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 sen-
sors, an adversary can drive the probability of detection in a sen-
nor 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 sen-
sors, cyber security, sensor networks.

I. INTRODUCTION

Sensor networks for surveillance and environmental moni-
toring [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 measure-
ments of the monitored physical phenomenon exhibit spatial
smoothness [5]. This allows for parsimonious parametric mod-
eling 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.

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 malici-
sous 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 $A$, $A^\dagger$ denotes its pseudoinverse,
and $\mathcal{R}(A)$, $\mathcal{R}^\perp(A)$ its column and null spaces, re-
spectively. The perpendicular projection matrices onto $\mathcal{R}(A)$
and $\mathcal{R}^\perp(A)$ are respectively denoted by $P_A = AA^\dagger$ and
$P_A^\perp = I - P_A$. The Gaussian distribution with mean $\mu$
covariance $\Sigma$ is denoted by $\mathcal{N}(\mu, \Sigma)$. By $F_{\nu_1, \nu_2}$,
and $F_{\nu_1, \nu_2}(\lambda_1, \lambda_2, \nu_2)$ we respectively denote the central,
non-central, and doubly noncentral $F$-distributions with $\nu_1$ and
$\nu_2$ degrees of freedom (d.o.f) and noncentrality parameters $\lambda_1$, $\lambda_2$ [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 $y \doteq [y_1 \cdots y_n]^T$, modeled as

$$y = Hx + a + w \in \mathbb{R}^n,$$ (1)

where $H \in \mathbb{R}^{n \times m}$ is a known full-rank matrix, $x \in \mathbb{R}^m$ is related
to the monitored physical process, and $w \sim \mathcal{N}(0, \sigma^2 I_n)$ is
measurement noise with unknown variance $\sigma^2$. The attack
vector $a = [a_1 \cdots a_n]^T \in \mathbb{R}^n$ 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 $x = 0$. The basis expansion
model $Hx$ 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 \( x \) \((m \ll n)\). It fits a wide range of signal representations based on Fourier or Discrete Cosine Transforms, polynomial bases, splines, etc.\([6, 8]\). Thus, \( H \) 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 \((x \neq 0)\), 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
\[
\begin{align*}
    y &= A\theta + v, \quad A \in \mathbb{R}^{n \times p}, \quad \theta \in \mathbb{R}^p, \\
    v &\sim N(B\phi, \sigma^2 I_n), \quad B \in \mathbb{R}^{n \times q}, \quad \phi \in \mathbb{R}^q,
\end{align*}
\]
with \( A, B \) known, and \( \theta, \phi, \sigma^2 \) unknown. The matrices \( A, B \) and \( [A \ B] \) are assumed full-rank. We wish to test for the presence of the signal component \( A\theta; \ H_0 : y \sim N(B\phi, \sigma^2 I_n) \) vs. \( H_1 : y \sim N(A\theta + B\phi, \sigma^2 I_n) \). The component \( B\phi \) can be regarded as an interference term. Thus, in this model, the signal term is known to lie in the subspace \( \mathcal{R}(A) \), whereas the interference is known to lie in the subspace \( \mathcal{R}(B) \).

Let \( G = \overline{P_B A} \). Provided that \( p + q < n \), the Generalized Likelihood Ratio Test (GLRT) exists and is given by
\[
T = \frac{n - p - q}{p} \frac{\|P_G \overline{P_B y}\|^2}{\|P_G \overline{P_B y}\|^2} \sim F_{n - p - q, \gamma},
\]
and it is uniformly most powerful invariant (UMPI) \([17]\). In (3), the interference is first cancelled by \( \overline{P_B} \), after which the signal component is known to lie in \( \mathcal{R}(G) \). Then (3) compares the energies of \( \overline{P_B} y \) in this subspace and in its orthogonal complement \( \mathcal{R}^\perp(G) \). The test statistic \( T \) is \( F \)-distributed as \( T \mid H_0 \sim F_{p, n - p - q} \) and \( T \mid H_1 \sim F_{p, n - p - q} \left( \frac{\|G\|^2}{\sigma^2} \right) \).

For fixed \( \frac{\|G\|^2}{\sigma^2} \), 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 \( a = 0 \). Under this assumption, the problem of testing for field presence, i.e., \( H_0 : x = 0 \) vs. \( H_1 : x \neq 0 \), can be cast in the framework of (2) with \( A\theta = Hx \) and \( B\phi = 0 \) (and thus \( p = m \) and \( q = 0 \)). The corresponding GLRT is
\[
T_F = \frac{n - m}{m} \frac{\|P_H y\|^2}{\|P_H y\|^2} \sim F_{n - m - p, \gamma_F},
\]
and in the absence of attacks, one has \( T_F \mid H_0 \sim F_{m, n - m} \) and \( T_F \mid H_1 \sim F_{m, n - m}(\rho) \), where \( \rho = \frac{\|Hx + P_H a\|^2}{\sigma^2} \) is the Signal-to-Noise Ratio (SNR). However, as shown in \([9]\), the presence of an attack changes the distribution of \( T_F \) under \( H_1 \) to
\[
T_F \sim F_{m, n - m}(\rho ||, \rho_\perp) \quad \text{with} \quad \begin{cases} 
\rho_|| \triangleq \frac{\|Hx + P_H a\|^2}{\sigma^2} \\
\rho_\perp \triangleq \frac{\|P_H a\|^2}{\sigma^2}.
\end{cases}
\]

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 \( H \) (which could be inferred from sensor locations), but not \( x \) (which depends on unknown environmental variables affecting field diffusion). Thus, to avoid accidentally increasing \( \rho_|| \), the adversary may choose an orthogonal attack \( a \in \mathcal{R}^\perp(H) \); 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 \( \|a\|^2 \) \([9]\). Clearly, appropriate countermeasures are needed.

Note that directly applying the results in Sec. II-A to (1) (by regarding \( a \) 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 \( a \) must lie. At most, by the reasoning in Sec. II-B, one could assume that \( a \in \mathcal{R}^\perp(H) \), 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 \( a = [0 \bar{a}_1 \bar{a}_2 \ldots]^T \) with \( \bar{a} \in \mathbb{R}^{n-t} \). Letting \( H \in \mathbb{R}^{t \times m} \) and \( H \in \mathbb{R}^{(n-t) \times m} \) comprise the first \( t \) and last \( n - t \) rows of \( H \) respectively, model (1) can be split as
\[
\begin{align*}
\bar{y} &\equiv \bar{H} \bar{x} + \bar{w}, \\
\tilde{y} &\equiv \bar{H} \bar{x} + \bar{a} + \tilde{w},
\end{align*}
\]
with \( \bar{y} \equiv [y_1 \ldots y_t]^T, \tilde{y} \equiv [y_{t+1} \ldots y_n]^T \) (data from trusted and untrusted nodes respectively), and \( \tilde{w}, \bar{w} \) 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 \( \bar{a} \), whereas the second assumes a structured (orthogonal) attack.

A. Unstructured attack

With no assumptions on \( \bar{a} \), then in the framework of Sec. II-A, one may take
\[
A = H, \quad B = \begin{bmatrix} 0 \\ \bar{H} \\ \bar{n}_{n-t} \end{bmatrix} \Rightarrow \begin{cases} 
 p = m, \\
 q = n - t.
\end{cases}
\]
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

$$T_{FU} \triangleq \frac{t - m}{m} \frac{\| P_H \bar{y} \|^2}{\| P_H^+ \bar{y} \|^2} \sim F'_{m,t-m} \left( \frac{\| Hx \|^2}{\sigma^2} \right). \quad (8)$$

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

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

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

Let $\tilde{r} = \text{rank } \tilde{H}$, and let the columns of $\tilde{U}_1 \in \mathbb{R}^{(n-t) \times (n-t-\tilde{r})}$ be an orthonormal basis for $\mathcal{R}^\perp(H)$, so that $P_{\tilde{H}} = \tilde{U}_1 \tilde{U}_1^T$. Assume that $a \in \mathcal{R}^\perp(H)$.

Then, in view of Lemma 1, field detection can be cast in the framework of Sec. II-A by setting

$$A = H, \quad B = \begin{bmatrix} 0 & \tilde{U}_1 \end{bmatrix} \Rightarrow \begin{cases} p = m, \\ q = n - t - \tilde{r}. \end{cases} \quad (9)$$

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

$$P_{B}^+ = I - P_{G} \quad \text{with} \quad P = \begin{bmatrix} 0 & P_{G}^+ \\ P_{G} & H \end{bmatrix} \quad (10)$$

Since $P_{H} P = 0$, it follows that $P_{G} P_{B}^+ = P_{H}$ and $P_{G}^+ P_{B} = P_{H}^T - P$. This results in the GLRT statistic

$$T_{FS} \triangleq \frac{t + \tilde{r} - m}{m} \frac{\| P_H \bar{y} \|^2}{\| P_H^+ \bar{y} \|^2} \sim F'_{m,t+\tilde{r}-m} \left( \frac{\| Hx \|^2}{\sigma^2} \right), \quad (11)$$

which coincides with the distribution of (4) under no attacks, $T_{F} \sim F'_{m,n-m} \left( \frac{\| Hx \|^2}{\sigma^2} \right)$, 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 $a \in \mathcal{R}^\perp(H)$, note from Lemma 1 that this implies $a = [0_t^T \bar{a}^T]^T$ with $\bar{a} \notin \mathcal{R}^\perp(H)$. Hence, let

$$a = \begin{bmatrix} 0_t & \bar{a}_1^T + \bar{a}_\perp^T \\ \bar{a}_1 & \bar{a}_\perp \end{bmatrix} \begin{cases} \bar{a}_1^T & \in \mathcal{R}(H) \\ \bar{a}_\perp & \in \mathcal{R}^\perp(H). \end{cases} \quad (13)$$

(Note that, in general, $\bar{a}_1^T$ does not lie either in $\mathcal{R}(H)$ nor in $\mathcal{R}^\perp(H)$). Then the distribution of $T_{FS}$ changes from (12) to

$$T_{FS} \sim F'_{m,t+\tilde{r}-m} \left( \frac{\| Hx + P_H \bar{a}_1 \|^2}{\| P_H a \|^2}, \frac{\| P_H a \|^2}{\sigma^2} \right). \quad (14)$$

The detection probability is monotonically increasing (resp. decreasing) in the first (resp. second) noncentrality parameter of (14). Again, since $x$ is unknown to the adversary, he is unable to design the attack to reduce $\| Hx + P_H a \|^2$, with a "poor" choice for $a_1$, this term may actually increase. Thus, a sensible approach for the attacker would be to maximize $\| P_{H}^+ a_1 \|^2$ subject to $P_{H} a_1 = 0$. However, one has:

**Lemma 2.** If $a_1 = [0_t^T \bar{a}_1^T]^T$ with $\bar{a}_1 \in \mathcal{R}(H)$, then $P_{H} a_1 = 0$ implies $a_1 = 0$.

**Proof.** $P_{H} a_1 = 0$ means $a_1 \in \mathcal{R}^\perp(H)$, i.e., $H^T a_1 = H^T \bar{a}_1 = 0$. Thus, $\bar{a}_1 \in \mathcal{R}^\perp(H)$, but since $\bar{a}_1 \notin \mathcal{R}(H)$ by assumption, it follows that $\bar{a}_1 = 0$, hence $a_1 = 0$. ■

Thus, any attack with $\| P_{H}^+ a_1 \|^2 > 0$ must yield $P_{H} a_1 \neq 0$, 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 $a$ 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., $H_0 : Hx + a = 0$ vs. $H_1 : Hx + a \neq 0$. Again, nodes $1, \ldots, t$ are assumed trusted, so that $a = [0_t^T \bar{a}^T]^T$.

A. Unstructured attack

With no assumptions on $\bar{a}$, then in the framework of Sec. II-A, $B$ is empty (thus, $P_{B} = I$) and $A \theta = Hx + a$. Hence,

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

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

$$T_{FS} \triangleq \frac{t - m}{m+n-t} \frac{\| P_H \bar{y} \|^2}{\| P_H^+ \bar{y} \|^2} \sim F'_{m,n-m} \left( \frac{\| Hx + a \|^2}{\sigma^2} \right). \quad (16)$$

B. Structured attack

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

$$A = \begin{bmatrix} H & 0 \\ H & \tilde{U}_1 \end{bmatrix} \Rightarrow \begin{cases} p = m + n - t - \tilde{r}, \\ q = 0. \end{cases} \quad (18)$$

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

$$T_{JS} \triangleq \frac{t + \tilde{r} - m}{m+n-t-\tilde{r}} \frac{\| P_{H} \bar{y} \|^2 + \| P_H^+ \bar{y} \|^2}{\| P_{H} \bar{y} \|^2 - \| P_H^+ \bar{y} \|^2} \sim F'_{m+n-t-\tilde{r},t+\tilde{r}-m} \left( \frac{\| Hx \|^2}{\sigma^2} + \frac{\| a \|^2}{\sigma^2} \right). \quad (19)$$
with (20) valid if $a \in \mathcal{R}^\perp(H)$. Otherwise, with $a$ given by (13), the distribution of $T_{IS}$ changes to

$$f_{m+n-t-\bar{r},t+\bar{r}-m}^\prime \left( \frac{\|Hx + PHa\|}{\sigma^2}^2 + \frac{\|a\|}{\sigma^2}^2, \frac{\|P_H^\perp a\|}{\sigma^2}^2 \right)$$

Any $a \perp 0$ will contribute to increasing the probability of detection. Regarding the component $a_i$, in view of Lemma 2, the adversary cannot increase the term $\|P_H^\perp a_i\|^2$ in (21) without the risk of increasing the term $\|Hx + PHa_i\|^2$.

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 $H_0 : a = 0$ vs. $H_1 : a \neq 0$ 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,

$$A = \begin{bmatrix} 0 \\ \tilde{U}_\perp \end{bmatrix}, \quad B = H \Rightarrow \begin{cases} \rho = n - t - \bar{r}, \\ q = m. \end{cases}$$

The GLRT exists if $t > m - \bar{r}$. Since $P_B^\perp = P_H^\perp$, one has $G = P_H^\perp A = A$, which has orthonormal columns; hence, $P_G = AA^T = P$, with $P$ given by (10). Then $P_GP_B^\perp = PP_H^\perp = P$ and $P_GP_B^\perp = P - P_H^\perp$, yielding the test statistic

$$T_{AS} \triangleq \frac{t + \bar{r} - m}{n - t - m} \frac{\|P_H^\perp \tilde{y}\|^2}{\|P_H^\perp \tilde{y}\|^2 + \|P_H^\perp \tilde{y}\|^2}$$

$$\sim \mathcal{F}_{m+n-t-\bar{r},t+\bar{r}-m}^\prime \left( \frac{\|a\|^2}{\sigma^2}^2, \frac{\|P_H^\perp a\|^2}{\sigma^2}^2 \right),$$

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

$$T_{AS} \sim \mathcal{F}_{m+n-t-\bar{r},t+\bar{r}-m}^\prime \left( \frac{\|a\|^2}{\sigma^2}^2, \frac{\|P_H^\perp a\|^2}{\sigma^2}^2 \right).$$

The detection performance of (25) degrades with $\|P_H^\perp a\|^2$, but again by virtue of Lemma 2, the adversary cannot increase such term without the risk of increasing the term $\rho_i$ 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 $T_F$ from (4) and $T_{AS}$ 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 $P_{FA} = P_{FA,1} + P_{FA,2} - P_{FA,1}P_{FA,2}$, with $P_{FA,i}$ the probabilities of false alarm of the individual tests. A degree of freedom is available to choose $P_{FA,i}$; for a target $P_{FA}$, we suggest taking $P_{FA,1} = P_{FA,2} = 1 - \sqrt{1 - P_{FA}} \approx \frac{1}{2} P_{FA}$, 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 $T_F$ and $T_{AS}$ may not hold when under attack, we propose the approximation $P_{D} \approx P_{D,1} + P_{D,2} - P_{D,1}P_{D,2}$, with $P_{D,i}$ the detection probabilities of the individual tests.

V. RESULTS

We consider a network with $n = 400$ nodes and $m = 20$, setting $P_{FA} = 0.01$. It is assumed that $H$ is full rank ($\bar{r} = m$). Fig. 1 shows the detection probability of the different schemes for 15 dB SNR, as a function of the Distortion-to-Noise Ratio (DNR) $\|a\|^2$ under orthogonal attacks $a \in \mathcal{R}^\perp(H)$. Analytical results are shown together with those from Monte Carlo simulations (with entries of $H$ independently drawn from a $N(0,1)$ 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 trade-offs between the low- and high-DNR regimes. Since it only uses data from trusted sensors, the detector $T_{FU}$ from (8) performs poorly if $t/n$ is small. The behavior of $T_{JS}$ from (16) and $T_{IS}$ from (19) is better for high DNR, but not in the low DNR region for small $t/n$. In contrast, the detector $T_{FS}$ from (11) and the OR detector from Sec. IV-C provide a much better performance-security tradeoff. Whereas $T_{FS}$ 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 $T_{FS}$ 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.
To better gauge the improvement of the proposed schemes with the number of trusted nodes, Fig. 2 plots $P_D$ vs. $t$ at 15 dB SNR in the absence of attacks. This shows the loss with respect to the $T_F$ 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 $P_D$ vs. DNR curve becomes less pronounced as $t$ increases, as seen in Fig. 1). The performance of $T_{JS}$ improves very slowly with $t$. The improvement is significantly faster for $T_{FS}$, although the number of trusted nodes it requires to outperform the OR detector is relatively large (of the order of $\frac{1}{2}$). Thus, the OR detector is a good choice when $t/n$ is small; otherwise, the $T_{FS}$ detector may be preferable.

We also checked (via Monte Carlo) the impact of blind attacks: entries of $\alpha$ for compromised nodes are $\alpha_i \sim \mathcal{N}(0, 1)$ i.i.d., and scaled to meet a given DNR. As Fig. 3 shows, even for detectors which assume orthogonal attacks ($T_{FS}$ 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 $T_{FS}$ 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