar behavior can be observed for the (undesired) honest SNs $mis - detection$ probability. We also can observe that our proposed detection scheme outperforms the scheme proposed in [26] (or at least for the simulation setup considered in this paper), $\forall K$ in terms of $\bar{P}_d^{i,true} - \bar{P}_d^{false}$ quantity (e.g., for $\Lambda_f = 7$, $\bar{P}_d^{i,true} - \bar{P}_d^{false} \leq 0, \forall K$ for the scheme proposed in [26]). We note that in practice we would like to have $\bar{P}_d^{i,true}$ close to 1 and $\bar{P}_d^{false}$ close to 0 (i.e., $\bar{P}_d^{i,true} - \bar{P}_d^{false}$ close to 1).

In Fig. 7 we plot the same (i.e., $\bar{P}_d^{i,true}$ and $\bar{P}_d^{false}$ performances) but now parametrized on the fraction of compromised SNs (β). Clearly, the quantity $\bar{P}_d^{i,true} - \bar{P}_d^{false}$ improves when the fraction of compromised SNs (β) decreases. This behavior (as expected) results in a robust compromised SNs detection scheme.

Now, to give more validity to the results, in Fig. 8 we show the difference between the system detection and the system false alarm probability ($P_d - P_{fa}$) versus the time window length (K) parametrized on the FC detection threshold (Λ_f). Clearly, as K increases, the performance of the $P_d - P_{fa}$ metric improves for all the presented cases. Also, we can observe that our proposed scheme outperforms the one proposed in [26]. For example, targeting a rate of 0.16, the proposed scheme requires roughly a time window of length 5 while the scheme in [26] requires a time window of length 11. Then, to better understand how these two important metrics (i.e., P_d and P_{fa}) evolve with K, in Fig. 9 we show both the system detection probability (P_d) and the system false alarm probability (P_{fa}) versus the time window length (K) parametrized on the FC detection threshold (Λ_f). As expected, the larger is the time window length K, the better the detection performance. However, increasing K, results in an increase in the P_{fa} metric. Hence, while selecting K, one has to consider the allowable system false alarm probability.

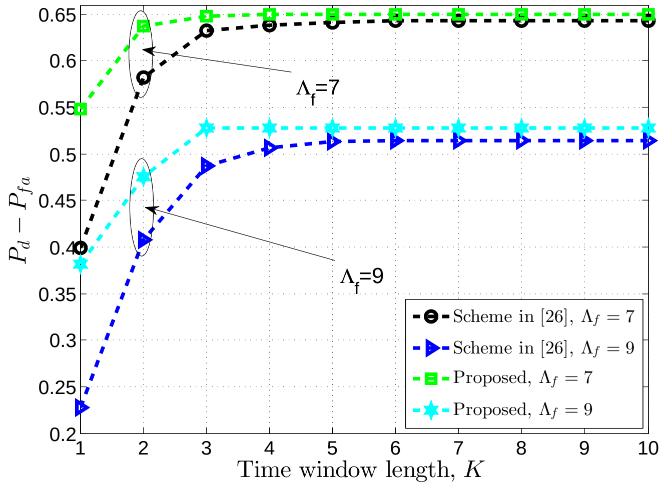


Fig. 10. The $P_d - P_{fa}$ metric versus the time window length (K), parametrized on the FC detection threshold (Λ_f), with $M = 40$, $N = 20$, $\beta = 0.25$, $P_C^{flip} = 0.2$, $\delta = 0.95$ and $\mu = 10$.

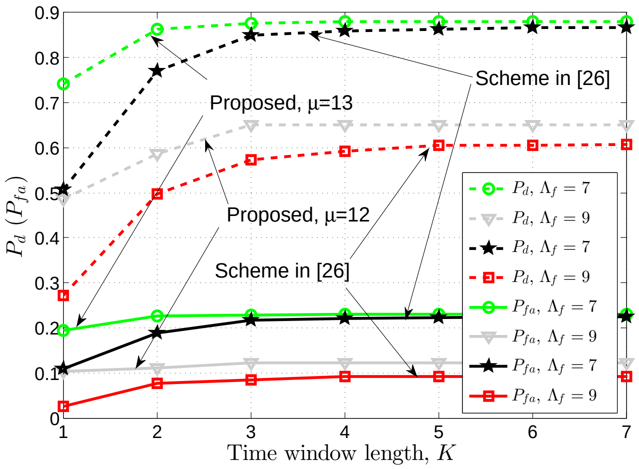


Fig. 11. Probability of detection (false alarm) P_d (P_{fa}) versus the time window length (K), parametrized on the FC detection threshold (Λ_f), with $M = 40$, $N = 20$, $\beta = 0.25$, $P_C^{flip} = 0.2$, and $\delta = 0.95$.

In Fig. 10 and in Fig. 11, we show the same (as in Fig. 8 and in Fig. 9 respectively) but now for (the attacker flipping probability) $P_C^{flip} = 0.2$ (see (8)). As expected, the $P_d - P_{fa}$ metric improves up to $K = 4$ whereas after that (i.e., for $K \geq 4$) a performance saturation gain is observed. We also note that the time window length (K^*) where this performance saturation gain is observed increases with the attacker flipping probability (P_C^{flip}) (see Fig. 8-Fig. 11). This is as expected, because increasing the (attacking) flipping probability one would require a larger time window length (K) for the FC in order to reduce as much as possible the attacker influence. However, increasing the value of K may introduce a delay to the FC detection algorithm. As a result, a careful choice for K should be selected in practice. Nevertheless, our proposed algorithm clearly requires a short time window span to converge.

B. Impact of reliability detection threshold and weight penalty parameter on the system detection performance

As previously mentioned, the reliability detection threshold and the weight penalty (i.e., δ and μ) (see (37)) are the two important parameters that will significantly affect the system detection performance at the FC.

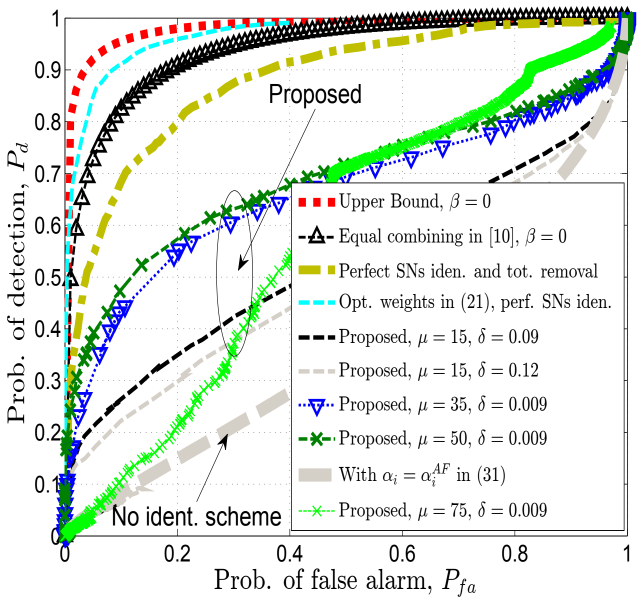


Fig. 12. Probability of detection (P_d) versus probability of false alarm (P_{fa}), parametrized on μ and δ, with $M = 40$, $N = 20$, $\beta = 0.5$ (unless otherwise stated), $P_C^{flip} = 1$ and $K = 5$.

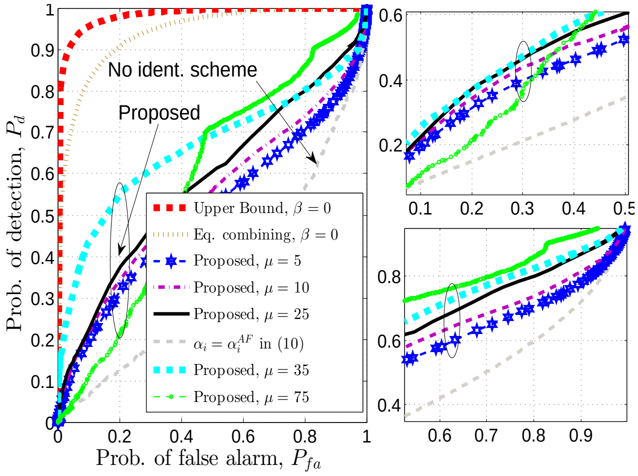


Fig. 13. Probability of detection (P_d) versus probability of false alarm (P_{fa}), parametrized on μ, with $M = 40$, $N = 20$, $\beta = 0.5$ (unless otherwise stated), $P_C^{flip} = 1$, $K = 5$, and $\delta = 0.009$.

So, in Fig. 12 we plot the ROC performance for different choices of the reliability detection threshold (δ) and for a fixed μ in (37). Obviously, there is an optimum value of δ such that P_d is maximized (for all the P_{fa} values). The detection performance using the weights derived under the $attack-free$ scenario (i.e., $\alpha_i = \alpha_i^{AF}$, see (31)) in (10) is also plotted. This corresponds to the case when no SNs identification scheme is used (i.e., $\mu_i = 0$ in (37)). Clearly, by appropriately choosing the reliability detection threshold (δ), the proposed identification scheme performance gain is significant compared to that when no identification scheme is used. Now, in Fig. 13 we show the same (but now for a fixed reliability detection threshold (δ)) and by varying the weight penalty parameter (μ). Clearly, there does exist an optimum value of μ that maximizes the ROC performance. Furthermore, the performance improvement parametrized on μ is shown to be significant for $P_{fa} \geq 0.1$.

C. Detection Performance Comparison

We now compare the system detection performance of the proposed strategy with the existing schemes.

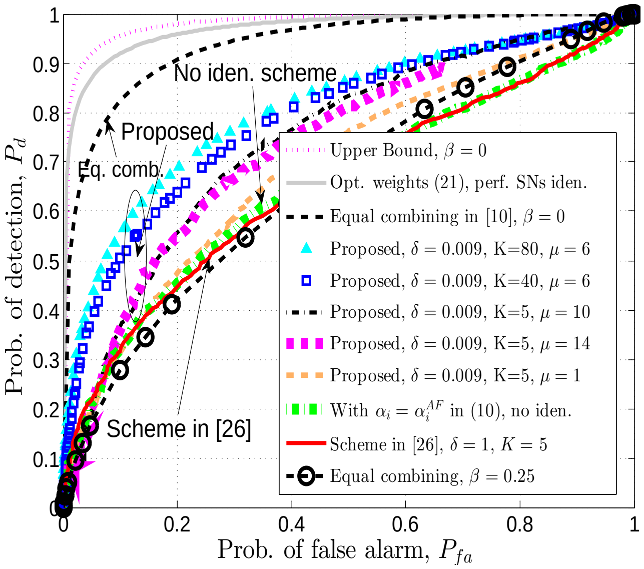


Fig. 14. Probability of detection (P_d) versus probability of false alarm (P_{fa}), parametrized on K, δ and μ, with $M = 40$, $N = 20$, $\beta = 0.25$ (unless otherwise stated), and $P_C^{flip} = 1$.

In Fig. 14, selecting some optimum value for δ and μ (more precisely, $\delta = 0.009$ and varying μ), we now compare our proposed strategy with the existing ones such as an equal combining scheme, the proposed scheme in [26] and the proposed scheme in [12] (i.e., with $\alpha_i = \alpha_i^{AF}$ in (10)) derived under the $attack - free$ scenario. We can observe that the performance of the proposed approach improves up to $\mu = 10$ whereas after that a performance degradation is noticed. Also, we can observe that by further increasing the time window length K, it is possible to further improve the detection performance. However, a careful selection of K should be made in practice as increasing the value of K introduces a delay to the FC decision making process. Clearly, the proposed scheme has a significant detection performance improvement compared to the case where no identification scheme is applied and also outperforms the existing strategy [12] and [26].

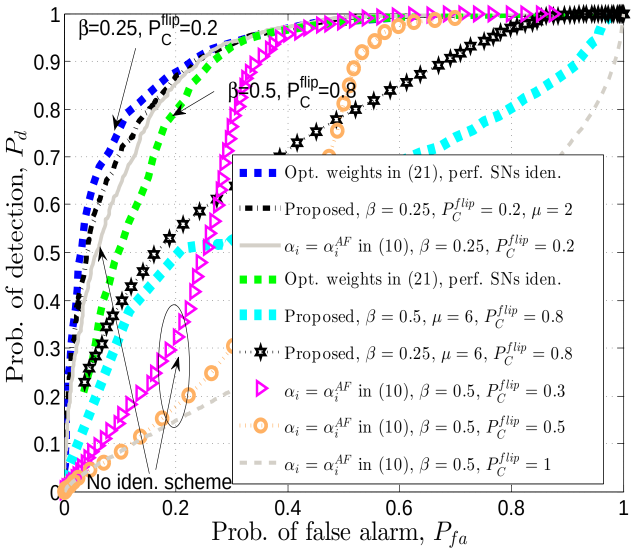


Fig. 15. Probability of detection (P_d) versus probability of false alarm (P_{fa}), parametrized on β, μ, and P_C^{flip}, with $M = 40$, $N = 20$, $K = 5$, and $\delta = 0.009$.

In Fig. 15, we report the ROC for the two different schemes (i.e., the one derived under an $attack - free$ scenario and the proposed one in this paper) parametrized on the fraction of compromised SNs (β) and flipping probability (P_C^{flip}) parameters. As expected (refer to (25)), the worst detection performance is observed for the case when $\beta = 0.5$ and $P_C^{flip} = 1$ as this is the case where the attacker causes the maximum possible FC degradation. Clearly, for a fixed β (i.e., $\beta = 0.5$), the detection performance improves as the flipping probability decreases. A significant improvement is observed in particular for high probability of false alarm (P_{fa}) values. Now, for low probability of false alarm (P_{fa}) (e.g., choosing $\beta = 0.25$ and $P_C^{flip} = 0.2$), the proposed scheme significantly outperforms the case when no identification scheme is applied (i.e., $\alpha_i = \alpha_i^{AF}$ in (10)) while for high P_{fa} its performance approaches the effective upper bound (i.e., when optimum weights in (21) are used and perfect SNs identification is assumed). Similarly, for e.g., $\beta = 0.25$ and $\beta = 0.5$ (for (fixed) $P_C^{flip} = 0.8$), the proposed approach possesses a remarkable detection performance gain compared to that of where no identification scheme is applied.

VI. Conclusion

In this paper, we have considered some of the key issues related to $under-attack$ WSNs. We have extended the results presented in our previous work [33] by considering a more realistic scenario where perfect knowledge of the true hypothesis is not required by the attacker. Optimal strategies from the FC's and the attacker's perspective have been characterized and some bounds have been derived.

We also proposed a new reliability metric and based on this, a reliability-based scheme was presented to identify the compromised SNs in the network and to control their contributions towards the FC's final decision. This new approach decreases the weights of the compromised SNs proportional to the reputation metric whereas the existing schemes totally exclude the compromised SNs (i.e., a zero weight is assigned) from the fusion process. Simulation results have shown that the proposed approach significantly outperforms, in terms of detection performance improvement, the existing FC rules and the compromised SNs identification schemes.

While this work and the other related publications assume that during the SNs identification stage, the attackers' parameters (i.e., β and P_C^{flip}) are fixed (i.e., not dynamic), there are interesting questions as to how the dynamic attackers' parameters will affect the network and how well the existing schemes can isolate the compromised SNs in the network. In this case, the dynamic optimum FC rules and the dynamic attacker strategies will be of particular interest and will be considered and investigated in future work in order to cope with such dynamic scenarios.

References

[1] D. Estrin, L. Girod, G. Pottie, and M. Srivastava, Instrumenting the world with wireless sensor networks, in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Salt Lake City, UT, United States, 7-11 May 2001.

[2] O. Songhwai, C. Phoebus, M. Michael, M. Srivastava, and S. Shankar, Instrumenting Wireless Sensor Networks for Real-time Surveillance, in Proc. of the International Conference on Robotics and Automation (ICRA), May 2006

[3] L. Zhang, G. Ding, Q. Wu, Y. Zou, Z. Han, and J. Wang, "Byzantine Attack and Defense in Cognitive Radio Networks: A Survey", *IEEE Communications Surveys & Tutorials*, vol. 17, no. 3, pp. 1342-1363, thirdquarter 2015.

[4] P. K. Varshney, *Distributed Detection and Data Fusion*. NewYork: Springer, 1997.

[5] J. N. Tsitsiklis, "Decentralized detection," *in Advances in Signal Processing*, H. V. Poor and J. B. Thomas, Eds. New York: JAI, 1993, vol. 2, pp. 297-344.

[6] R. Blum, S. Kassam, and H. Poor, "Distributed detection with multiple sensors: Part II-advanced topics," *Proc. IEEE*, vol. 85, no. 1, pp. 64-79, Jan. 1997.

[7] Z. Quan, S. Cui, and A. H. Sayed,"Optimal linear cooperation for spectrum sensing in cognitive radio networks," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 1, pp. 28-40, Feb. 2008.

[8] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "*Distributed Detection and Estimation in Wireless Sensor Networks*," In Rama Chellappa and Sergios Theodoridis eds., Academic Press Library in Signal Processing, Vol. 2, Communications and Radar Signal Processing, pp. 329-408, 2014.

[9] J. F. Chamberland and V. V. Veeravalli, "Asymptotic results for decentralized detection in power constrained wireless sensor networks," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 6, pp. 1007-1015, Aug. 2004.

[10] E. Nurellari, D. McLernon and M. Ghogho, "Distributed Two-Step Quantized Fusion Rules Via Consensus Algorithm for Distributed Detection in Wireless Sensor Networks," *in IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 3, pp. 321-335, Sept. 2016.

[11] A. Ribeiro and G. B. Giannakis, "Bandwidth-constrained distributed estimation for wireless sensor networks, part I: Gaussian case," *IEEE Transactions on Signal Processing*, vol. 54, no. 3, pp.1131-1143, 2006.

[12] E. Nurellari, D. McLernon, M. Ghogho and S. Aldalahmeh, "Optimal quantization and power allocation for energy-based distributed sensor detection," *Proc. EUSIPCO*, Lisbon, Portugal, 1-5 Sept. 2014.

[13] X. Zhang, H. V. Poor, and M. Chiang, "Optimal power allocation for distributed detection over MIMO channels in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 56, no. 9, pp. 41244140, Sep. 2008.

[14] E. Nurellari, S. Aldalahmeh, M. Ghogho and D. McLernon, "Quantized Fusion Rules for Energy-Based Distributed Detection in Wireless Sensor Networks ," *Proc. SSPD*, Edinburgh, Scotland, 8-9 Sept. 2014.

[15] T. Karygiannis and L. Owens, "Wireless network security," *NIST special publication*, 2002.

[16] S. Marano, V. Matta, and L. Tong, "Distributed detection in the presence of byzantine attacks," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 16-29, Oct 2009.

[17] A. Vempaty, L. Tong, and P. K. Varshney, "Distributed inference with Byzantine data," *IEEE Signal Process. Mag.,* vol. 30, no. 5, pp. 65-75, Sep. 2013.

[18] L. Lamport, R. Shostak, and M. Pease,"The byzantine generals problem," *ACM Trans. Program. Lang. Syst.,* vol. 4, no. 3, pp. 382-401, Jul. 1982. [Online]. Available: http://doi.acm.org/10.1145/357172.357176.

[19] B. Kailkhura, S. Brahma and P. K. Varshney, "Data Falsification Attacks on Consensus-Based Detection Systems," *in IEEE Transactions on Signal and Information Processing over Networks*, vol. 3, no. 1, pp. 145-158, Mar. 2017.

[20] V. S. S. Nadendla, Y. S. Han, and P. K. Varshney, " Distributed Inference With M-Ary Quantized Data in the Presence of Byzantine Attacks," *IEEE Transactions on Signal Processing*, vol. 62, no. 10, pp. 2681-2695, May 2014.

[21] L. Zhang, Q. Wu, G. Ding, S. Feng, and J. Wang, "Performance analysis of probabilistic soft SSDF attack in cooperative spectrum sensing," *EURASIP J. Adv. Signal Process.*, vol. 2014, no. 1, pp. 81, May 2014.

[22] S. Cui, Z. Han, S. Kar, T.T. Kim, H. Poor, and A. Tajer, "Coordinated data-injection attack and detection in smart grid," *IEEE Signal Process. Mag.* vol. 29, no. 5, pp. 106-115, Sep. 2012.

[23] P. Kaligineedi, M. Khabbazian, and V. K. Bhargava, "Secure cooperative sensing techniques for cognitive radio systems," *Proc. ICC*, pp. 3406- 3410, Beijing, China, 19-23 May 2008.

[24] A. Vempaty, L. Tong, and P. Varshney,"Distributed Inference with Byzantine Data: State-of-the-Art Review on Data Falsification Attacks," *Signal Processing Magazine, IEEE*, vol. 30, no. 5, pp. 65-75, 2013.

[25] R. Chen, J. Park, K. Bian, "Robust distributed spectrum sensing in cognitive radio networks," *Proc. INFOCOM*, pp. 1876-1884, Apr. 2008.

[26] A. S. Rawat, P. Anand, H. Chen, and P. K. Varshney, "Collaborative spectrum sensing in the presence of byzantine attacks in cognitive radio networks," *IEEE Trans. Signal Process.*, vol. 59, no. 2, pp. 774-786, Jan. 2011.

[27] H. Chen , X. Jin, and L. Xie, "Reputation-based Collaborative Spectrum Sensing Algorithm in Cognitive Radio Networks," *proc. PIRMC*, pp. 582-587, Tokyo, Japan, 13-16 Sep. 2009.

[28] R. Gentz, S. X. Wu, H. T. Wai, A. Scaglione, and A. Leshem, "Data Injection Attacks in Randomized Gossiping," *in IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 4, pp. 523-538, Dec. 2016.

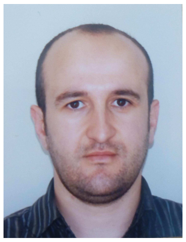
[29] H. Urkowitz, "Energy detection of unknown deterministic signals, *Proc. IEEE* , vol. 55, pp. 523-531, Apr. 1967.

[30] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory* , Englewood Cliffs, NJ: Prentice-Hall PTR, 1993.

[31] A. Vempaty, O. Ozdemir, K. Agrawal, H. Chen, and P. Varshney, "Localization in wireless sensor networks: Byzantines and mitigation techniques," *IEEE Trans. Signal Process.*, vol. 61, no. 6, pp. 1495-1508, Mar. 2013.

[32] T. Zhao and Y. Zhao, "A new cooperative detection technique with malicious user suppression," *Proc. ICC*, Dresden, Germany, 14-18 Jun. 2009.

[33] E. Nurellari, D. McLernon, M. Ghogho, and S. Aldalahmeh, "Distributed Binary Event Detection Under Data-Falsification and Energy-Bandwidth Limitation," *IEEE Sensors Journal*, vol. 16, no. 16, pp. 6298-6309, Aug. 15, 2016.



Edmond Nurellari received his B.Sc and his M.Sc degree in Electrical and Electronic Engineering, both from Eastern Mediterranean University, Northern Cyprus, in 2010 and in 2012 respectively.

From September 2010 to February 2013, he served as a Research and Teaching Assistant in the department of Electrical and Electronic Engineering at Eastern Mediterranean University. In 2013, he was awarded the Leeds International Research Scholarship (LIRS) to pursue his Ph. D. at the School of Electronics and Electrical Engineering, University of Leeds, United Kingdom. Since April 2017, he has been a faculty member with the School of Engineering at the University of Lincoln, United Kingdom, where he is currently a Lecturer in Electrical Engineering/Robotics.

His research interests includes distributed signal processing, signal processing on graphs, resource allocations, distributed decisions and network security analysis in wireless sensor networks by employing tools from graph theory and game theory. He has served as an Invited Reviewer for the IEEE Transactions on Signal and Information Processing over Networks, IEEE Communication Letter, Springers Wireless Networks Journal, Springers Digital Signal Processing Journal and IEEE Flagship conferences.

Des McLernon (M'94) received his B.Sc in Electronic and Electrical Engineering and his M.Sc. in Electronics, both from Queens University of Belfast, N. Ireland.

He then worked on radar research and development with Ferranti Ltd. in Edinburgh, Scotland, and later joined the Imperial College, University of London, where he received his Ph.D. in signal processing. After first lecturing at South Bank University, London, UK, he moved to the School of Electronic and Electrical Engineering at the University of Leeds, UK, where he is a Reader in Signal Processing.

His research interests are broadly within the domain of signal processing for communications, in which area he has published over 300 journal and conference papers.

Mounir Ghogho (SM'96) received the M.Sc degree (DEA) in 1993 and the Ph.D. degree in 1997 from the National Polytechnic Institute of Toulouse, France. He was an EPSRC Research Fellow with the University of Strathclyde, Glasgow (Scotland), from September 1997 to November 2001.

Since December 2001, he has been a faculty member with the school of Electronic and Electrical Engineering at the University of Leeds (England), where he is currently a Professor. He is also currently a Research Director and a Scientific Advisor to the President at the International University of Rabat (Morocco). He was awarded the UK Royal Academy of Engineering Research Fellowship in September 2000. He is a recipient of the 2013 IBM Faculty Award. He is currently an Associate Editor of the Signal Processing Magazine. He served as an Associate Editor of the IEEE Signal Processing Letters from 2001 to 2004, the IEEE Transactions on Signal Processing from 2005 to 2008, and the Elsevier Digital Signal Processing journal from 2011 to 2012. He served as a member of the IEEE Signal Processing Society SPCOM Technical Committee from 2005 to 2010, a member of IEEE Signal Processing Society SPTM Technical Committee from 2006 to 2011, and is currently a member of the IEEE Signal Processing Society SAM Technical Committee. He was the General Chair of the European Signal Processing conference Eusipco2013 and the IEEE workshop on Signal Processing for Advanced Wireless Communications SPAWC2010, the technical co-chair of the MIMO symposium of IWCMC 2007 and IWCMC 2008, and a technical area co-chair of Eusipco 2008, Eusipco 2009 and ISCCSP05. He is the general Chair of IEEE WCNC 2019.

His research interests are in signal processing and communication networks. He has published over 260 journal and conference papers. He held invited scientist/professor positions at Telecom Paris-Tech (France), NII (Japan), BUPT (China), University Carlos 3rd of Madrid (Spain), ENSICA (Toulouse), Darmstadt Technical University (Germany), and Minnesota University (USA). He is the Eurasip Liaison in Morocco.