Defending Surveillance Sensor Networks Against Data-Injection Attacks via Trusted Nodes

Roberto López-Valcarce
University of Vigo, Spain
valcarce@gts.uvigo.es

Daniel Romero
University of Agder, Norway
daniel.romero@uia.no

***Abstract*—By injecting false data through compromised sensors, an adversary can drive the probability of detection in a sensor network-based spatial field surveillance system to arbitrarily low values. As a countermeasure, a small subset of sensors may be secured. Leveraging the theory of Matched Subspace Detection, we propose and evaluate several detectors that add robustness to attacks when such trusted nodes are available. Our results reveal the performance-security tradeoff of these schemes and can be used to determine the number of trusted nodes required for a given performance target.**

***Index Terms*—Adversarial signal processing, Byzantine sensors, cyber security, sensor networks.**

I. Introduction

Sensor networks for surveillance and environmental monitoring [1], [2] comprise a large number of nodes measuring some physical phenomena and reporting to a fusion center (FC). Unattended nodes in typical deployments over large areas are susceptible to external attacks [3], [4], e.g, an adversary may capture some nodes and change the content of data packets, or directly alter the environment around some sensors. Standard cryptographic measures are ineffective against such data-injection attacks.

We focus on typical dense deployments, for which measurements of the monitored physical phenomenon exhibit spatial smoothness [5]. This allows for parsimonious parametric modeling of the corresponding spatial field, which can be exploited for inference purposes [6]–[8]. As in [9], a linear model in spatially independent Gaussian noise with unknown variance is adopted. The FC tests for the presence of a spatially smooth field (bacterial activity, toxic chemical spill, radioactivity, etc.). A related model with data-injection attacks for state *estimation* (rather than *detection*) in cyber-physical systems has been used in [10], [11]. An adversary may inject false data in a number of sensors to preclude the FC from detecting the spatial field. If the number of compromised sensors is strictly larger than the signal subspace dimension, the adversary will be able to maximize system degradation [9]. Since the effects can be disastrous, it is important to devise adequate defense mechanisms against such attacks.

Work funded by the Agencia Estatal de Investigación (Spain) and the European Regional Development Fund (ERDF) (projects TEC2013-47020-C2-1-R, TEC2015-69648-REDC, TEC2016-76409-C2-2-R), and by the Xunta de Galicia and ERDF (projects GRC2013/009, R2014/037 and ED431G/04 Agrupación Estratéxica Consolidada de Galicia accreditation 2016-2019).

To this end, we propose to endow the network with *trusted nodes*. By investing additional resources, network deployment may be planned so that a number of sensors are placed at secure locations, out of reach to adversaries. In this extended form of *tamper resistance* [12], not only physical protection against unauthorized attempts to read or modify the content of the device must be provided, but also against intentional alteration of the measured field near the node. For example, in clustered network topologies [13], [14], cluster heads are natural candidates for becoming trusted nodes. Alternatively, the FC may obtain estimates of node reliability via reputation metrics based on previous interactions[15], [16], so that only those nodes with a sufficiently high reputation metric are labeled as trusted. Our goal is to exploit the availability of trusted nodes to make spatial field detection robust to malicious actions. To this end, a number of schemes are developed based on the framework of Matched Subspace Detection [17]. These schemes will be shown to exhibit different tradeoffs between robustness and performance.

Notation: For a matrix $\boldsymbol{A}$, $\boldsymbol{A}^\dagger$ denotes its pseudoinverse, and $\mathcal{R}(\boldsymbol{A})$, $\mathcal{R}^\perp(\boldsymbol{A})$ its column and left null spaces, respectively. The perpendicular projection matrices onto $\mathcal{R}(\boldsymbol{A})$ and $\mathcal{R}^\perp(\boldsymbol{A})$ are respectively denoted by $\boldsymbol{P_A} = \boldsymbol{A}\boldsymbol{A}^\dagger$ and $\boldsymbol{P_A^\perp} = \boldsymbol{I} - \boldsymbol{P_A}$. The Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{C}$ is denoted by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C})$. By $\mathcal{F}_{\nu_1,\nu_2}$, $\mathcal{F}'_{\nu_1,\nu_2}(\lambda_1)$ and $\mathcal{F}''_{\nu_1,\nu_2}(\lambda_1,\lambda_2)$ we respectively denote the central, noncentral, and doubly noncentral F-distributions with ν_1 and ν_2 degrees of freedom (d.o.f.) and noncentrality parameters λ_1, λ_2 [18].

II. Problem Statement and Preliminaries

Consider a surveillance network deployed to monitor the presence or absence of some spatial field, consisting of an FC collecting data from n sensors. These data are comprised in vector $\boldsymbol{y} \triangleq [\, y_1 \, \cdots \, y_n \,]^T$, modeled as

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{a} + \boldsymbol{w} \quad \in \quad \mathbb{R}^n, \tag{1}$$

where $\boldsymbol{H} \in \mathbb{R}^{n\times m}$ is a known full-rank matrix, $\boldsymbol{x} \in \mathbb{R}^m$ is related to the monitored physical process, and $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ is measurement noise with unknown variance σ^2. The *attack vector* $\boldsymbol{a} = [\, a_1 \, \cdots \, a_n \,]^T \in \mathbb{R}^n$ is injected by the adversary. The goal of the network is to detect the presence of the spatial field, i.e., to decide whether or not $\boldsymbol{x} = \boldsymbol{0}$. The basis expansion model $\boldsymbol{H}\boldsymbol{x}$ of the signal component is fairly general; it hinges

on the assumption that the spatial field is sufficiently smooth so it can be parameterized by a low-dimensional $\boldsymbol{x}$ ($m \ll n$). It fits a wide range of signal representations based on Fourier or Discrete Cosine Transforms, polynomial bases, splines, etc. [6], [8]. Thus, $\boldsymbol{H}$ depends on the chosen basis expansion and sensor locations.

After instigating some event (e.g., chemical spill) which will eventually generate a spatial field measurable by the network ($\boldsymbol{x} \neq \boldsymbol{0}$), the adversary's goal is to prevent this event from being detected, by injecting false data into a subset of $k \ll n$ compromised sensors. Thus, a_i is freely selected by the adversary if the i-th node was compromised; else, $a_i = 0$.

Our goal is to investigate defense mechanisms to make the network robust to these data-injection attacks. To this end, first we briefly review Matched Subspace Detection [17], since the problems studied in the sequel fit in this class. Subsequently, we will review the effect of false data injection in an attack-unaware network [9], which will motivate the search for alternatives robust to this kind of attacks.

A. Matched Subspace Detectors

Consider the general model

$$\begin{array}{rcll} \boldsymbol{y} & = & \boldsymbol{A\theta} + \boldsymbol{v}, & \boldsymbol{A} \in \mathbb{R}^{n\times p}, \quad \boldsymbol{\theta} \in \mathbb{R}^p, \\ \boldsymbol{v} & \sim & \mathcal{N}(\boldsymbol{B\phi}, \sigma^2 \boldsymbol{I}_n), & \boldsymbol{B} \in \mathbb{R}^{n\times q}, \quad \boldsymbol{\phi} \in \mathbb{R}^q, \end{array} \tag{2}$$

with $\boldsymbol{A}$, $\boldsymbol{B}$ known, and $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, σ^2 unknown. The matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $[\boldsymbol{A}\ \boldsymbol{B}]$ are assumed full-rank. We wish to test for the presence of the signal component $\boldsymbol{A\theta}$: $\mathcal{H}_0 : \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{B\phi}, \sigma^2\boldsymbol{I}_n)$ vs. $\mathcal{H}_1 : \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{A\theta} + \boldsymbol{B\phi}, \sigma^2\boldsymbol{I}_n)$. The component $\boldsymbol{B\phi}$ can be regarded as an interference term. Thus, in this model, the signal term is known to lie in the subspace $\mathcal{R}(\boldsymbol{A})$, whereas the interference is known to lie in the subspace $\mathcal{R}(\boldsymbol{B})$.

Let $\boldsymbol{G} = \boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{A}$. Provided that $p + q < n$, the Generalized Likelihood Ratio Test (GLRT) exists and is given by

$$T \triangleq \frac{n-p-q}{p} \frac{\|\boldsymbol{P}_{\boldsymbol{G}}\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{y}\|^2}{\|\boldsymbol{P}_{\boldsymbol{G}}^{\perp}\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{y}\|^2} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma, \tag{3}$$

and it is uniformly most powerful invariant (UMPI) [17]. In (3), the interference is first cancelled by $\boldsymbol{P}_{\boldsymbol{B}}^{\perp}$, after which the signal component is known to lie in $\mathcal{R}(\boldsymbol{G})$. Then (3) compares the energies of $\boldsymbol{P}_{\boldsymbol{B}}^{\perp}\boldsymbol{y}$ in this subspace and in its orthogonal complement $\mathcal{R}^{\perp}(\boldsymbol{G})$. The test statistic T is F-distributed as $T \,|\, \mathcal{H}_0 \sim \mathcal{F}_{p,n-p-q}$ and $T \,|\, \mathcal{H}_1 \sim \mathcal{F}'_{p,n-p-q}\left(\frac{\|\boldsymbol{G\theta}\|^2}{\sigma^2}\right)$. For fixed $\frac{\|\boldsymbol{G\theta}\|^2}{\sigma^2}$, detection performance improves as the first number of d.o.f. (signal subspace dimension p) decreases, and as the second number of d.o.f. $n-p-q$ increases.

B. Field detection with an unprotected network

Consider now the original model (1), and suppose that the network is oblivious to potential attacks, meaning that it assumes $\boldsymbol{a} = \boldsymbol{0}$. Under this assumption, the problem of testing for field presence, i.e., $\mathcal{H}_0 : \boldsymbol{x} = \boldsymbol{0}$ vs. $\mathcal{H}_1 : \boldsymbol{x} \neq \boldsymbol{0}$, can be cast in the framework of (2) with $\boldsymbol{A\theta} = \boldsymbol{Hx}$ and $\boldsymbol{B\phi} = \boldsymbol{0}$ (and thus $p = m$ and $q = 0$). The corresponding GLRT is

$$T_{\mathrm{F}} \triangleq \frac{n-m}{m} \frac{\|\boldsymbol{P}_{\boldsymbol{H}}\boldsymbol{y}\|^2}{\|\boldsymbol{P}_{\boldsymbol{H}}^{\perp}\boldsymbol{y}\|^2} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma_{\mathrm{F}}, \tag{4}$$

and in the absence of attacks, one has $T_{\mathrm{F}} \,|\, \mathcal{H}_0 \sim \mathcal{F}_{m,n-m}$ and $T_{\mathrm{F}} \,|\, \mathcal{H}_1 \sim \mathcal{F}'_{m,n-m}(\rho)$, where $\rho \triangleq \frac{\|\boldsymbol{Hx}\|^2}{\sigma^2}$ is the Signal-to-Noise Ratio (SNR). However, as shown in [9], the presence of an attack changes the distribution of T_{F} under $\mathcal{H}_1$ to

$$T_{\mathrm{F}} \sim \mathcal{F}''_{m,n-m}(\rho_{||}, \rho_{\perp}) \ \text{ with } \begin{cases} \rho_{||} & \triangleq & \frac{\|\boldsymbol{Hx}+\boldsymbol{P}_{\boldsymbol{H}}\boldsymbol{a}\|^2}{\sigma^2}, \\ \rho_{\perp} & \triangleq & \frac{\|\boldsymbol{P}_{\boldsymbol{H}}^{\perp}\boldsymbol{a}\|^2}{\sigma^2}. \end{cases} \tag{5}$$

Under (5), the detection probability of (4) is monotonically increasing in $\rho_{||}$ and decreasing in $\rho_{\perp}$. As in [9], we assume that the adversary knows $\boldsymbol{H}$ (which could be inferred from sensor locations), but not $\boldsymbol{x}$ (which depends on unknown environmental variables affecting field diffusion). Thus, to avoid accidentally increasing $\rho_{||}$, the adversary may choose an *orthogonal attack* $\boldsymbol{a} \in \mathcal{R}^{\perp}(\boldsymbol{H})$; for this, it suffices to capture $k > m$ sensors [9]. The performance of (4) can be drastically degraded by this kind of attack, in the sense that its detection probability can be made arbitrarily small by increasing the attack power $\|\boldsymbol{a}\|^2$ [9]. Clearly, appropriate countermeasures are needed.

Note that directly applying the results in Sec. II-A to (1) (by regarding $\boldsymbol{a}$ as the interference term) is not possible: the subset of compromised sensors is unknown to the FC, so it does not know any subspace in which $\boldsymbol{a}$ must lie. At most, by the reasoning in Sec. II-B, one could assume that $\boldsymbol{a} \in \mathcal{R}^{\perp}(\boldsymbol{H})$, but then, in the framework of Sec. II-A, $p = m$ and $q = n-m$: this violates the constraint $p + q < n$, so the GLRT does not exist. As shown in the sequel, this situation changes once trusted nodes become available.

III. Field detection with trusted nodes

One sensible way to provide robustness to attacks is to secure a subset of nodes. Thus, suppose that a number t of them, say nodes $1, \ldots, t$, are trusted. Then the attack vector must be of the form $\boldsymbol{a} = [\ \boldsymbol{0}_t^T\ \tilde{\boldsymbol{a}}^T\ ]^T$ with $\tilde{\boldsymbol{a}} \in \mathbb{R}^{n-t}$. Letting $\bar{\boldsymbol{H}} \in \mathbb{R}^{t\times m}$ and $\tilde{\boldsymbol{H}} \in \mathbb{R}^{(n-t)\times m}$ comprise the first t and last $n-t$ rows of $\boldsymbol{H}$ respectively, model (1) can be split as

$$\bar{\boldsymbol{y}} = \bar{\boldsymbol{H}}\boldsymbol{x} + \bar{\boldsymbol{w}}, \qquad \tilde{\boldsymbol{y}} = \tilde{\boldsymbol{H}}\boldsymbol{x} + \tilde{\boldsymbol{a}} + \tilde{\boldsymbol{w}}, \tag{6}$$

with $\bar{\boldsymbol{y}} \triangleq [\ y_1\ \cdots\ y_t\ ]^T$, $\tilde{\boldsymbol{y}} \triangleq [\ y_{t+1}\ \cdots\ y_n\ ]^T$ (data from trusted and untrusted nodes respectively), and $\bar{\boldsymbol{w}}$, $\tilde{\boldsymbol{w}}$ the corresponding noise vectors. The availability of trusted nodes constrains the attack vector to a certain subspace, allowing the application of the theory from Sec. II-A. Now, two strategies are possible: the first one makes no assumptions on the attack vector $\tilde{\boldsymbol{a}}$, whereas the second assumes a structured (orthogonal) attack.

A. Unstructured attack

With no assumptions on $\tilde{\boldsymbol{a}}$, then in the framework of Sec. II-A, one may take

$$\boldsymbol{A} = \boldsymbol{H}, \quad \boldsymbol{B} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{I}_{n-t} \end{bmatrix} \Rightarrow \begin{cases} p & = & m, \\ q & = & n-t. \end{cases} \tag{7}$$

Hence, to have $p + q < n$ one needs $t > m$, i.e., the number of trusted nodes must exceed the dimension of the signal

subspace. Then, the GLRT statistic for field detection is given by

$$T_{\rm FU} \triangleq \frac{t-m}{m}\frac{\|\boldsymbol{P}_{\bar{\boldsymbol{H}}}\bar{\boldsymbol{y}}\|^2}{\|\boldsymbol{P}^{\perp}_{\bar{\boldsymbol{H}}}\bar{\boldsymbol{y}}\|^2} \sim \mathcal{F}'_{m,t-m}\left(\frac{\|\bar{\boldsymbol{H}}\boldsymbol{x}\|^2}{\sigma^2}\right). \quad (8)$$

Attacks have no effect on the performance of (8), but all data from untrusted nodes is discarded, which is clearly inefficient. Alternatives exhibiting a better robustness-performance trade-off will be investigated next.

B. Structured attack

If an orthogonal attack $\boldsymbol{a} \in \mathcal{R}^{\perp}(\boldsymbol{H})$ is assumed, then the resulting interference subspace can be further constrained. This subspace can be characterized as follows.

Lemma 1. *If* $\boldsymbol{a} = [\,\boldsymbol{0}_t^T\ \tilde{\boldsymbol{a}}^T\,]^T$ *and* $\boldsymbol{a} \in \mathcal{R}^{\perp}(\boldsymbol{H})$*, then* $\tilde{\boldsymbol{a}} \in \mathcal{R}^{\perp}(\tilde{\boldsymbol{H}})$.

Proof. $\boldsymbol{a}\in\mathcal{R}^{\perp}(\boldsymbol{H})$ means $\boldsymbol{H}^T\boldsymbol{a}=\boldsymbol{0}$. Since $\boldsymbol{a}=[\,\boldsymbol{0}_t^T\ \tilde{\boldsymbol{a}}^T\,]^T$, one has $\boldsymbol{H}^T\boldsymbol{a}=\boldsymbol{0}\Rightarrow\tilde{\boldsymbol{H}}^T\tilde{\boldsymbol{a}}=\boldsymbol{0}$, hence $\tilde{\boldsymbol{a}}\in\mathcal{R}^{\perp}(\tilde{\boldsymbol{H}})$. ■

Let $\tilde{r}=\operatorname{rank}\tilde{\boldsymbol{H}}$, and let the columns of $\tilde{\boldsymbol{U}}_{\perp}\in\mathbb{R}^{(n-t)\times(n-t-\tilde{r})}$ be an orthonormal basis for $\mathcal{R}^{\perp}(\tilde{\boldsymbol{H}})$, so that $\boldsymbol{P}^{\perp}_{\tilde{\boldsymbol{H}}}=\tilde{\boldsymbol{U}}_{\perp}\tilde{\boldsymbol{U}}_{\perp}^T$. Assume that $\boldsymbol{a}\in\mathcal{R}^{\perp}(\boldsymbol{H})$. Then, in view of Lemma 1, field detection can be cast in the framework of Sec. II-A by setting

$$\boldsymbol{A}=\boldsymbol{H},\quad \boldsymbol{B}=\begin{bmatrix}\boldsymbol{0}\\ \tilde{\boldsymbol{U}}_{\perp}\end{bmatrix} \Rightarrow \begin{cases} p &=& m,\\ q &=& n-t-\tilde{r}.\end{cases} \quad (9)$$

The GLRT exists if $p+q<n$, i.e., $t>m-\tilde{r}$ (in particular, if the matrix $\tilde{\boldsymbol{H}}$ has full column rank $\tilde{r}=m$, then the GLRT exists for any number of trusted nodes $t>0$), in which case

$$\left.\begin{array}{rcl}\boldsymbol{P}^{\perp}_{\boldsymbol{B}} &=& \boldsymbol{I}-\boldsymbol{P}\\ \boldsymbol{G} &=& (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{H}=\boldsymbol{H}\end{array}\right\} \text{ with } \boldsymbol{P}=\begin{bmatrix}\boldsymbol{0} & \\ & \boldsymbol{P}^{\perp}_{\tilde{\boldsymbol{H}}}\end{bmatrix}. \quad (10)$$

Since $\boldsymbol{P}_{\boldsymbol{H}}\boldsymbol{P}=\boldsymbol{0}$, it follows that $\boldsymbol{P}_{\boldsymbol{G}}\boldsymbol{P}^{\perp}_{\boldsymbol{B}}=\boldsymbol{P}_{\boldsymbol{H}}$ and $\boldsymbol{P}^{\perp}_{\boldsymbol{G}}\boldsymbol{P}^{\perp}_{\boldsymbol{B}}=\boldsymbol{P}^{\perp}_{\boldsymbol{H}}-\boldsymbol{P}$. This results in the GLRT statistic

$$T_{\rm FS} \triangleq \frac{t+\tilde{r}-m}{m}\frac{\|\boldsymbol{P}_{\boldsymbol{H}}\boldsymbol{y}\|^2}{\|\boldsymbol{P}^{\perp}_{\boldsymbol{H}}\boldsymbol{y}\|^2-\|\boldsymbol{P}^{\perp}_{\tilde{\boldsymbol{H}}}\tilde{\boldsymbol{y}}\|^2} \quad (11)$$

$$\sim \mathcal{F}'_{m,t+\tilde{r}-m}\left(\frac{\|\boldsymbol{H}\boldsymbol{x}\|^2}{\sigma^2}\right), \quad (12)$$

which coincides with the distribution of (4) under no attacks, $T_{\rm F}\sim\mathcal{F}'_{m,n-m}\left(\frac{\|\boldsymbol{H}\boldsymbol{x}\|^2}{\sigma^2}\right)$, if n is replaced by $t+\tilde{r}$. The operational SNR of both tests is the same, in contrast with (8).

To analyze the effect of an attack not complying with the original assumption $\boldsymbol{a}\in\mathcal{R}^{\perp}(\boldsymbol{H})$, note from Lemma 1 that this implies $\boldsymbol{a}=[\,\boldsymbol{0}_t^T\ \tilde{\boldsymbol{a}}^T\,]^T$ with $\tilde{\boldsymbol{a}}\notin\mathcal{R}^{\perp}(\tilde{\boldsymbol{H}})$. Hence, let

$$\boldsymbol{a}=\underbrace{\begin{bmatrix}\boldsymbol{0}_t\\ \tilde{\boldsymbol{a}}_{||}\end{bmatrix}}_{=\boldsymbol{a}_{||}}+\underbrace{\begin{bmatrix}\boldsymbol{0}_t\\ \tilde{\boldsymbol{a}}_{\perp}\end{bmatrix}}_{=\boldsymbol{a}_{\perp}} \text{ with } \begin{cases}\tilde{\boldsymbol{a}}_{||}\in\mathcal{R}(\tilde{\boldsymbol{H}}),\\ \tilde{\boldsymbol{a}}_{\perp}\in\mathcal{R}^{\perp}(\tilde{\boldsymbol{H}}).\end{cases} \quad (13)$$

(Note that, in general, $\boldsymbol{a}_{||}$ does not lie either in $\mathcal{R}(\boldsymbol{H})$ nor in $\mathcal{R}^{\perp}(\boldsymbol{H})$). Then the distribution of $T_{\rm FS}$ changes from (12) to

$$T_{\rm FS}\sim\mathcal{F}''_{m,t+\tilde{r}-m}\left(\frac{\|\boldsymbol{H}\boldsymbol{x}+\boldsymbol{P}_{\boldsymbol{H}}\boldsymbol{a}_{||}\|^2}{\sigma^2},\frac{\|\boldsymbol{P}^{\perp}_{\boldsymbol{H}}\boldsymbol{a}_{||}\|^2}{\sigma^2}\right). \quad (14)$$

The detection probability is monotonically increasing (resp. decreasing) in the first (resp. second) noncentrality parameter of (14). Again, since $\boldsymbol{x}$ is unknown to the adversary, he is unable to design the attack to reduce $\|\boldsymbol{H}\boldsymbol{x}+\boldsymbol{P}_{\boldsymbol{H}}\boldsymbol{a}_{||}\|^2$; with a "poor" choice for $\boldsymbol{a}_{||}$, this term may actually increase. Thus, a sensible approach for the attacker would be to maximize $\|\boldsymbol{P}^{\perp}_{\boldsymbol{H}}\boldsymbol{a}_{||}\|^2$ subject to $\boldsymbol{P}_{\boldsymbol{H}}\boldsymbol{a}_{||}=\boldsymbol{0}$. However, one has:

Lemma 2. *If* $\boldsymbol{a}_{||}=[\,\boldsymbol{0}_t^T\,\tilde{\boldsymbol{a}}_{||}^T\,]^T$ *with* $\tilde{\boldsymbol{a}}_{||}\in\mathcal{R}(\tilde{\boldsymbol{H}})$*, then* $\boldsymbol{P}_{\boldsymbol{H}}\boldsymbol{a}_{||}=\boldsymbol{0}$ *implies* $\boldsymbol{a}_{||}=\boldsymbol{0}$.

Proof. $\boldsymbol{P}_{\boldsymbol{H}}\boldsymbol{a}_{||}=\boldsymbol{0}$ means $\boldsymbol{a}_{||}\in\mathcal{R}^{\perp}(\boldsymbol{H})$, i.e., $\boldsymbol{H}^T\boldsymbol{a}_{||}=\tilde{\boldsymbol{H}}^T\tilde{\boldsymbol{a}}_{||}=\boldsymbol{0}$. Thus, $\tilde{\boldsymbol{a}}_{||}\in\mathcal{R}^{\perp}(\tilde{\boldsymbol{H}})$, but since $\tilde{\boldsymbol{a}}_{||}\in\mathcal{R}(\tilde{\boldsymbol{H}})$ by assumption, it follows that $\tilde{\boldsymbol{a}}_{||}=\boldsymbol{0}$, hence $\boldsymbol{a}_{||}=\boldsymbol{0}$. ■

Thus, any attack with $\|\boldsymbol{P}^{\perp}_{\boldsymbol{H}}\boldsymbol{a}_{||}\|^2>0$ must yield $\boldsymbol{P}_{\boldsymbol{H}}\boldsymbol{a}_{||}\neq\boldsymbol{0}$, so the effect on detection probability cannot be determined a priori: it may very well *increase* due to the attack. This establishes the robustness of (11) to data-injection attacks.

IV. Joint Field-Attack detection

The schemes in Sec. III attempt to block $\boldsymbol{a}$ before testing for field presence. Noting that the adversary is only interested in launching an attack if a field is present suggests an alternative approach, by which the FC jointly tests for the presence of the spatial field *or* the attack vector, i.e., $\mathcal{H}_0:\boldsymbol{H}\boldsymbol{x}+\boldsymbol{a}=\boldsymbol{0}$ vs. $\mathcal{H}_1:\boldsymbol{H}\boldsymbol{x}+\boldsymbol{a}\neq\boldsymbol{0}$. Again, nodes $1,\ldots,t$ are assumed trusted, so that $\boldsymbol{a}=[\,\boldsymbol{0}_t^T\ \tilde{\boldsymbol{a}}^T\,]^T$.

A. Unstructured attack

With no assumptions on $\tilde{\boldsymbol{a}}$, then in the framework of Sec. II-A, $\boldsymbol{B}$ is empty (thus, $\boldsymbol{P}^{\perp}_{\boldsymbol{B}}=\boldsymbol{I}$) and $\boldsymbol{A}\boldsymbol{\theta}=\boldsymbol{H}\boldsymbol{x}+\boldsymbol{a}$. Hence,

$$\boldsymbol{A}=\begin{bmatrix}\bar{\boldsymbol{H}} & \boldsymbol{0}\\ \tilde{\boldsymbol{H}} & \boldsymbol{I}_{n-t}\end{bmatrix} \Rightarrow \begin{cases} p &=& m+n-t,\\ q &=& 0.\end{cases} \quad (15)$$

The GLRT exists if $t>m$, and its test statistic is given by

$$T_{\rm JU} \triangleq \frac{t-m}{m+n-t}\frac{\|\boldsymbol{P}_{\bar{\boldsymbol{H}}}\bar{\boldsymbol{y}}\|^2+\|\tilde{\boldsymbol{y}}\|^2}{\|\boldsymbol{P}^{\perp}_{\bar{\boldsymbol{H}}}\bar{\boldsymbol{y}}\|^2} \quad (16)$$

$$\sim \mathcal{F}'_{m+n-t,t-m}\left(\frac{\|\boldsymbol{H}\boldsymbol{x}+\boldsymbol{a}\|^2}{\sigma^2}\right). \quad (17)$$

B. Structured attack

Assuming an orthogonal attack $\boldsymbol{a}\in\mathcal{R}^{\perp}(\boldsymbol{H})$ (and therefore $\tilde{\boldsymbol{a}}\in\mathcal{R}^{\perp}(\tilde{\boldsymbol{H}})$, in view of Lemma 1), we again set $\boldsymbol{B}$ to be empty, whereas now

$$\boldsymbol{A}=\begin{bmatrix}\bar{\boldsymbol{H}} & \boldsymbol{0}\\ \tilde{\boldsymbol{H}} & \tilde{\boldsymbol{U}}_{\perp}\end{bmatrix} \Rightarrow \begin{cases} p &=& m+n-t-\tilde{r},\\ q &=& 0.\end{cases} \quad (18)$$

Then $\boldsymbol{G}=\boldsymbol{A}$ and, with $\boldsymbol{P}$ as in (10), one has $\boldsymbol{P}_{\boldsymbol{G}}=\boldsymbol{P}_{\boldsymbol{H}}+\boldsymbol{P}$ and $\boldsymbol{P}^{\perp}_{\boldsymbol{G}}=\boldsymbol{P}^{\perp}_{\boldsymbol{H}}-\boldsymbol{P}$. The GLRT exists if $t>m-\tilde{r}$ and, since $\boldsymbol{P}_{\boldsymbol{H}}\boldsymbol{P}=\boldsymbol{0}$, its statistic can be written as

$$T_{\rm JS} \triangleq \frac{t+\tilde{r}-m}{m+n-t-\tilde{r}}\frac{\|\boldsymbol{P}_{\boldsymbol{H}}\boldsymbol{y}\|^2+\|\boldsymbol{P}^{\perp}_{\tilde{\boldsymbol{H}}}\tilde{\boldsymbol{y}}\|^2}{\|\boldsymbol{P}^{\perp}_{\boldsymbol{H}}\boldsymbol{y}\|^2-\|\boldsymbol{P}^{\perp}_{\tilde{\boldsymbol{H}}}\tilde{\boldsymbol{y}}\|^2} \quad (19)$$

$$\sim \mathcal{F}'_{m+n-t-\tilde{r},t+\tilde{r}-m}\left(\frac{\|\boldsymbol{H}\boldsymbol{x}\|^2}{\sigma^2}+\frac{\|\boldsymbol{a}\|^2}{\sigma^2}\right) \quad (20)$$

with (20) valid if $\boldsymbol{a} \in \mathcal{R}^{\perp}(\boldsymbol{H})$. Otherwise, with $\boldsymbol{a}$ given by (13), the distribution of $T_{\rm JS}$ changes to

$$\mathcal{F}''_{m+n-t-\tilde{r},t+\tilde{r}-m}\left(\frac{\|\boldsymbol{H}\boldsymbol{x}+\boldsymbol{P}_{\boldsymbol{H}}\boldsymbol{a}_{||}\|^2}{\sigma^2}+\frac{\|\boldsymbol{a}_{\perp}\|^2}{\sigma^2},\frac{\|\boldsymbol{P}^{\perp}_{\boldsymbol{H}}\boldsymbol{a}_{||}\|^2}{\sigma^2}\right) \tag{21}$$

Any $\boldsymbol{a}_{\perp} \neq \boldsymbol{0}$ will contribute to increasing the probability of detection. Regarding the component $\boldsymbol{a}_{||}$, in view of Lemma 2, the adversary cannot increase the term $\|\boldsymbol{P}^{\perp}_{\boldsymbol{H}}\boldsymbol{a}_{||}\|^2$ in (21) without the risk of increasing the term $\|\boldsymbol{H}\boldsymbol{x}+\boldsymbol{P}_{\boldsymbol{H}}\boldsymbol{a}_{||}\|^2$.

C. OR detector

In this approach, the field detection test (4), which assumes attacks are absent, is applied simultaneously with a second detector testing for the presence of an attack, as described below, and then the results of both tests are fused using the OR rule. The rationale for this is that, whereas the performance of the field detector (4) will degrade as the attack power increases, that of the attack detector should improve. Since the adversary will only launch an attack to cover up the presence of the spatial field, an OR fusion rule is well motivated.

The detector for $\mathcal{H}_0 : \boldsymbol{a} = \boldsymbol{0}$ vs. $\mathcal{H}_1 : \boldsymbol{a} \neq \boldsymbol{0}$ can be developed as per Sec. II-A, by regarding the attack as signal and the field as interference. We assume an orthogonal attack (a similar approach is possible for unstructured attacks). Thus,

$$\boldsymbol{A}=\begin{bmatrix}\boldsymbol{0}\\ \tilde{\boldsymbol{U}}_{\perp}\end{bmatrix},\quad \boldsymbol{B}=\boldsymbol{H}\ \Rightarrow\ \begin{cases}p &= n-t-\tilde{r},\\ q &= m.\end{cases} \tag{22}$$

The GLRT exists if $t > m-\tilde{r}$. Since $\boldsymbol{P}^{\perp}_{\boldsymbol{B}} = \boldsymbol{P}^{\perp}_{\boldsymbol{H}}$, one has $\boldsymbol{G} = \boldsymbol{P}^{\perp}_{\boldsymbol{H}}\boldsymbol{A} = \boldsymbol{A}$, which has orthonormal columns; hence, $\boldsymbol{P}_{\boldsymbol{G}} = \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{P}$, with $\boldsymbol{P}$ given by (10). Then $\boldsymbol{P}_{\boldsymbol{G}}\boldsymbol{P}^{\perp}_{\boldsymbol{B}} = \boldsymbol{P}\boldsymbol{P}^{\perp}_{\boldsymbol{H}} = \boldsymbol{P}$, and $\boldsymbol{P}^{\perp}_{\boldsymbol{G}}\boldsymbol{P}^{\perp}_{\boldsymbol{B}} = \boldsymbol{P}^{\perp}_{\boldsymbol{H}} - \boldsymbol{P}$, yielding the test statistic

$$T_{\rm AS} \triangleq \frac{t+\tilde{r}-m}{n-t-\tilde{r}}\,\frac{\|\boldsymbol{P}^{\perp}_{\tilde{\boldsymbol{H}}}\tilde{\boldsymbol{y}}\|^2}{\|\boldsymbol{P}^{\perp}_{\tilde{\boldsymbol{H}}}\boldsymbol{y}\|^2-\|\boldsymbol{P}^{\perp}_{\tilde{\boldsymbol{H}}}\tilde{\boldsymbol{y}}\|^2} \tag{23}$$

$$\sim \mathcal{F}'_{n-t-\tilde{r},t+\tilde{r}-m}\left(\frac{\|\boldsymbol{a}\|^2}{\sigma^2}\right), \tag{24}$$

with (24) valid if $\boldsymbol{a} \in \mathcal{R}^{\perp}(\boldsymbol{H})$. Otherwise, let $\boldsymbol{a}$ be given by (13). Then, the distribution of $T_{\rm AS}$ changes to

$$T_{\rm AS} \sim \mathcal{F}''_{n-t-\tilde{r},t+\tilde{r}-m}\left(\frac{\|\boldsymbol{a}_{\perp}\|^2}{\sigma^2},\frac{\|\boldsymbol{P}^{\perp}_{\boldsymbol{H}}\boldsymbol{a}_{||}\|^2}{\sigma^2}\right). \tag{25}$$

The detection performance of (25) degrades with $\|\boldsymbol{P}^{\perp}_{\boldsymbol{H}}\boldsymbol{a}_{||}\|^2$, but again by virtue of Lemma 2, the adversary cannot increase such term without the risk of increasing the term $\rho_{||}$ in (5) and, in turn, the detection probability of the spatial field detector.

The following result holds now:

Lemma 3. *In the absence of attacks, the test statistics* $T_{\rm F}$ *from* (4) *and* $T_{\rm AS}$ *from* (23) *are statistically independent.*

We skip the proof for lack of space. By Lemma 3, the probability of false alarm of the OR detector is $P_{\rm FA} = P_{{\rm FA},1} + P_{{\rm FA},2} - P_{{\rm FA},1}P_{{\rm FA},2}$, with $P_{{\rm FA},i}$ the probabilities of false alarm of the individual tests. A degree of freedom is available to choose $P_{{\rm FA},i}$; for a target $P_{\rm FA}$, we suggest taking $P_{{\rm FA},1} = P_{{\rm FA},2} = 1-\sqrt{1-P_{\rm FA}} \approx \frac{1}{2}P_{\rm FA}$, as this usually yields an acceptable tradeoff between the two individual tests. Regarding the probability of detection of the OR detector, even though independence of $T_{\rm F}$ and $T_{\rm AS}$ may not hold when under attack, we propose the approximation $P_{\rm D} \approx P_{{\rm D},1} + P_{{\rm D},2} - P_{{\rm D},1}P_{{\rm D},2}$, with $P_{{\rm D},i}$ the detection probabilities of the individual tests.

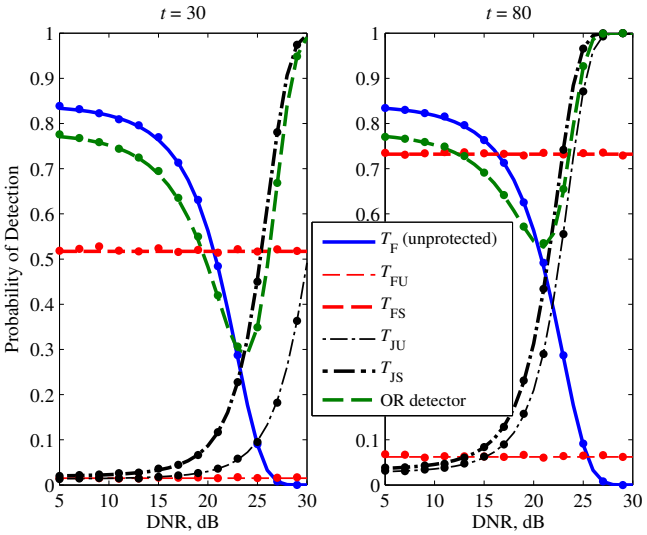

Fig. 1. $P_{\rm D}$ vs. DNR. Analytical (lines), simulation (markers). $n = 400$, $m = 20$, $P_{\rm FA} = 0.01$, SNR $= 15$ dB.

V. Results

We consider a network with $n = 400$ nodes and $m = 20$, setting $P_{\rm FA} = 0.01$. It is assumed that $\tilde{\boldsymbol{H}}$ is full rank ($\tilde{r} = m$). Fig. 1 shows the detection probability of the different schemes for 15 dB SNR, as a function of the Distortion-to-Noise Ratio (DNR) $\frac{\|\boldsymbol{a}\|^2}{\sigma^2}$ under orthogonal attacks $\boldsymbol{a} \in \mathcal{R}^{\perp}(\boldsymbol{H})$. Analytical results are shown together with those from Monte Carlo simulations (with entries of $\boldsymbol{H}$ independently drawn from a $\mathcal{N}(0,1)$ distribution, and $k = 22$ compromised sensors) for 30 and 80 trusted nodes. As expected, the unprotected detector (4), which does not account for attacks, performs best for low DNR but its detection probability goes to zero as DNR increases. The remaining schemes present different performance tradeoffs between the low- and high-DNR regimes. Since it only uses data from trusted sensors, the detector $T_{\rm FU}$ from (8) performs poorly if t/n is small. The behavior of $T_{\rm JU}$ from (16) and $T_{\rm JS}$ from (19) is better for high DNR, but not in the low DNR region for small t/n. In contrast, the detector $T_{\rm FS}$ from (11) and the OR detector from Sec. IV-C provide a much better performance-security tradeoff. Whereas $T_{\rm FS}$ is insensitive to orthogonal attacks, it may present a sizable loss for low DNR if t/n is small. Although the OR detector is affected by attacks, it performs well in the low- and high-DNR regions, and only for an interval of medium DNR values it is outperformed by the $T_{\rm FS}$ detector. The proposed analytical approximation for the probability of detection of the OR scheme is seen to accurately match the empirical results. This was the case in all settings investigated.

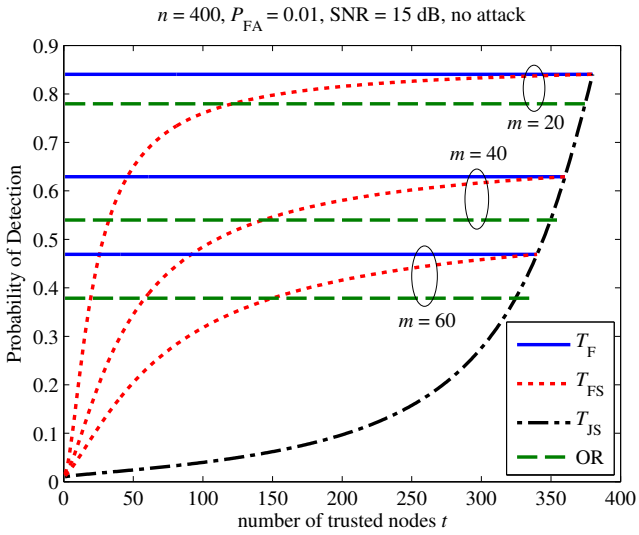

Fig. 2. $P_{\rm D}$ vs. t in the absence of attacks.

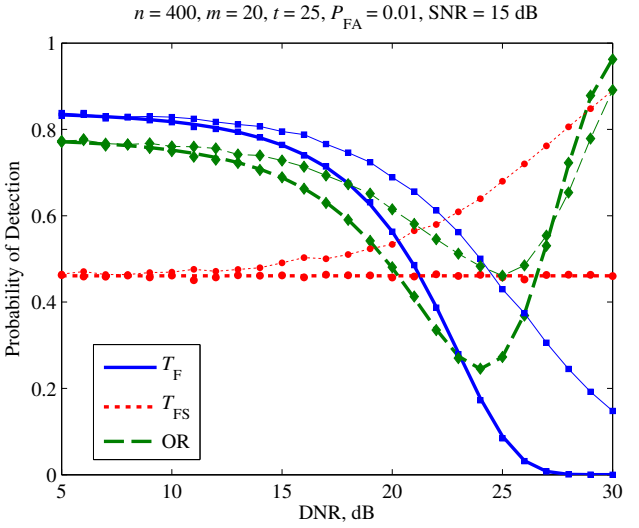

Fig. 3. $P_{\rm D}$ for orthogonal (thick) and blind attacks (thin lines).

To better gauge the improvement of the proposed schemes with the number of trusted nodes, Fig. 2 plots $P_{\rm D}$ vs. t at 15 dB SNR in the absence of attacks. This shows the loss with respect to the $T_{\rm F}$ benchmark incurred in the low DNR region (the price paid for robustness against attacks). In this regime, the loss of the OR detector is constant with t (for this detector, the benefit of a larger t is reflected in the medium DNR regime, in which the "dip" in the $P_{\rm D}$ vs. DNR curve becomes less pronounced as t increases, as seen in Fig. 1). The performance of $T_{\rm JS}$ improves very slowly with t. The improvement is significantly faster for $T_{\rm FS}$, although the number of trusted nodes it requires to outperform the OR detector is relatively large (of the order of $\frac{n}{3}$). Thus, the OR detector is a good choice when t/n is small; otherwise, the $T_{\rm FS}$ detector may be preferrable.

We also checked (via Monte Carlo) the impact of *blind attacks:* entries of $\boldsymbol{a}$ for compromised nodes are $a_i \sim \mathcal{N}(0,1)$ i.i.d., and scaled to meet a given DNR. As Fig. 3 shows, even for detectors which assume orthogonal attacks ($T_{\rm FS}$ and OR), with blind attacks the probability of detection is actually larger than with orthogonal attacks (because the adversary cannot successfully exploit the attack component in the signal subspace), which justifies such worst-case assumption.

VI. Conclusions

With trusted nodes available, robust detectors against malicious attacks have been developed and analyzed under the umbrella of matched subspace detection. Exploiting the structure of the attack vector may significantly reduce the number of required trusted nodes. The $T_{\rm FS}$ and OR detectors emerge as attractive choices with different performance-robustness tradeoffs. Our results can be used in network design to evaluate these tradeoffs, offering guidelines as to how many sensors must be secured for a given performance target.

References

[1] P. Corke et al., "Environmental wireless sensor networks," *Proc. IEEE*, vol. 98, no. 11, pp. 1903–1916, Nov. 2010.

[2] J.-F. Chamberland and V. V. Veeravalli, "Wireless sensors in distributed detection applications," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 16–25, May 2007.

[3] A. Perrig, J. Stankovic, and D. Wagner, "Security in wireless sensor networks," *Commun. ACM*, vol. 47, no. 6, pp. 53–57, Jun. 2004.

[4] Xiangqian Chen, K. Makki, Kang Yen, and N. Pissinou, "Sensor network security: a survey," *IEEE Commun. Surveys & Tutorials*, vol. 11, no. 2, pp. 52–73, 2nd quarter 2009.

[5] M. C. Vuran, Ö. B. Akan, and I. F. Akyildiz, "Spatio-temporal correlation: theory and applications for wireless sensor networks," *Computer Networks*, vol. 45, pp. 245–259, 2004.

[6] C. Guestrin et al., "Distributed regression: an efficient framework for modeling sensor network data," in *Int. Symp. Info. Process. Sensor Networks (IPSN)*, 2004, pp. 2394–2398.

[7] A. Ribeiro and G. Giannakis, "Bandwidth-constrained distributed estimation for wireless sensor networks–Part II," *IEEE Trans. Signal Process.*, vol. 54, pp. 2784–2796, Jul. 2006.

[8] A. Dogandžić and K. Qiu, "Decentralized random-field estimation for sensor networks using quantized spatially correlated data and fusion-center feedback," *IEEE Trans. Signal Process.*, vol. 56, no. 12, pp. 6069–6085, Dec. 2008.

[9] R. López-Valcarce and D. Romero, "Design of data-injection adversarial attacks against spatial field detectors," in *IEEE Workshop Stat. Signal Process. (SSP)*, 2016.

[10] S. Cui et al., "Coordinated data-injection attack and detection in the smart grid," *IEEE Signal Process. Mag.*, vol. 29, no. 5, pp. 106–115, Sep. 2012.

[11] J. Kim, L. Tong, and R. J. Thomas, "Subspace methods for data attack on state estimation: a data driven approach," *IEEE Trans. Signal Process.*, vol. 63, pp. 1102–1114, Mar. 2015.

[12] A. Pfitzmann, B. Pfitzmann, M. Schunter, and M. Waidner, "Trusting mobile user devices and security modules," *IEEE Computer*, vol. 30, no. 2, pp. 61–68, 1997.

[13] P. Schaffer, K. Farkas, A. Horvath, T. Holczer, and L. Buttyán, "Secure and reliable clustering in wireless sensor networks: A critical survey," *Computer Networks*, vol. 56, no. 11, pp. 2726–2741, Jul. 2012.

[14] R. López-Valcarce and D. Romero, "Attack detectors for data aggregation in clustered sensor networks," in *European Signal Process. Conf. (EUSIPCO)*, 2015.

[15] S. Ganeriwal, L. K. Balzano, and M. B. Srivastava, "Reputation-based framework for high integrity sensor networks," *ACM Trans. Sensor Networks*, vol. 4, no. 3, May 2008.

[16] J. López, R. Román, I. Agudo, and C. Fernández-Gago, "Trust management systems for wireless sensor networks: Best practices," *Computer Communications*, vol. 33, no. 9, pp. 1086–1093, Jun. 2010.

[17] L. L. Scharf and B. Friedlander, "Matched subspace detectors," *IEEE Trans. Signal Process.*, vol. 42, no. 8, pp. 2146–2157, Aug. 1994.

[18] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, vol. 2, John Wiley, 2nd edition, 1995.