Adversarial Detection as a Zero-Sum Game

K. G. Vamvoudakis[1], Member, IEEE J. P. Hespanha[1], Fellow IEEE B. Sinopoli[2], Member, IEEE Yilin Mo[2]

Abstract— **We propose a new game theoretic approach to estimate a binary random variable based on a vector of sensor measurements that may be corrupted by an adversary. The problem is formulated as a zero-sum partial information game in which a detector attempts to minimize the probability of error and an attacker attempts to maximize this probability. Explicit mixed policies are computed using the matrix form of the game and exploiting sensor symmetry to reduce complexity.**

Index Terms— **Adversarial detection, computer security, zero-sum games, estimation, mixed policies.**

I. Introduction

Embedded sensing [11], computation, and communication have enabled the proliferation of sophisticated sensing devices for a wide range of applications that include safety monitoring, health-care related applications, surveillance, traffic monitoring, military applications, and cyber-physical systems [6]. While computer-based networked sensors provide a tremendous level of flexibility for these application, they also introduce significant security vulnerabilities because a sensor can be compromised without physical access to the device. Of particular concern are scenarios in which an attacker *infiltrates a sensing device and manipulates its output in a manner that cannot be easily detected by the system.* These scenarios force system designers to rethink basic estimation problems in light of network security concerns.

In traditional estimation problems one attempts to determine the value of a variables that cannot be measured directly based on a set of "noisy" measurements of that variable. Typically, some form of probabilistic structure is assumed to model how the measurements relate to the true value of the variable to be estimated. This type of framework is adequate, e.g., when the measurements fluctuate around the variable's true value due to microscopic thermal fluctuations. However, things can be quite different in scenarios where measurement can be controlled by an entity that actively attempts to degrade the estimation process.

The most basic mechanism to overcome stochastic measurements errors relies on the use of redundancy. When multiple sensors provide redundant and independent measurements about a variable that needs to be estimated, the confidence on the estimate increases with the number of sensors. When some of these sensors are under control of an adversary that wants to maximize the estimation error, the independence assumption is generally not valid and the probability of an estimation error scales differently with the number of sensors. The goal of this paper is to provide insights regarding what happens in such situations.

To focus our attention on the issues that arise when one needs to do estimation using potentially compromised sensors, we consider a prototypical problem in which one wants to estimate the value of a binary random variable based on measurements provided by a group of binary sensors. We assume that such measurements incorporate two types of errors: purely stochastic error that are responsible for bit-flips with a pre-specified probability and adversarial errors that are controlled by an adversary that has infiltrated a subset of the sensors. Which sensors have been infiltrated is not known a-priori to the system. We shall see that the (optimal) adversarial errors may actually be stochastic, with probability distributions carefully selected by the attacker to maximize the probability of estimation error (corresponding to mixed policies). In general, these distributions will be a function of the value of the variable to be estimated.

The adversarial estimation problem described above is formulated as a zero-sum game between a player that wants to estimate the binary random variable with minimal probability of error, henceforth called the *detector*, and a player that wants to maximize the same probability of error, henceforth called the *attacker*. This is a game of partial information in that the detector only has access to a "noisy" sensor measurement that has been corrupted both by stochastic and by adversarial errors. Similarly, the attacker also only has partial information since we assume that it knows the true value of the variable to be estimated, but not the measurements that are being reported by the sensors that she has not infiltrated.

By expanding the game into its matrix form and exploiting sensor symmetry to reduce complexity, we can obtain optimal estimation policies for the detector and optimal sensor manipulation policies for the attacker. Policy domination is used to reduce the apparent exponential complexity of the problem, eventually leading to simple detection and attack policies.

To model the fact that the detector may not be certain that an attacker may actually be infiltrating some of the sensors, we introduce a "probability of attack" p_{attack}, which reflects how certain the detector is about the existence of a malicious

*This material is based upon work supported in part by ARO MURI Grant number W911NF0910553.

[1] K. G. Vamvoudakis, and J. P. Hespanha are with the Center for Control, Dynamical-systems and Computation (CCDC), University of California, Santa Barbara, CA 93106-9560 USA e-mail: kyriakos@ece.ucsb.edu

[2] B. Sinopoli, and Yilin Mo are with the Electrical and Computer Engineering Department, Carnegie Mellon University, Pittsburgh, PA, USA.

attacker. An interesting feature of the solution obtained is that the optimal estimation policy is largely insensitive to this parameter p_{attack} that typically is hard to guess.

Related Work

Several game-theoretic approaches [5], [10] have been proposed to wired networks, WLANs, sensor networks, and ad hoc networks. In [3] the authors propose a game theoretic approach to intrusion detection in distributed virtual sensor networks, where each agent in the network has imperfect detection capabilities. This interaction between the defender and the attacker is modeled as a noncooperative non-zero sum game. A two-player noncooperative, non-zero-sum game has also been studied by [1] and [2] to address attack-defense problems in sensor networks. Kodialam et al. [7] have proposed a zero-sum game to model the intrusion detection game between the service provider and the intruder. The optimal solution for both players is to play the minimax strategy of the game. Game theoretic solutions for ad hoc networks based on cooperation and selfishness of the network have been reported in [9], [12], where each node decides whether to forward or not a packet based on payoff functions. Researchers in [8] propose a denial-of-service attack game where an attacker on the Internet is trying to deface the homepage on a given server. A stochastic game approach is proposed between the network administrator and the attacker where at each time step, both players choose their policies and the game moves to a new state according to some probability that depends on the chosen policies. Through simulations, the authors have shown that the game admits multiple Nash equilibria. Since a Nash equilibrium gives to the defender an idea about the attacker's best strategies, finding more Nash equilibria means having more information about the attack.

The remainder of the paper is structured as follows. In Section II, we provide a description of the problem. Section III discusses how one can use symmetry to reduce the complexity of the problem and Section IV contains the main results of the paper, providing optimal detection and attack policies. Finally, in Section V we conclude the paper and discuss future work.

II. Problem Formulation

The goal of this paper is to estimate the value of a binary random variable X with Bernoulli distribution

$$\mathrm{P}(X=1) = 1 - \mathrm{P}(X=1) = p \in (0,1),$$

based on a vector $Y \coloneqq (Y_1, Y_2, \ldots, Y_n)$ of n binary "noisy" sensor measurements, where the measurements Y_i, $i \in \{1, 2, \ldots, n\}$ are assumed conditionally independent and identically distributed (iid), given X. Specifically,

$$\mathrm{P}(Y_i = 1 \,|\, X, Y_{j\neq i}) = \begin{cases} p_{\text{err}} & X = 0, \\ 1 - p_{\text{err}} & X = 1, \end{cases}$$

where $p_{\text{err}} \in [0,1]$ denotes a per-sensor error probability. Setting us a part from standard estimation problems, we consider a scenario where an estimate $\hat{X}$ of X needs to be constructed based on version $Z \coloneqq (Z_1, Z_2, \ldots, Z_n)$ of the measurement vector Y that may have been "corrupted" by an attacker. It is assumed that, with a given probability $p_{\text{attack}} \in [0,1]$, the attacker manipulated the readings of $m \leqslant n$ sensors. Which sensors and in what way should the attacker manipulate them will be discussed below. The probability $p_{\text{attack}} \in [0,1]$ should be viewed as a design parameter that reflects how certain the detector is that the measurements may have been manipulated and, for $p_{\text{attack}} = 0$, we recover a standard estimation problem.

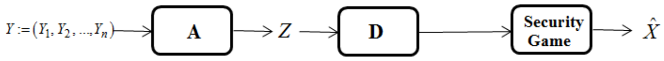

Fig. 1. Detection-Attack Model

The game under consideration is a partial information game for both players. The *detector* must select its estimate $\hat{X}$ based solely on the vector Z of possibly corrupted sensor readings, which corresponds to the selection of an *estimation policy* $\mu : \{0,1\}^n \to \{0,1\}$ that is used to compute the estimate

$$\hat{X} = \mu(Z).$$

Since the domain of μ has 2^n elements and its codomain has 2 elements, the set $\mathcal{U}$ of all possible estimation policies contains $2^{(2^n)}$ policies.

We assume that the *attacker* knows the true value of X and bases its decision on which sensors to compromise and how to corrupt their measurements as a function of X. However, the attacker is not assumed to know the values reported by the remaining sensors, thus also suffering from partial information. We can thus view an *attack policy* as a pair of functions $\delta_{\text{which}} : \{0,1\} \to \mathcal{S}_n^m$ and $\delta_{\text{how}} : \{0,1\} \to \{0,1\}^m$, where $\mathcal{S}_n^m$ denotes the set of all ordered subsets of $\{1, 2, \ldots, n\}$ with m elements, with the understanding that $\delta_{\text{which}}(X) \in \mathcal{S}_n^m$ determines which sensors will be attacked and $\delta_{\text{how}} \in \{0.1\}^m$ determines the corresponding readings set by the attacker. Specifically, if the kth element of $\delta_{\text{which}}(X)$ is equal to i, then the attacker sets Z_i equal to the kth element of $\delta_{\text{how}}(X)$. The total number of distinct functions δ_{which} and δ_{how} are equal to $\binom{n}{m}^2$ and $(2^m)^2$, respectively, so the set $\mathcal{D}$ of all possible attack policies contains $\binom{n}{m}^2 (2^m)^2$ policies.

The model just described is illustrated in Figure 1 and allow us to define adversarial estimation as a zero-sum game in which the detector selects a policy $\mu \in \mathcal{U}$ and the attacker a policy $\delta \in \mathcal{D}$ so to minimize and maximize, respectively, the probability of error

$$\mathrm{P}_{\mu,\delta}(\hat{X} \neq X), \tag{1}$$

where the subscript $_{\mu,\delta}$ in the probability measure emphasizes the fact that the probability of error depends on the players' policies. Since the sets of policies are finite, we have a (finite) matrix game defined by the matrix

$$A \coloneqq \big[a_{ij}\big]_{2^{(2^n)} \times \binom{n}{m}^2 (2^m)^2}, \tag{2}$$

where a_{ij} denotes the probability of error in (1) corresponding to the ith estimation policy in $\mathcal{U}$ and the jth attack policy in $\mathcal{D}$. In general, this game does not have pure saddle-point equilibria so the players will seek for mixed policies, which correspond to selecting probability distributions over the sets of actions $\mathcal{U}$ and $\mathcal{D}$.

III. Symmetric Games

Since all sensors are equal in their probability of error and in their vulnerability to attacks, all entries of the vector Y and Z should be treated similarly by the attacker and the detector, respectively. This allows one to significantly reduce the size of the matrix game:

1) The estimation policy $\mu(Z)$ should only depend on the total number of 0's and 1's in the vector Z and therefore can be written as

$$\mu(Z) = \bar{\mu}\Big(\sum_{i=1}^{n} Z_i\Big), \tag{3}$$

for some function $\bar{\mu} : \{0, 1, \dots, n\} \to \{0, 1\}$. The total number of such functions is given by 2^{n+1}.

2) For the attack policies, all sensor selection functions $\delta_{\text{which}}(X)$ are equally good and can therefore be selected with equal probability. Moreover, the function $\delta_{\text{how}}(X)$ only needs to decide how many sensors will be set equal to 0 and how many will be set equal to 1, with the understanding that these 0s and 1s will be distributed with equal probability among the sensors selected. The selection of how many sensors will be set equal to 0 corresponds to the selection of a function $\bar{\delta}_{\#0} : \{0, 1\} \to \{0, 1, \dots, m\}$. The total number of such functions is given by $(m+1)^2$.

These observations lead to a zero-sum game defined by a matrix $\bar{A}$ that is only $2^{n+1} \times (m+1)^2$, with one row for each policy $\bar{\mu}$ and one column for each policy $\bar{\delta}_{\text{how}}$. The following result, proved in the Appendix, can be used to compute such matrix.

Lemma 1: When the detector utilizes a policy $\mu(Z)$ of the form (3) and the attacker a policy $\delta(X)$ that tries to set to 0 and to 1 a number of sensors equal to $\bar{\delta}_{\#0}(X)$ and $m - \bar{\delta}_{\#0}(X)$, respectively (all sensors selected with equal probability), we obtain the following probability of error:

$$\begin{aligned}
\mathrm{P}_{\mu,\delta}(\hat{X} \neq X) = (1-p)\bigg(& p_{\text{attack}} \sum_{k=m-\bar{\delta}_{\#0}(0)}^{n-\bar{\delta}_{\#0}(0)} \bar{\mu}(k) \\
& \binom{n-m}{k-m+\bar{\delta}_{\#0}(0)} p_{\text{err}}^{k-m+\bar{\delta}_{\#0}(0)} (1-p_{\text{err}})^{n-k-\bar{\delta}_{\#0}(0)} \\
& + (1-p_{\text{attack}}) \sum_{k=0}^{n} \bar{\mu}(k) \binom{n}{k} p_{\text{err}}^{k} (1-p_{\text{err}})^{n-k} \bigg) \\
& + p\bigg(p_{\text{attack}} \sum_{k=m-\bar{\delta}_{\#0}(1)}^{n-\bar{\delta}_{\#0}(1)} \big(1-\bar{\mu}(k)\big) \binom{n-m}{k-m+\bar{\delta}_{\#0}(1)} \\
& (1-p_{\text{err}})^{k-m+\bar{\delta}_{\#0}(1)} p_{\text{err}}^{n-k-\bar{\delta}_{\#0}(1)} \\
& + (1-p_{\text{attack}}) \sum_{k=0}^{n} \big(1-\bar{\mu}(k)\big) \binom{n}{k} (1-p_{\text{err}})^{k} p_{\text{err}}^{n-k} \bigg).
\end{aligned} \tag{4}$$

□

IV. Main Result

It turns out that the exponential complexity in the number of sensors n can be removed using policy domination. For simplicity of presentation, we show this for the case where it is equally likely that X is 0 or 1 (i.e., $p = 1/2$) and the number of sensors is odd (allowing for tie-breaking). In this case, we can provide explicit formulas for mixed saddle-point policies and for the value of the game. This result if formulated in terms of the following (pure policies):

1) We define the detector's *majority rule* to be its pure policy $\mu(Z)$ of the form (3), defined by

$$\mu_{\text{majority}}(Z) = \begin{cases} 0 & \sum_{i=1}^{n} Z_i \leqslant \frac{n-1}{2} \\ 1 & \sum_{i=1}^{n} Z_i \geqslant \frac{n+1}{2}, \end{cases}$$

which corresponds to deciding on $\hat{X} = 0$ if more than half the sensors reported the value 0.

2) We define the detector's *no-consensus rule* to be its pure policy $\mu(Z)$ of the form (3), defined by

$$\begin{aligned} &\mu_{\text{no}-\text{consensus}}(Z) = \\ &\begin{cases} 0 & 0 < \sum_{i=1}^{n} Z_i \leqslant \frac{n-1}{2} \text{ or } \sum_{i=1}^{n} Z_i = n \\ 1 & n > \sum_{i=1}^{n} Z_i \geqslant \frac{n+1}{2} \text{ or } \sum_{i=1}^{n} Z_i = 0, \end{cases} \end{aligned}$$

this somewhat unexpected policy is like the majority rule, except that if all sensors agree on a particular value, the estimate $\hat{X}$ should take the opposite value.

3) We define the attacker's *deception rule* to be its pure policy $\delta_{\text{deception}}(X)$ that, when $X = 0$ sets all m sensors equal to 1 and when $X = 1$ sets all m sensors equal to 0.

4) We define the attacker's *no-deception rule* to be its pure policy $\delta_{\text{no}-\text{deception}}(X)$ that, when $X = 0$ sets all m sensors equal to 10 and when $X = 1$ sets all m sensors equal to 1.

Theorem 1: Suppose that $p = 1/2$, the number of sensors n is odd, and

$$m \leqslant \frac{n+1}{2}, \tag{5}$$

$$p_{\text{err}} \leqslant 1 - \frac{n-1}{2(n-m)}, \tag{6}$$

$$p_{\text{attack}} \leqslant \frac{1}{1 + \frac{1}{n}\binom{n-m}{\frac{n-1}{2}} \frac{p_{\text{err}}^{n-m-\frac{n-1}{2}} (1-p_{\text{err}})^{\frac{n-1}{2}}}{p_{\text{err}}(1-p_{\text{err}})^{n-1} - p_{\text{err}}^{n-1}(1-p_{\text{err}})}}. \tag{7}$$

In this case, the value of the game is given by

$$v^* = \alpha + p_{\text{attack}}$$

$$\min\left\{\gamma, \frac{\gamma(1-p_{\text{err}})^{n-m}+\beta\gamma+\rho(p_{\text{err}}^{n-m}-\beta)}{(1-p_{\text{err}})^{n-m}+p_{\text{err}}^{n-m}}\right\}$$

and a mixed saddle-point policy corresponds to selecting

$$\begin{cases}\mu_{\text{majority}} & \text{w.p. } 1-y_2\\ \mu_{\text{no-majority}} & \text{w.p. } y_2,\end{cases}$$

$$\begin{cases}\delta_{\text{deception}} & \text{w.p. } 1-z_2\\ \delta_{\text{no-deception}} & \text{w.p. } z_2,\end{cases}$$

where

$$y_2 = \begin{cases}\Pi\Big(\frac{\gamma-\rho}{(1-p_{\text{err}})^{n-m}+p_{\text{err}}^{n-m}}\Big) & \beta \leqslant p_{\text{err}}^{n-m}\\ 0 & \beta > p_{\text{err}}^{n-m}\end{cases}$$

$$z_2 = \Pi\Big(\frac{p_{\text{err}}^{n-m}-\beta}{(1-p_{\text{err}})^{n-m}+p_{\text{err}}^{n-m}}\Big)$$

$$\alpha \coloneqq (1-p_{\text{attack}})\sum_{k=0}^{\frac{n-1}{2}}\binom{n}{k}p_{\text{err}}^{n-k}(1-p_{\text{err}})^{k}$$

$$\rho \coloneqq \sum_{k=m}^{\frac{n-1}{2}}\binom{n-m}{k-m}p_{\text{err}}^{n-k}(1-p_{\text{err}})^{k-m}$$

$$\gamma \coloneqq \sum_{k=0}^{\frac{n-1}{2}}\binom{n-m}{k}p_{\text{err}}^{n-m-k}(1-p_{\text{err}})^{k}$$

$$\beta = \frac{1-p_{\text{attack}}}{p_{\text{attack}}}\big((1-p_{\text{err}})^{n}-p_{\text{err}}^{n}\big).$$

and $\Pi:\mathbb{R}\to\mathbb{R}$ denotes the projection function

$$\Pi(x) = \begin{cases}0 & x<0\\ x & x\in[0,1]\\ 1 & x>1.\end{cases} \qquad \square$$

A. Discussion

Conveniently, the optimal policy for the estimator is largely independent of the attack probability p_{attack}, which may be difficulty to know. In essence, as long as the probability of attack does not exceed the bound (7), the estimator's policy only depends on p_{attack} because of the threshold condition

$$\beta \coloneqq \frac{1-p_{\text{attack}}}{p_{\text{attack}}}\big((1-p_{\text{err}})^{n}-p_{\text{err}}^{n}\big) > p_{\text{err}}^{n-m},$$

which when true leads to a pure majority rules, and otherwise leads to the mixed policy given in Theorem 1. While the estimator's policy may depends little on p_{attack}, that is obviously not the case for the probability of error corresponding to the saddle-point solution. For example, for very small probabilities of error, the saddle point is essentially given by

$$v^* \approx p_{\text{attack}}\binom{n-m}{\frac{n-1}{2}}p_{\text{err}}^{\frac{n+1}{2}-m},$$

which shows a probability of error that scales linearly with the attack probability. This formula also shows that the scaling law for the probability of error scales with the number of sensors is of the form

$$p_{\text{err}}^{\frac{n+1-2m}{2}}. \tag{8}$$

In the absence of attacks (for which the majority rule would be optimal), we can conclude from Lemma 1 that the probability of error is given by

$$\mathrm{P}_{\mu,\delta}(\hat{X}\neq X) = \sum_{k=\frac{n+1}{2}}^{n}\binom{n}{k}p_{\text{err}}^{k}(1-p_{\text{err}})^{n-k},$$

which, for a small probability of error, scales with the number of sensors as

$$p_{\text{err}}^{\frac{n+1}{2}}. \tag{9}$$

From the perspective of these scaling laws, it is as if each one of the m sensors compromised effectively decreases the total number of sensors by $2m$.

B. Proof of Theorem 1

The following simple proposition is needed to prove Theorem 1.

Proposition 1: Given an integer n and a scalar $p\in(0,1)$, for every integer k such that $0\leqslant k\leqslant \ell\leqslant np$,

$$\binom{n}{k}p^{k}(1-p)^{n-k} \leqslant \binom{n}{\ell}p^{\ell-1}(1-p)^{n-\ell+1},$$

and, for every integer k such that $1\leqslant k\leqslant n-1$,

$$\binom{n}{k}\big(p^{k}(1-p)^{n-k}-p^{n-k}(1-p)^{k}\big) \leqslant n\big(p(1-p)^{n-1}-p^{n-1}(1-p)\big). \qquad \square$$

Proof of Theorem 1. When $p=1/2$, we have perfect symmetry between the case $X=0$ and $X=1$, which means that both players can treat 0 and 1 similarly. In particular, if the detector uses the estimate $\hat{X}=1$ when the vector Z has k 1's, then it should use the estimate $\hat{X}=0$ when the vector Z has k 0's, which means that we must have

$$\bar{\mu}(k) = 1-\bar{\mu}(n-k).$$

Similarly, if the attacker decides to set to 0 a certain number of sensors when $X=0$, then it should set to 1 the same number of sensors when $X=1$, which means that we must have

$$\bar{\delta}_{\#0}(0) = m-\bar{\delta}_{\#0}(1).$$

In this case, (4) becomes

$$\begin{aligned}\mathrm{P}_{\mu,\delta}(\hat{X}\neq X) = \frac{1}{2}\Big(& p_{\text{attack}}\sum_{k=m-\bar{\delta}_{\#0}(0)}^{n-\bar{\delta}_{\#0}(0)}\bar{\mu}(k)\\ &\binom{n-m}{k-m+\bar{\delta}_{\#0}(0)}p_{\text{err}}^{k-m+\bar{\delta}_{\#0}(0)}(1-p_{\text{err}})^{n-k-\bar{\delta}_{\#0}(0)}\\ &+(1-p_{\text{attack}})\sum_{k=0}^{n}\bar{\mu}(k)\binom{n}{k}p_{\text{err}}^{k}(1-p_{\text{err}})^{n-k}\\ &+p_{\text{attack}}\sum_{k=\bar{\delta}_{\#0}(0)}^{n-m+\bar{\delta}_{\#0}(0)}\bar{\mu}(n-k)\binom{n-m}{k-\bar{\delta}_{\#0}(0)}\\ &(1-p_{\text{err}})^{k-\bar{\delta}_{\#0}(0)}p_{\text{err}}^{n-m-k+\bar{\delta}_{\#0}(0)}\end{aligned}$$

$$+ (1 - p_{\text{attack}}) \sum_{k=0}^{n} \bar{\mu}(n-k) \binom{n}{k} (1 - p_{\text{err}})^{k} p_{\text{err}}^{n-k} \Big).$$

Making the change of variable $n - k \to \ell$ in the third and fourth summation, we further obtain

$$\mathrm{P}_{\mu,\delta}(\hat{X} \neq X) = \frac{1}{2} \Big(p_{\text{attack}} \sum_{k=m-\bar{\delta}_{\#0}(0)}^{n-\bar{\delta}_{\#0}(0)} \bar{\mu}(k)$$

$$\binom{n-m}{k-m+\bar{\delta}_{\#0}(0)} p_{\text{err}}^{k-m+\bar{\delta}_{\#0}(0)} (1 - p_{\text{err}})^{n-k-\bar{\delta}_{\#0}(0)}$$

$$+ (1 - p_{\text{attack}}) \sum_{k=0}^{n} \bar{\mu}(k) \binom{n}{k} p_{\text{err}}^{k} (1 - p_{\text{err}})^{n-k}$$

$$+ p_{\text{attack}} \sum_{\ell=m-\bar{\delta}_{\#0}(0)}^{n-\bar{\delta}_{\#0}(0)} \bar{\mu}(l) \binom{n-m}{n-\ell-\bar{\delta}_{\#0}(0)}$$

$$p_{\text{err}}^{\ell-m+\bar{\delta}_{\#0}(0)} (1 - p_{\text{err}})^{n-\ell+\bar{\delta}_{\#0}(0)}$$

$$+ (1 - p_{\text{attack}}) \sum_{\ell=0}^{n} \bar{\mu}(\ell) \binom{n}{n-\ell} p_{\text{err}}^{\ell} (1 - p_{\text{err}})^{n-\ell} \Big)$$

$$= p_{\text{attack}} \sum_{k=m-\bar{\delta}_{\#0}(0)}^{n-\bar{\delta}_{\#0}(0)} \bar{\mu}(k) \binom{n-m}{k-m+\bar{\delta}_{\#0}(0)} p_{\text{err}}^{k-m+\bar{\delta}_{\#0}(0)}$$

$$(1 - p_{\text{err}})^{n-k-\bar{\delta}_{\#0}(0)}$$

$$+ (1 - p_{\text{attack}}) \sum_{k=0}^{n} \bar{\mu}(k) \binom{n}{k} p_{\text{err}}^{k} (1 - p_{\text{err}})^{n-k}.$$

Using the fact that n is odd, we can break the summations as follows

$$\mathrm{P}_{\mu,\delta}(\hat{X} \neq X) = p_{\text{attack}} \sum_{k=m-\bar{\delta}_{\#0}(0)}^{\frac{n-1}{2}} \bar{\mu}(k) \binom{n-m}{k-m+\bar{\delta}_{\#0}(0)}$$

$$p_{\text{err}}^{k-m+\bar{\delta}_{\#0}(0)} (1 - p_{\text{err}})^{n-k-\bar{\delta}_{\#0}(0)}$$

$$+ p_{\text{attack}} \sum_{k=\frac{n+1}{2}}^{n-\bar{\delta}_{\#0}(0)} \bar{\mu}(k) \binom{n-m}{k-m+\bar{\delta}_{\#0}(0)}$$

$$p_{\text{err}}^{k-m+\bar{\delta}_{\#0}(0)} (1 - p_{\text{err}})^{n-k-\bar{\delta}_{\#0}(0)}$$

$$+ (1 - p_{\text{attack}}) \sum_{k=0}^{\frac{n-1}{2}} \bar{\mu}(k) \binom{n}{k} p_{\text{err}}^{k} (1 - p_{\text{err}})^{n-k}$$

$$+ (1 - p_{\text{attack}}) \sum_{k=\frac{n+1}{2}}^{n} \bar{\mu}(k) \binom{n}{k}$$

$$p_{\text{err}}^{k} (1 - p_{\text{err}})^{n-k}$$

$$= p_{\text{attack}} \sum_{k=m-\bar{\delta}_{\#0}(0)}^{\frac{n-1}{2}} \bar{\mu}(k) \binom{n-m}{k-m+\bar{\delta}_{\#0}(0)}$$

$$p_{\text{err}}^{k-m+\bar{\delta}_{\#0}(0)} (1 - p_{\text{err}})^{n-k-\bar{\delta}_{\#0}(0)}$$

$$+ p_{\text{attack}} \sum_{k=\frac{n+1}{2}}^{n-\bar{\delta}_{\#0}(0)} \big(1 - \bar{\mu}(n-k)\big) \binom{n-m}{k-m+\bar{\delta}_{\#0}(0)}$$

$$p_{\text{err}}^{k-m+\bar{\delta}_{\#0}(0)} (1 - p_{\text{err}})^{n-k-\bar{\delta}_{\#0}(0)}$$

$$+ (1 - p_{\text{attack}}) \sum_{k=0}^{\frac{n-1}{2}} \bar{\mu}(k) \binom{n}{k} p_{\text{err}}^{k} (1 - p_{\text{err}})^{n-k}$$

$$+ (1 - p_{\text{attack}}) \sum_{k=\frac{n+1}{2}}^{n} \big(1 - \bar{\mu}(n-k)\big) \binom{n}{k}$$

$$p_{\text{err}}^{k} (1 - p_{\text{err}})^{n-k}$$

$$= p_{\text{attack}} \sum_{k=m-\bar{\delta}_{\#0}(0)}^{\frac{n-1}{2}} \bar{\mu}(k) \binom{n-m}{k-m+\bar{\delta}_{\#0}(0)}$$

$$p_{\text{err}}^{k-m+\bar{\delta}_{\#0}(0)} (1 - p_{\text{err}})^{n-k-\bar{\delta}_{\#0}(0)}$$

$$+ p_{\text{attack}} \sum_{\ell=\bar{\delta}_{\#0}(0)}^{\frac{n-1}{2}} \big(1 - \bar{\mu}(\ell)\big) \binom{n-m}{n-\ell-m+\bar{\delta}_{\#0}(0)}$$

$$p_{\text{err}}^{n-m-\ell+\bar{\delta}_{\#0}(0)} (1 - p_{\text{err}})^{\ell-\bar{\delta}_{\#0}(0)}$$

$$+ (1 - p_{\text{attack}}) \sum_{k=0}^{\frac{n-1}{2}} \bar{\mu}(k) \binom{n}{k} p_{\text{err}}^{k} (1 - p_{\text{err}})^{n-k}$$

$$+ (1 - p_{\text{attack}}) \sum_{\ell=0}^{\frac{n-1}{2}} \big(1 - \bar{\mu}(\ell)\big) \binom{n}{n-\ell} p_{\text{err}}^{n-\ell} (1 - p_{\text{err}})^{\ell}$$

$$= p_{\text{attack}} \sum_{k=m-\bar{\delta}_{\#0}(0)}^{\frac{n-1}{2}} \bar{\mu}(k) \binom{n-m}{k-m+\bar{\delta}_{\#0}(0)} p_{\text{err}}^{k-m+\bar{\delta}_{\#0}(0)}$$

$$(1 - p_{\text{err}})^{n-k-\bar{\delta}_{\#0}(0)}$$

$$+ p_{\text{attack}} \sum_{k=\bar{\delta}_{\#0}(0)}^{\frac{n-1}{2}} \binom{n-m}{n-m-k+\bar{\delta}_{\#0}(0)}$$

$$p_{\text{err}}^{n-m-k+\bar{\delta}_{\#0}(0)} (1 - p_{\text{err}})^{k-\bar{\delta}_{\#0}(0)}$$

$$- p_{\text{attack}} \sum_{k=\bar{\delta}_{\#0}(0)}^{\frac{n-1}{2}} \bar{\mu}(k) \binom{n-m}{n-m-k+\bar{\delta}_{\#0}(0)}$$

$$p_{\text{err}}^{n-m-k+\bar{\delta}_{\#0}(0)} (1 - p_{\text{err}})^{k-\bar{\delta}_{\#0}(0)}$$

$$+ (1 - p_{\text{attack}}) \sum_{k=0}^{\frac{n-1}{2}} \bar{\mu}(k) \binom{n}{k} \Big(p_{\text{err}}^{k} (1 - p_{\text{err}})^{n-k}$$

$$- p_{\text{err}}^{n-k} (1 - p_{\text{err}})^{k} \Big) + \alpha. \tag{10}$$

With this formula, we can proceed to exclude some of the estimator policies $\bar{\mu}$ based on policy domination. To achieve this, we compute the derivative of the probability of error with respect to the values of $\bar{\mu}(k)$. When this derivative is positive for every attacker policy δ, we know that we can restrict our attention to estimator policies for which $\bar{\mu}(k) = 0$ since the policies with $\bar{\mu}(k) = 1$ would be dominated. We recall that the estimator is the minimizer.

We consider separately three cases that differ by which summations in (10) include specific values of k:

1) For k such that $m - \bar{\delta}_{\#0}(0) \leqslant k < \bar{\delta}_{\#0}(0)$ and $1 \leqslant$

$k \leqslant \frac{n-1}{2}$, we have

$$\begin{aligned}\frac{d\,\mathrm{P}_{\mu,\delta}(\hat{X} \neq X)}{d\bar{\mu}(k)} &= p_{\mathrm{attack}}\binom{n-m}{k-m+\bar{\delta}_{\#0}(0)}\\ &\quad p_{\mathrm{err}}^{k-m+\bar{\delta}_{\#0}(0)}(1-p_{\mathrm{err}})^{n-k-\bar{\delta}_{\#0}(0)}\\ &\quad + (1-p_{\mathrm{attack}})\binom{n}{k}\Big(p_{\mathrm{err}}^{k}(1-p_{\mathrm{err}})^{n-k}\\ &\quad - p_{\mathrm{err}}^{n-k}(1-p_{\mathrm{err}})^{k}\Big) \geqslant 0,\end{aligned}$$

where the last inequality is a consequence of Proposition 1.

2) For k such that $\bar{\delta}_{\#0}(0) \leqslant k < m - \bar{\delta}_{\#0}(0)$ and $1 \leqslant k \leqslant \frac{n-1}{2}$, we have

$$\begin{aligned}\frac{d\,\mathrm{P}_{\mu,\delta}(\hat{X} \neq X)}{d\bar{\mu}(k)} &= -p_{\mathrm{attack}}\binom{n-m}{k-\bar{\delta}_{\#0}(0)}\\ &\quad p_{\mathrm{err}}^{n-m-k+\bar{\delta}_{\#0}(0)}(1-p_{\mathrm{err}})^{k-\bar{\delta}_{\#0}(0)}\\ &\quad + (1-p_{\mathrm{attack}})\binom{n}{k}\Big(p_{\mathrm{err}}^{k}(1-p_{\mathrm{err}})^{n-k}\\ &\quad - p_{\mathrm{err}}^{n-k}(1-p_{\mathrm{err}})^{k}\Big)\\ &\geqslant -p_{\mathrm{attack}}\binom{n-m}{\frac{n-1}{2}}p_{\mathrm{err}}^{n-m-\frac{n-1}{2}}(1-p_{\mathrm{err}})^{\frac{n-1}{2}}\\ &\quad + (1-p_{\mathrm{attack}})n\Big(p_{\mathrm{err}}(1-p_{\mathrm{err}})^{n-1}\\ &\quad - p_{\mathrm{err}}^{n-1}(1-p_{\mathrm{err}})\Big) \geqslant 0,\end{aligned}$$

where the second to last inequality is a consequence of Proposition 1 and the fact that, in this case, $k \leqslant m-1$, which because of (5)–(6) implies that $k - \bar{\delta}_{\#0}(0) \leqslant m-1 \leqslant \frac{n-1}{2} \leqslant (n-m)(1-p_{\mathrm{err}})$. The last inequality is then a consequence of (7).

3) For k such that $\max\{m-\bar{\delta}_{\#0}(0), \bar{\delta}_{\#0}(0)\} \leqslant k$ and $1 \leqslant k \leqslant \frac{n-1}{2}$, we have

$$\begin{aligned}\frac{d\,\mathrm{P}_{\mu,\delta}(\hat{X} \neq X)}{d\bar{\mu}(k)} &= p_{\mathrm{attack}}\Bigg(\binom{n-m}{k-m+\bar{\delta}_{\#0}(0)}\\ &\quad p_{\mathrm{err}}^{k-m+\bar{\delta}_{\#0}(0)}(1-p_{\mathrm{err}})^{n-k-\bar{\delta}_{\#0}(0)}\\ &\quad - \binom{n-m}{k-\bar{\delta}_{\#0}(0)}p_{\mathrm{err}}^{n-m-k+\bar{\delta}_{\#0}(0)}(1-p_{\mathrm{err}})^{k-\bar{\delta}_{\#0}(0)}\Bigg)\\ &\quad + (1-p_{\mathrm{attack}})\binom{n}{k}\Big(p_{\mathrm{err}}^{k}(1-p_{\mathrm{err}})^{n-k}\\ &\quad - p_{\mathrm{err}}^{n-k}(1-p_{\mathrm{err}})^{k}\Big)\\ &\geqslant -p_{\mathrm{attack}}\binom{n-m}{\frac{n-1}{2}}p_{\mathrm{err}}^{n-m-\frac{n-1}{2}}(1-p_{\mathrm{err}})^{\frac{n-1}{2}}\\ &\quad + (1-p_{\mathrm{attack}})n\Big(p_{\mathrm{err}}(1-p_{\mathrm{err}})^{n-1}\\ &\quad - p_{\mathrm{err}}^{n-1}(1-p_{\mathrm{err}})\Big) \geqslant 0,\end{aligned}$$

where the second to last inequality is a consequence of Proposition 1 and the fact that $k-\bar{\delta}_{\#0}(0) \leqslant \frac{n-1}{2} \leqslant (n-m)(1-p_{\mathrm{err}})$, because of (6). The last inequality is then a consequence of (7).

We then conclude that we only need to consider policies for which $\bar{\mu}(k) = 0$ for $1 \leqslant k \leqslant \frac{n-1}{2}$, so we are only left with two policies for the detector: the majority rule and no-consensus rule. For these pure policies, the probability of error (10) simplifies as follows: When $\bar{\delta}_{\#0}(0) = 0$ (which corresponds to the deception rule),

$$\mathrm{P}_{\mu,\delta}(\hat{X} \neq X) = \bar{\mu}(0)p_{\mathrm{attack}}(\beta - p_{\mathrm{err}}^{n-m}) + p_{\mathrm{attack}}\gamma + \alpha;$$

when $\bar{\delta}_{\#0}(0) = m$ (which corresponds to the no-deception rule),

$$\begin{aligned}&\mathrm{P}_{\mu,\delta}(\hat{X} \neq X)\\ &= \bar{\mu}(0)p_{\mathrm{attack}}\big(\beta + (1-p_{\mathrm{err}})^{n-m}\big) + p_{\mathrm{attack}}\rho + \alpha; \quad (11)\end{aligned}$$

and when $0 < \bar{\delta}_{\#0}(0) < m$,

$$\mathrm{P}_{\mu,\delta}(\hat{X} \neq X) = \bar{\mu}(0)p_{\mathrm{attack}}\beta + \alpha. \qquad (12)$$

Comparing (11) with (12), we conclude that the deception rule leads to a higher probability of error than the policy with $0 < \bar{\delta}_{\#0}(0) < m$, and therefore the former dominates the latter. We are thus left, we the following 2×2 zero-sum game where the first row corresponds to the majority rule ($\bar{\mu}(0) = 0$), the second row to the no-consensus rule ($\bar{\mu}(0) = 1$), the first column to the deception rule ($\bar{\delta}_{\#0}(0) = 0$), and the second column to the no-deception rule ($\bar{\delta}_{\#0}(0) = m$):

$$\begin{aligned}\bar{A} \coloneqq p_{\mathrm{attack}}&\begin{bmatrix}\gamma & \rho\\ \beta+\gamma-p_{\mathrm{err}}^{n-m} & \beta+\rho+(1-p_{\mathrm{err}})^{n-m}\end{bmatrix}\\ &+ \alpha\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}.\end{aligned}$$

It is straightforward to show that this matrix has a mixed saddle-point

$$y^* \coloneqq \begin{bmatrix}1-y_2\\ y_2\end{bmatrix}, \qquad z^* = \begin{bmatrix}1-z_2\\ z_2\end{bmatrix},$$

with value v^*, from which the result follows. ■

V. Conclusions

We propose a new game approach to estimate a binary random variable based on a vector of sensor measurements that may have been corrupted by an attacker. The problem is formulated as a zero-sum partial information game in which a detector attempts to minimize the probability of error and an attacker attempts to maximize this probability. We are currently extending these results to dynamic estimation problems.

Appendix

Proof of Lemma 1. By the law of total probability, we can expand

$$\begin{aligned}\mathrm{P}_{\mu,\delta}(\hat{X} \neq X) = \sum_{k=0}^{n}\mathrm{P}_{\mu\delta}\left(\hat{X} = 1 | X = 0, \sum_{i=1}^{n} Z_i = k\right)\\ \mathrm{P}_{\mu\delta}\left(\sum_{i=1}^{n} Z_i = k | X = 0\right)\mathrm{P}_{\mu\delta}(X = 0)\end{aligned}$$

$$+\sum_{k=0}^{n} \mathrm{P}_{\mu\delta}\Big(\hat{X} = 0|X=1,\sum_{i=1}^{n} Z_i = k\Big)$$

$$\mathrm{P}_{\mu\delta}\Big(\sum_{i=1}^{n} Z_i = k|X=1\Big)\,\mathrm{P}_{\mu\delta}(X=1)$$

$$\mathrm{P}_{\mu,\delta}(\hat{X} \neq X) = (1-p)\sum_{k=0}^{n} \bar{\mu}(k)\,\mathrm{P}_{\delta}\Big(\sum_{i=1}^{n} Z_i = k|X=0\Big)$$

$$+\,p\sum_{k=0}^{n}\big(1-\bar{\mu}(k)\big)\,\mathrm{P}_{\delta}\Big(\sum_{i=1}^{n} Z_i = k|X=1\Big), \tag{13}$$

where we used the facts that

$$\mathrm{P}_{\mu\delta}\Big(\hat{X} = 1|X=0,\sum_{i=1}^{n} Z_i = k\Big) = \begin{cases} 1 & \bar{\mu}(k)=1 \\ 0 & \bar{\mu}(k)=0 \end{cases} = \bar{\mu}(k),$$

$$\mathrm{P}_{\mu\delta}\Big(\hat{X} = 0|X=0,\sum_{i=1}^{n} Z_i = k\Big) = \begin{cases} 1 & \bar{\mu}(k)=0 \\ 0 & \bar{\mu}(k)=1 \end{cases} = 1-\bar{\mu}(k).$$

We now proceed to compute the conditional probabilities in the formula above, that can also be expanded as follows:

$$\mathrm{P}_{\delta}\Big(\sum_{i=1}^{n} Z_i = k|X\Big) = \mathrm{P}_{\delta}\Big(\sum_{i=1}^{n} Z_i = k|X,\mathcal{E}_{\text{attack}}\Big)p_{\text{attack}}$$

$$+\,\mathrm{P}_{\delta}\Big(\sum_{i=1}^{n} Z_i = k|X,\neg\mathcal{E}_{\text{attack}}\Big)(1-p_{\text{attack}}),$$

where $\mathcal{E}_{\text{attack}}$ denotes the events that the attacker manipulated measurements and $\neg\mathcal{E}_{\text{attack}}$ the complementary event. When no measurements have been manipulated, we simply have that

$$\mathrm{P}_{\delta}\Big(\sum_{i=1}^{n} Z_i = k|X,\neg\mathcal{E}_{\text{attack}}\Big) = \binom{n}{k}\begin{cases} p_{\text{err}}^{k}(1-p_{\text{err}})^{n-k} & X=0, \\ (1-p_{\text{err}})^{k}p_{\text{err}}^{n-k} & X=1. \end{cases}$$

Otherwise, since $\delta(X)$ sets to 0 and to 1 a number of sensors equal to $\bar{\delta}_{\#0}(X)$ and to $m-\bar{\delta}_{\#0}(X)$, respectively, we have that

$$\mathrm{P}_{\delta}\Big(\sum_{i=1}^{n} Z_i = k|X,\mathcal{E}_{\text{attack}}\Big) = \begin{cases} f_1, & m-\bar{\delta}_{\#0}(0) \leqslant k \leqslant n-\bar{\delta}_{\#0}(0), X=0 \\ f_2, & m-\bar{\delta}_{\#0}(1) \leqslant k \leqslant n-\bar{\delta}_{\#0}(1), X=1 \\ 0, & \text{otherwise} \end{cases}$$

where

$$f_1 = \binom{n-m}{k-m+\bar{\delta}_{\#0}(0)} p_{\text{err}}^{k-m+\bar{\delta}_{\#0}(0)}(1-p_{\text{err}})^{n-k-\bar{\delta}_{\#0}(0)}$$

and

$$f_2 = \binom{n-m}{k-m+\bar{\delta}_{\#0}(1)}(1-p_{\text{err}})^{k-m+\bar{\delta}_{\#0}(1)}p_{\text{err}}^{n-k-\bar{\delta}_{\#0}(1)}.$$

Therefore

$$\mathrm{P}_{\delta}\Big(\sum_{i=1}^{n} Z_i = k|X\Big) = p_{\text{attack}}\begin{cases} f_1, & m-\bar{\delta}_{\#0}(0) \leqslant k \leqslant n-\bar{\delta}_{\#0}(0), X=0 \\ f_2, & m-\bar{\delta}_{\#0}(1) \leqslant k \leqslant n-\bar{\delta}_{\#0}(1), X=1 \\ 0, & \text{otherwise.} \end{cases}$$

$$+\,(1-p_{\text{attack}})\binom{n}{k}\begin{cases} p_{\text{err}}^{k}(1-p_{\text{err}})^{n-k} & X=0, \\ (1-p_{\text{err}})^{k}p_{\text{err}}^{n-k} & X=1. \end{cases}$$

The result follows from this and (13).

References

[1] A. Agah, S. K. Das, K. Basu, and M. Asadi, "Intrusion detection in sensor networks: A noncooperative game approach," in *Proc. of the 3rd Intl. Symp. on Network Computing and Applications (NCA)*, Washington, DC, USA: IEEE Computer Society, pp. 343-346, 2004.

[2] T. Alpcan and T. Basar, "A game theoretic approach to decision and analysis in network intrusion detection," in *Proc. of the 42nd IEEE Conf. on Decision and Control (CDC)*, Maui, HI, pp. 2595-2600, 2003.

[3] T. Alpcan, and T. Basar, "A Game Theoretic Analysis of Intrusion Detection in Access Control Systems," in *Proc. 43rd IEEE Conference on Decision and Control (CDC)*, December, vol. 2, pp.1568-1573.

[4] T. Basar, "Optimum performance levels for H_∞ filters, predictors, and smoothers," *Syst. Contr. Lett.*, vol. 16, pp. 309317, 1991.

[5] T. Basar and G. J. Olsder, *Dynamic Noncooperative Game Theory*, 2nd ed. Philadelphia, PA:SIAM, 1999.

[6] E. Byres and J. Lowe, "The myths and facts behind cyber security risks for industrial control systems," *VDE Congress*, 2004.

[7] M. Kodialam, and T.V. Lakshman, "Detecting Network Intrusions via Sampling: A Game Theoretic Approach," *IEEE INFOCOM*, March-April, Vol. 3, pp.1880-1889, 2003.

[8] K.-w. Lye and J. M. Wing, "Strategies in Network Security," *International Journal of Information Security*, Vol. 4, pp. 71-86, 2005.

[9] V. Srinivasan, V. Nuggehalli, C-F. Chiasserini, and R. R. Rao, "An Analytical Approach to the Study of Cooperation in Wireless Ad Hoc Networks,"in *IEEE Transactions on Communications*, March, vol. 4, No. 2, pp.722-733, 2005.

[10] S. Tijs, *Introduction to Game Theory*, Hindustan Book Agency, India, 2003.

[11] M. Tubaishat and S. Madria, "Sensor Networks: an Overview," *IEEE Potentials*, vol. 22, no. 2, 20-23, April 2003.

[12] A. Urpi, M. Bonuccelli, and S. Giordano, "Modelling Cooperation in Mobile Ad Hoc Networks: A formal description of selfishness," *WiOpt'03 Workshop: Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, 2003.