Distributed Detection in the Presence of Byzantine Attacks

Stefano Marano, Vincenzo Matta, and Lang Tong, *Fellow, IEEE*

***Abstract*—Distributed detection in the presence of cooperative (Byzantine) attack is considered. It is assumed that a fraction of the monitoring sensors are compromised by an adversary, and these compromised (Byzantine) sensors are reprogrammed to transmit fictitious observations aimed at confusing the decision maker at the fusion center. For detection under binary hypotheses with quantized sensor observations, the optimal attacking distributions for Byzantine sensors that minimize the detection error exponent are obtained using a "water-filling" procedure. The smallest error exponent, as a function of the Byzantine sensor population, characterizes the power of attack. Also obtained is the minimum fraction of Byzantine sensors that destroys the consistency of detection at the fusion center. The case when multiple measurements are made at the remote nodes is also considered, and it is shown that the detection performance scales with the number of sensors differently from the number of observations at each sensor.**

***Index Terms*—Byzantine attack, distributed detection, network defense.**

I. Introduction

We consider the classical problem of distributed detection but under the assumption that some of the sensors have been compromised by an intruder. The compromised sensors are referred to as Byzantine and they can be reprogrammed by the intruder to attack the fusion center by transmitting fictitious observations. The rest of the sensors are referred to as honest, and they follow the expected rule of operation.

In the context of distributed detection, the Byzantine sensor problem is motivated by applications of envisioned wireless sensor networks where sensors are more vulnerable to tempering. In particular, wireless sensors may be made of low cost devices with severe constraints on battery power. Such practical limitations may make the use of sophisticated encryption unrealistic. Furthermore, the wireless transmission medium is more vulnerable to eavesdropping, which makes it possible for the attacker to extract information from sensor transmissions. As a result, the adversary can employ a wide range of strategies including deploying its own sensors aimed at jamming the transmission of honest sensors or, in a more sophisticated way, transmitting optimally designed signals to confuse the fusion center.

We are interested in an analytical characterization of the ability that Byzantine sensors can affect the decision at the fusion center. Specifically, from the intruder's perspective, what is the most effective attacking strategy by the Byzantine sensors? It seems obvious that if too many sensors are compromised, the fusion center will loose its ability to detect the underlying phenomenon. But what is the minimum population size of the Byzantine sensors such that the fusion network is rendered ineffective completely? From the decision maker's perspective, given the Byzantine sensor population, or an upper bound thereof, what is the achievable performance not knowing which sensor is compromised?

We adopt a standard model in distributed detection under binary hypotheses $\mathcal{H}_0$ versus $\mathcal{H}_1$, with known distributions [1]. All sensors draw observations that are independent and identically distributed (i.i.d.) conditioned on the unknown hypothesis. The classical assumption of conditional i.i.d. observations may not always be valid in practice and the complications of correlated observations are well known [2]. Recognizing its limitations, we make the conditional i.i.d. assumption for analytical tractability and gaining insights into how Byzantine sensors can affect the overall performance.

For the Byzantine sensors, we adopt an approach that grants the intruder with more power than usually allowed in practice, which leads to a conservative assessment of security risk but gains in analytical tractability. To this end, we assume that Byzantine sensors in fact know the true hypothesis and they use this knowledge to construct the most effective fictitious observations aimed at confusing the fusion center. This assumption obviously is difficult (but not impossible) to satisfy in practice; it would require that the attacker has a separate network that allows Byzantine sensors to cooperate among themselves.[1]

For the fusion center, we assume that it is not compromised, and it is able to collect data from n sensors.[2] We are not concerned in this paper about how individual sensors deliver their data to the fusion center except that what the fusion center receives is what transmitted by the sensors (Byzantine or honest). This simplifying assumption also has practical implications:

Manuscript received April 05, 2007; revised August 09, 2008. First published October 31, 2008; current version published January 06, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chong-Meng Samson See. A portion of this work was presented at MILCOM2006 and at ASILOMAR 2006. The work of L. Tong was supported in part by the National Science Foundation under Contract CCF-0635070.

S. Marano and V. Matta are with the Department of Information and Electrical Engineering (DIIIE), University of Salerno, Fisciano (SA) 84084, Italy (e-mail: marano@unisa.it; vmatta@unisa.it).

L. Tong is with the Electrical and Computer Engineering Department, Cornell University, Ithaca, NY 14853 USA (e-mail: ltong@ece.cornell.edu).

Digital Object Identifier 10.1109/TSP.2008.2007335

[1]For example, the compromised sensors may send their observations to the intruder and the intruder detects the underlying phenomenon and inform the Byzantine sensors so that they can collaboratively attack the fusion center.

[2]It is possible that more than n sensors have transmitted and some transmissions are not successful.

transmissions of sensors may need to be protected by error control mechanisms, and the Byzantine sensors are not able to alter the transmissions of honest sensors.

We assume that the fusion center does not know which sensor is Byzantine, but it knows the average percentage of compromised sensors, or at least an upper bound, and that the Byzantine sensors may create fictitious samples according to some unknown (possibly optimized) distribution. The fusion center makes the detection under a variant of Neyman-Pearson (NP) setup, which the adversary knows. Different from the standard NP problem, not knowing what distribution the attacker adopts, the fusion center caps the false alarm probability to a for all possible attacking distributions. Minimizing miss detection probability for all possible attacking strategies is not possible. A reasonable approach is to minimize the worst miss detection probability, which guarantees that the miss detection probability will not exceed that advertised the worst case, no matter which distribution is used by the Byzantine sensors.

A. Summary of Results

We first consider the case when the fusion center receives one observation from each sensor, and no more than an average fraction α of the received samples are from the Byzantine sensors. The parameter α represents the power of the adversary of affecting the detection performance.

Formulating the problem as one of minimizing Kullback-Leibler (KL) divergence, we obtain the optimal attacking distribution by the Byzantine sensors through a "water-filling" procedure. In the NP setup, the resulting KL divergence $\Delta(\alpha)$—a function of α—represents the worst rate of exponential decay (the error exponent) of the miss detection probability. When $\Delta(\alpha) = 0$, however, the fusion is "blinded" by the Byzantine sensors, loosing its ability to distinguish the two hypotheses even as the number of samples, we denote it by n, goes to infinite. We show that $\Delta(\alpha)$ is a decreasing convex function, reaching zero at the "blinding" attacking power $\alpha_b \leq 1/2$.

We then consider the case of multiple observations per sensor. Differently form the single-sample case, now the blinding power α_b is always 1/2, regardless of the initial data distributions. This implies that an intruder owing less than 50% of the nodes can never completely blind the system. One somewhat surprising result here is that the performances of the network under attack do not scale with m, the sample size at each sensor.

B. Related Work

Distributed detection is a classical subject in signal processing (see [3]–[5]) and has attracted recent interest due to the potential deployment of wireless sensors for a variety of applications from environmental monitoring to military surveillance. See [6] for a survey of recent sensor network research from signal processing and communications perspectives. While there is a vast literature on secure networking for general ad hoc and sensor networks, see, e.g., [7] and references therein, reported work on distributed detection and data fusion in the presence of Byzantine sensors is still limited, see [8]–[10]. Relevant to the application considered in this paper is the witness-based approach proposed by Du *et al.* [11] where the fusion center and a set of witnesses jointly authenticate the fusion data by the use of the *Message Authentication Code*. Our focus is considerably different in that we do not try to authenticate the data; we consider most effective attacking strategies and distributed detection schemes that are robust to attack.

The Byzantine model assumed in this paper was originally proposed by Lamport, Shostak and Pease [12] and further developed by Dolev [13] and later in the information theoretic context by Pfitzmann and Waidner [14]. Byzantine models have also been used in recent work on network security, see [15]. Here we focus on the impact of Byzantine nodes on distributed detection, which has not been considered in the past. In someway, our problem in the presence of compromised sensors is similar to the original Byzantine general problem in the sense that a set of sensors try to interfere the fusion center to reach reliable detection, and the compromised sensors, like the traitorous general, are given full options (including collaboration) to disrupt the sensor network. A key difference is the presence of the fusion center (which is always honest).

An information theoretic investigation of data fusion in the presence of Byzantine sensors is considered in [16]. The authors of [16], however, are interested not in the detection performance but in recovering measurements from honest sensors at the fusion center.

The signal processing problem considered in this paper is most relevant to robust statistical inference [17]. In his seminal work [18], Huber considered the problem of binary hypothesis testing with α-contaminated distributions. The Byzantine sensor model used in this paper fits naturally into Huber's robust detection framework, and results in classical robust detection apply to the Byzantine sensor problem. In particular, Huber showed that the likelihood ratio test based on the worst distribution pair has the minimax property. It minimizes the maximum miss detection error probability (among all possible α-contaminated distributions) while all the false alarm probabilities are below a preset bound a.

Our results, however, are not a direct application of those of Huber. We have used the miss detection error exponent as our performance metric in our analysis. While as Huber we too are interested in the worst distribution pair, our techniques of finding them are different. Our technique leads to a "water-filling" solution whereas Huber's technique is algebraic. Indeed, finding the worst distribution pair is only the first step toward characterizing the power of the Byzantine attack. For example, our result allows us to obtain the relation between the size of the Byzantine sensor population and the worst detection error exponent. We have also investigated the effects of multiple sensor measurements and the scaling behavior, which are not considered in classical robust detection.

The paper is organized as follows. In Section II the addressed problem is formalized. Sections III and IV contain the main results of the paper, respectively for single and multiple observations. These results are presented in the form of two theorems whose implications are discussed in due depth; the proofs of the theorems are provided in two appendices at the end of the paper. Numerical examples are given in Section V, while Section VI contains conclusions and hints for future works.

II. Problem Formulation

In the following $K^{(j)},\ j=1,2,\ldots,n$, denotes the observation made at sensor j. We denote p, q, x, y, w and z as probability distributions defined over a discrete finite common alphabet that, without loss of generality, can be taken as $\mathcal{K}=\{0,1,2,\ldots,|\mathcal{K}|-1\}$. The entries of the probability vector p will be denoted by $p_k,\ k\in\mathcal{K}$, and the same definition applies to other probability vectors. $D(z\|w)$ denotes the divergence or Kullback-Leibler (KL) distance [19] between the two probability mass functions (pmfs) z and w. Logarithms are to base e and divergences are measured in nats. It is assumed that $0\log(0/w_k)=0,\ \forall w_k$.

A. Models and Assumptions

We use the conventional distributed detection model where we consider two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$. We make the conditional i.i.d. assumption under which observations from honest sensors are conditionally independent and identically distributed. If sensor j is honest, it observes $K^{(j)}$ according to distributions p and q under $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively

$$\begin{aligned}\mathcal{H}_0 &: \Pr\{K^{(j)}=k|\mathcal{H}_0\}=q_k\\ \mathcal{H}_1 &: \Pr\{K^{(j)}=k|\mathcal{H}_1\}=p_k\end{aligned}\tag{1}$$

where $k=0,\ 1,\ldots,|\mathcal{K}|-1$, and, for the time being, $K^{(j)}$ is a scalar.

For Byzantine sensors, we assume that, through collaborations, they know the true hypothesis. This allows the Byzantine sensors to transmit fictitious observations to the fusion center according to distributions different from that given by nature. In particular, if a sensor j is Byzantine, it generates observation $K^{(j)}$ as follows:

$$\begin{aligned}\mathcal{H}_0 &: \Pr\{K^{(j)}=k|\mathcal{H}_0\}=y_k\\ \mathcal{H}_1 &: \Pr\{K^{(j)}=k|\mathcal{H}_1\}=x_k\end{aligned}\tag{2}$$

where y and x are two pmfs defined over $\mathcal{K}$. We assume that all Byzantine sensors use the same distribution and transmit fake observations independently. Note that this problem formulation can be used also when the Byzantine bases its emission on the original observation owned by the sensor. That is, if under $\mathcal{H}_0$ the original observation taken from nature by sensor j is $\tilde{K}^{(j)}=i$, the Byzantine may transmit to the fusion center a value $K^{(j)}$ using the conditional pmf $\Pr\{K^{(j)}=k|\tilde{K}^{(j)}=i,\mathcal{H}_0\}$. However, since the fusion center does not have access to the true $\tilde{K}^{(j)}$, the distribution seen by the fusion center should be accordingly averaged, such that y_k can be defined as

$$\Pr\{K^{(j)}=k|\mathcal{H}_0\}=\sum_{i=0}^{|\mathcal{K}|-1}\Pr\{K^{(j)}=k|\tilde{K}^{(j)}=i,\mathcal{H}_0\}\,q_i.$$

The same holds true under the alternative hypothesis.

We assume that the fusion center receives n observations from sensors and with probability α is Byzantine. The jth sample then has the distribution

$$\begin{aligned}\mathcal{H}_0 &: \Pr\{K^{(j)}=k|\mathcal{H}_0\}=z_k,\ \text{with } z=(1-\alpha)q+\alpha y\\ \mathcal{H}_1 &: \Pr\{K^{(j)}=k|\mathcal{H}_1\}=w_k,\ \text{with } w=(1-\alpha)p+\alpha x.\end{aligned}\tag{3}$$

Again the n observations are conditional i.i.d.

B. Problem Statement

In characterizing the power of a Byzantine attack, we first take the perspective of the intruder and aim at degrading as much as possible the detection performance at the fusion center. By Stein's lemma [19], we know that the KL divergence $D(z\|w)$ represents the best error exponent of the missed detection error probability in the NP setup. In other words, for any detector of size[3] a, the miss detection error probability $P_M^{(n)}$ has the asymptotic representation

$$P_M^{(n)}=e^{-n(D(z\|w)+\epsilon(n))}$$

where n is the number of samples used in the detection. Thus it is natural that the intruder should minimize $D(z\|w)$ by choosing attacking distributions x and y optimally.

To highlight the dependencies of $D(z\|w)$ on x and y, define

$$\begin{aligned}d(y;x)&:=D(z\|w)=D((1-\alpha)q+\alpha y\|(1-\alpha)p+\alpha x)\\&=\sum_{k=0}^{|\mathcal{K}|-1}[(1-\alpha)q_k+\alpha y_k]\log\frac{(1-\alpha)q_k+\alpha y_k}{(1-\alpha)p_k+\alpha x_k}\end{aligned}\tag{4}$$

which is not to be confused with the divergence between y and x. From now on, unless otherwise specified, sums such as $\sum x_k$ run over the entries of the alphabet $\mathcal{K}$.

Also, let

$$\Delta(\alpha):=\min_{x,y}d(y;x)\tag{5}$$

be the minimum of the divergence between the two hypotheses with distributions w and z. We refer α as the attacking power, $\Delta(\alpha)$ the *attacking exponent*, and the optimizing (x,y) as the *attacking distributions*. In Theorem 1, we provide the analytical characterization of these quantities.

The KL divergence has the property that $D(z\|w)=0$ if and only if $z=w$. Therefore, if $\Delta(\alpha)=0$, the fusion center is unable to distinguish the two original hypotheses p and q. The intruder then is interested in the minimum attacking power that will destroy the ability of fusion center to detect. Thus, we call

$$\alpha_b=\min\{\alpha\ \text{such that}\ \Delta(\alpha)=0\}\tag{6}$$

the *blinding power*. A closed form expression can also be found in Theorem 1.

It should be obvious that $\Delta(\alpha)=0$ for any $\alpha>1/2$. In such a case, the intruder simply creates the parity by letting a fraction of the Byzantine sensors transmitting samples generated from the true hypothesis and the rest from the alternative. When $\alpha>$

[3]That is, for any detector whose false alarm probability does not exceed a, see, e.g., [19].

$1/2$, the fusion center sees with equal probability observations generated under $\mathcal{H}_0$ and under $\mathcal{H}_1$.

Next we take the perspective of the fusion center which knows that the distributions of the received samples follow (3), but it does not know the specific attacking distributions used by the Byzantine sensors. The specific forms of distributions in (3) allows us to use the framework of robust detection. In [18], Huber considered the hypothesis testing problem given in (3) and provide a minimax solution. He showed that there exists a pair of distributions such that a likelihood ratio detector will i) guarantee that false alarm probabilities under all possible z are below a preset bound a; ii) minimize the maximum miss detection probability (among all possible w).

It is not obvious that worst distribution pair from Huber's robust detection theory matches that obtained under (5). This turns out to be the case as shown in Theorem 1, and implies that if the fusion center adopts Huber's robust detector designed over the upper bound, say $\bar{\alpha}$, a miss detection error exponent no smaller than $\Delta(\bar{\alpha})$ is achievable. This coincidence gives the operational meaning of (5).

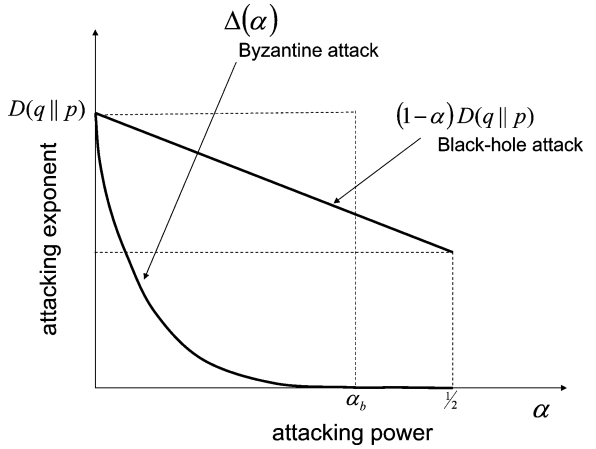

Fig. 1. The fundamental tradeoff between the attacking exponent of the network $\Delta(\alpha)$ and the attacking power of the intruder α. For $\alpha = 0$ the intruder does not affect the network. By increasing the fraction of infected nodes α, the intruder becomes more pervasive and the detection capabilities of the network are accordingly reduced. At $\alpha = \alpha_b$ the intruder blinds the system, and $\Delta(\alpha) = 0$. Also shown is the exponent of a black-hole attack.

III. Attacking Distributions and Detection Error Exponents

Now we describe how the intruder can minimize the test exponent, and discuss the cases in which this minimum is zero, meaning that the system is definitely impaired and useless.

Theorem 1:

i) The blinding power α_b defined in (6) has the form

$$\alpha_b := \frac{\sum(q_k - p_k)^+}{1 + \sum(q_k - p_k)^+} \leq \frac{1}{2} \tag{7}$$

where $(c)^+ = c$ when $c \geq 0$ and $(c)^+ = 0$ otherwise.

ii) If $\alpha = \alpha_b$, then the unique pair of distributions (x, y) that nullifies the attacking exponent is, $\forall k \in \mathcal{K}$,

$$\begin{cases} x_k = \frac{1-\alpha}{\alpha}(q_k - p_k)^+ \\ y_k = \frac{1-\alpha}{\alpha}(p_k - q_k)^+. \end{cases} \tag{8}$$

If $\alpha > \alpha_b$, there exist infinitely many solutions (x, y) that nullifies the test exponent.

iii) If $\alpha < \alpha_b$, $\Delta(\alpha) > 0$ and the unique pair of pmfs (x, y) that attains such minimum divergence is given by, $\forall k \in \mathcal{K}$,

$$\begin{cases} x_k = \frac{1-\alpha}{\alpha}\ (\gamma_x q_k - p_k)^+, \\ y_k = \frac{1-\alpha}{\alpha}\ (\gamma_y p_k - q_k)^+, \end{cases} \tag{9}$$

where $0 < \gamma_x,\ \gamma_y \leq 1$ are constants to be set in order to fulfill $\sum x_k = \sum y_k = 1$.

iv) The function $\Delta(\alpha)$ is continuous, decreasing, and convex $\cup$ over the interval $\alpha \in (0, \alpha_b)$, with $\Delta(0) = D(q\|p) > 0$ and $\lim_{\alpha \longrightarrow \alpha_b} \Delta(\alpha) = 0$. ■

Proof: See Appendix I. •

As a first comment we note that, according to part iv) of the theorem, the typical curve $\Delta(\alpha)$ is that depicted in Fig. 1, and it represents the best achievable asymptotic rate of the test, accounting for the malicious intruder that plagues the system. Also shown is the performance curve for the black-hole attack, amounting at simply destroying a fraction α of nodes. This yields an exponent $(1 - \alpha)D(q\|p)$, which is always less effective than the more sophisticated attack dealt with in Theorem 1, due to convexity of the divergence.

Solution (8) in part ii) is actually a special case of that in (9), obtained with $\gamma_x = \gamma_y = 1$, while the case $\alpha > \alpha_b$ leaves to the intruder additional degrees of freedom in choosing the attacking distributions. As to statement iii), it is straightforward to see that the likelihood ratio between w and z, with x and y given by (9), amounts to the product of terms in the form

$$\frac{w_k}{z_k} = \begin{cases} \frac{1}{\gamma_y} & \text{if } \frac{p_k}{q_k} \geq \frac{1}{\gamma_y} \\ \frac{p_k}{q_k} & \text{if } \gamma_x < \frac{p_k}{q_k} < \frac{1}{\gamma_y} \\ \gamma_x & \text{if } \frac{p_k}{q_k} \leq \gamma_x \end{cases} \tag{10}$$

that is a censored version of the original test between p and q. Such kind of tests naturally arises in the context of robust detection, see Huber [18]. In fact, statement iii) of the theorem can be found in the literature of robust inference [17], [18], [20], once that one recognizes w and z as being α-contaminated mixtures. Under a variety of test performance criteria (Bayes risk, NP, Chernoff exponent, etc.,) (9) are in fact known to yield the least favorable distributions.[4] For self-consistency, in the appendix we opt for providing a simple proof of statement iii), tailored to our special case of KL criterion and finite alphabets.

The analogy with robust theory also implies further properties. First, a Byzantine change of the attacking distributions when the network is not aware of this [i.e., when it still implements the censored test (10)], cannot provide any advantage from the intruder's viewpoint. In addition, would only an upper bound $\bar{\alpha}$ be available to the fusion center, the censoring thresholds γ_x and γ_y in test (10) should be simply computed using $\bar{\alpha}$. For any $\alpha \leq \bar{\alpha}$ the ensemble of attacking distributions x and y available to the Byzantines spans the α-contaminated class which, more or less obviously, is contained in the $\bar{\alpha}$-contaminated. This implies that the censored test designed with $\bar{\alpha}$ achieves an exponent no smaller than $\Delta(\bar{\alpha})$.

[4]Similarly, we can argue that some results in Theorem 1 can be easily extended to more general settings, such as continuous observations, different performance criteria, etc.

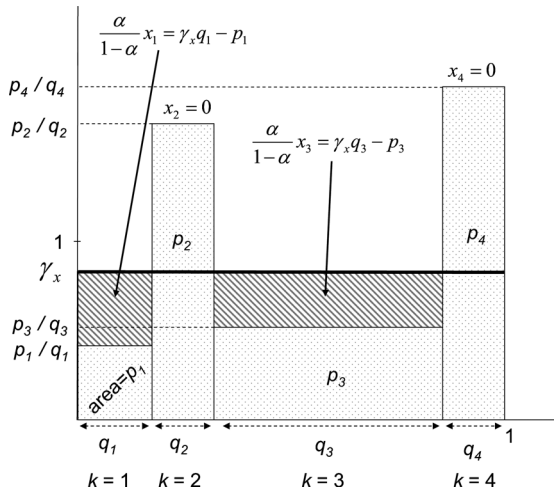

Fig. 2. Description of the "water-filling" procedure. For each k, we draw rectangular boxes of area p_k (bases q_k, and heights p_k/q_k). When γ_x increases from its initial value of 0, some dashed boxes begin to appear in correspondence of the indexes where p_k/q_k is smaller. The process of increasing γ_x is stopped when the total area of these dashed regions is equal to $\alpha/(1-\alpha)$. This is the level shown by the bold line.

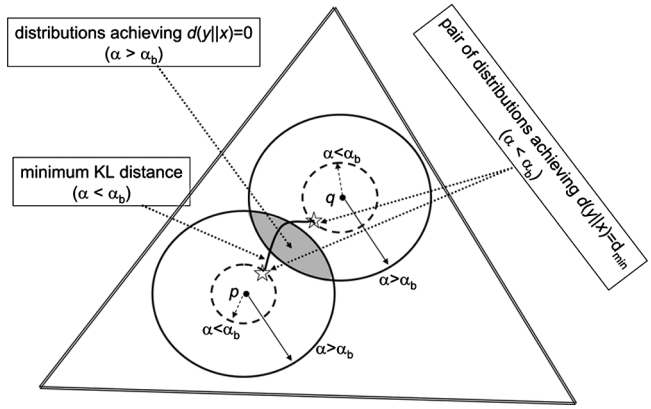

Fig. 3. Probability simplex and geometrical interpretation of the sets $\mathcal{S}_\alpha^p$ and $\mathcal{S}_\alpha^q$. The circles around p and q denote $\mathcal{S}_\alpha^p$ and $\mathcal{S}_\alpha^q$, respectively.

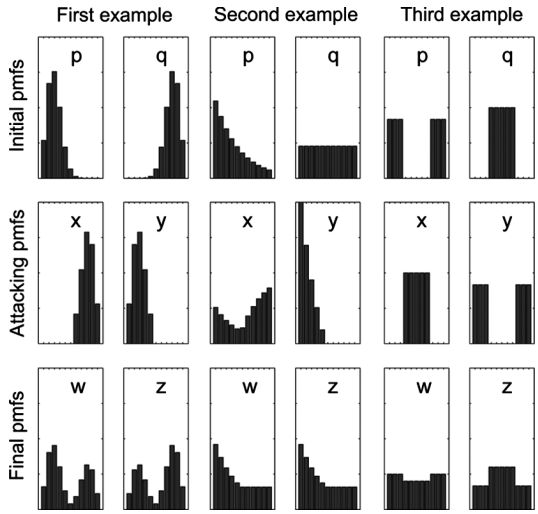

Fig. 4. Three examples of application of the theory. See the main text for details.

The values of γ_x and γ_y appearing in Theorem 1 result from a procedure that can be described in terms of a "water-filling," see Fig. 2. For each k we draw a rectangle of area p_k, whose basis and height are q_k and p_k/q_k, respectively. Then, we start to increase the *level of water* γ_x from its initial value of 0. When γ_x reaches the height of the shortest rectangle (in the example the leftmost rectangle whose height is p_1/q_1) the *water* begins filling the correspondent dashed region. By further increasing, γ_x passes the level p_3/q_3 so that the water simultaneously fills the dashed regions in correspondence of $k = 1$ and $k = 3$. The process of increasing γ_x is stopped when the total area of these dashed areas is equal to $\alpha/(1-\alpha)$. The final value of γ_x is depicted by a bold line in the figure and represents the sought value of the constant appearing in the first equation of (9). In fact, with such a value of γ_x we have found the desired pmf: $x_1 = (\gamma_x q_1 - p_1)(1-\alpha)/\alpha$, $x_3 = (\gamma_x q_3 - p_3)(1-\alpha)/\alpha$, $x_2 = x_4 = 0$. Note that, as prescribed by Theorem 1, only rectangles such that $p_k < q_k$ can be (but not necessarily are) filled, because $\gamma_x \leq 1$. Thus, only the correspondent vector entries can be positive; in the example this happens for x_1 and x_3. Obviously, similar arguments apply to the computation of γ_y. Note also that x_k and y_k are never both positive for a given k, as one can expect.

It is instructive to regard results in Theorem 1 from a geometrical viewpoint. Let $\mathcal{S}_\alpha^p = \{w : w = (1-\alpha)p + \alpha x\}$, for some pmf x over $\mathcal{K}$, and let $\mathcal{S}_\alpha^q = \{z : z = (1-\alpha)q + \alpha y\}$, for some pmf y over $\mathcal{K}$. With reference to Fig. 3, these sets admit the following geometric interpretation. In the regime $\alpha > \alpha_b$, the intruder has enough degrees of freedom to have the chance of selecting one of the infinitely many pairs (x, y) such that $w = z \in \mathcal{S}_\alpha^p \cap \mathcal{S}_\alpha^q$, so that $d(y;x) = 0$. Conversely, whenever $\alpha < \alpha_b$, the two sets in Fig. 3 are disjoint and the divergence cannot be nullified: a single solution (x, y) is available to the intruder. Such solution gives the pair (w, z) that achieves $d(y;x) = \Delta(\alpha)$. If $\alpha = \alpha_b$ the two circumferences in Fig. 3 would be tangent, the intersection $\mathcal{S}_\alpha^p \cap \mathcal{S}_\alpha^q$ would contain the single element $w = z$, and clearly $\Delta(\alpha)$ is zero.[5]

Before concluding this section, let us give some examples of application of the above theory. Assume that we are given some prescribed initial distributions p and q (possible states of the nature) and the intruder's power α. Then, the fictitious distributions for the intruder's attack x and y can be computed as described in Theorem 1, and these pmfs minimize the final divergence between the distributions w and z, as perceived by the FC. The situation is illustrated in Fig. 4, where three somehow peculiar situations are illustrated, namely: one case where the original pmfs have markedly different shapes; the second one where the Byzantines are able to blind the network; the last degenerate example where the original pmfs have disjoint support (i.e., infinite divergence). In the first example (six leftmost panels) we have initial distributions p and q such that $\alpha_b \approx 0.49$, and such that the divergence between the hypotheses is $D(q||p) \approx 8.3178$

[5]For $\alpha < \alpha_b$ the divergence minimization problem can be regarded in terms of successive minimization of the distance between two disjoint convex sets of probability distributions. The *alternating minimization algorithm* [21] is a general approach to solve this kind of problems.

nats. We now assume that the intruder power is $\alpha = 0.4$. The attacking distributions x and y are as shown in figure, and yield w and z as depicted in the two left-bottom panels. The final divergence (that between z and w) reduces to $\Delta(\alpha) \approx 6.4\ 10^{-2}$ nats. The second example illustrates a case where the divergence can be nullified; in fact here it is assumed $\alpha = 0.3$, while $\alpha_b \approx 0.22$. We start from $D(q\|p) \approx 0.237$ and the distributions x and y whose shape is illustrated (this is only one of the infinitely many possible pairs) yield identical final pmfs $w = z$, so that $\Delta(\alpha) = 0$. Finally, in the third example the initial divergence is $D(q\|p) = \infty$ and the attack leads to $\Delta(\alpha) \approx 8.1\ 10^{-2}$ nats; in fact, in this case $0.4 = \alpha < \alpha_b = 1/2$.

IV. Vector Observations

We now consider the case that m observations are collected at each sensor of the network. It is assumed that the m-vector that each node delivers to the FC is in any case made of i.i.d. samples, drawn from p or q, in the case of a honest sensor, or from x or y, for the Byzantines (i.e., no correlation can be imposed among the m samples of an infected node).

Let us extend the basic notation to the multiple-observations case. Let $\boldsymbol{K}^{(j)}$ be the m-dimensional vector of per-sensor observations, whose entries belong to $\mathcal{K}$. The generic statistical distributions of such a vector under $\mathcal{H}_1$ and $\mathcal{H}_0$ will be denoted by W and Z, respectively, and $D(Z\|W)$ is their multidimensional KL distance. Also, for any $\boldsymbol{k} \in \mathcal{K}$, we define $Q(\boldsymbol{k}) = \prod_{i=1}^{m} q_{k_i}$ $Y(\boldsymbol{k}) = \prod_{i=1}^{m} y_{k_i}$, $P(\boldsymbol{k}) = \prod_{i=1}^{m} p_{k_i}$ and $X(\boldsymbol{k}) = \prod_{i=1}^{m} x_{k_i}$. In what follows P, Q, X and Y (we often omit the argument $\boldsymbol{k}$) are always in product form. Thus, capital letters refer to multidimensional pmfs, while lower cases are used for their one-dimensional counterparts.[6]

Assume now that a vector of m i.i.d. observations is made available to each honest node. The Byzantine sensors, similarly, deliver to the FC a fictitious vector of i.i.d. samples drawn from the intruder's attacking distributions. As a matter of fact, this breaks down the similarity with robust detection formulation in α-contaminated classes, as our contaminating pmfs are constrained to be in product form. Otherwise stated, while the nominal distributions P and Q lie in a $|\mathcal{K}|^m$-dimensional space, the intruder is only allowed to act on $|\mathcal{K}|$-dimensional pmfs x and y. As far as we can tell, no standard results in simple analytical form are known for this contaminating model.

The pertinent hypothesis test can be now reformulated as follows. For $j = 1, 2, \ldots, n$, denoting the sensor index

$$\begin{aligned} \mathcal{H}_0 &: \Pr\{\boldsymbol{K}^{(j)} | \mathcal{H}_0 = \boldsymbol{k}\} = (1-\alpha)Q(\boldsymbol{k}) + \alpha Y(\boldsymbol{k}) \\ \mathcal{H}_1 &: \Pr\{\boldsymbol{K}^{(j)} | \mathcal{H}_1 = \boldsymbol{k}\} = (1-\alpha)P(\boldsymbol{k}) + \alpha X(\boldsymbol{k}). \end{aligned} \tag{11}$$

We can again refer to the KL distance as performance figure. The analogous of (4) for the case of $m > 1$ observations is

$$d_m(y;x) := D((1-\alpha)Q + \alpha Y \| (1-\alpha)P + \alpha X) \tag{12}$$

and represents the objective function to be minimized. We reiterate that $d_m(y;x)$ is a function of the *marginal* distributions x and y, as a consequence of the assumed i.i.d. property of the attacking distributions.

Let $d_\infty(y;x) = \lim_{m \to \infty} d_m(y;x)$. This limit may be infinite and in effect one would expect that it should be so. After all, we are saying that infinite observations are available so that error free decision should be possible, and the KL distance $d_\infty(y;x)$ should accordingly diverge. However there exist attacking distributions x and y such that $d_m(y;x)$ does not scale with m, implying that $d_\infty(y;x) < \infty$. The malicious intruder, obviously, will choose that. Let us see.

Define $\Delta_m(\alpha) := \min_{x,y} d_m(y;x)$ and $\Delta_\infty(\alpha) := \min_{x,y} d_\infty(y;x)$. In addition, denote with $h(\alpha)$ the KL distance between the binary pmfs $[1-\alpha, \alpha]$ and $[\alpha, 1-\alpha]$. We have the following result.

Theorem 2:

i) For any $\alpha \geq 1/2$ and $\forall m \geq 1$, $\Delta_m(\alpha) = 0$. The hypothesis-reversed emission strategy, that is, $x = q$, $y = p$, acting on exactly 50% of the nodes, achieves such minimum.

ii) For any $\alpha < 1/2$, and $m \geq 3$, $\Delta_m(\alpha)$ is strictly larger than zero.

iii) For any $\alpha < 1/2$, $\Delta_\infty(\alpha) = h(\alpha)$, and the pair (x, y) achieving such minimum again corresponds to the hypothesis-reversed emission strategy. ■

Proof: See Appendix II. •

To elaborate on the results of Theorem 2, we start considering $\alpha = 1/2$. In this case the "hypothesis-reversed" emission strategy $x = q$ and $y = p$, makes both the distribution appearing in the statistical test (11) equal to $1/2\,Q(\boldsymbol{k}) + 1/2\,P(\boldsymbol{k})$ and, therefore, the divergence in (12) goes to zero. This further implies that in the regime $\alpha > 1/2$ at least one pair of attacking distributions exists that nullifies the divergence, thus blinding the network (just infect 50% of the nodes and carry out the hypothesis-reversed attack). The conclusion is that the reversed emission strategy is optimal (from the intruder's viewpoint) and completely impairs the detection capabilities of the system, provided that $\alpha \geq 1/2$. This justifies statement i).

As to part ii), the implication is that, differently from the scenario considered in Theorem 1, an intruder infecting less than 50% of the nodes cannot completely impair the system, regardless on the initial distributions p and q. At rigor, this turns out to be true only if $m \geq 3$, but this is a technical condition with little impact on many practical systems where the typical setup is that of *large* m.

Now, let $\alpha < 1/2$, assume m finite, and consider again the hypothesis-reversed attack. The divergence $d_m(y;x)$ in (12) becomes $d_m(p;q) = D((1-\alpha)Q + \alpha P \| (1-\alpha)P + \alpha Q)$. A straightforward application of the log-sum inequality (see e.g., [19]) then reveals that $d_m(p;q) \leq h(\alpha)$, also implying that $d_m(p;q)$ does not scale with m. Theorem 2, part iii), simply ensures that, in the limit of increasingly large m, the bound is tight and that no other attacking distributions can do better (for the intruder) than that.

More specifically, the proof of Theorem 2, part iii), detailed in Appendix II-B, contains the following result. The limit $\lim_{m \to \infty} d_m(y;x)/m$ can only assume one of the four possible values shown in Table I. If such limit is strictly positive, then $d_m(y;x)$ scales linearly with the number of local

[6]We tolerate a slight asymmetry: p_k represents the pmf p evaluated at k, while $P(\boldsymbol{k})$ is the pmf P evaluated at $\boldsymbol{k}$, so that we use subindex in a case and parentheses in the other.

TABLE I
POSSIBLE RESULTS OF $\lim_{m\to\infty}(1/m)d_m(y;x)$ IN THEOREM 2, UNDER THE VARIOUS ASSUMPTIONS

assumptions	$D(q\|\|p) > D(q\|\|x)$	$D(q\|\|p) \le D(q\|\|x)$
$D(y\|\|p) > D(y\|\|x)$	$(1-\alpha)D(q\|\|x)+\alpha D(y\|\|x)$ nullified by $x=y=q$	$(1-\alpha)D(q\|\|p)+\alpha D(y\|\|x)$ minimized by $x=y$
$D(y\|\|p) \le D(y\|\|x)$	$(1-\alpha)D(q\|\|x)+\alpha D(y\|\|p)$ nullified by $x=q$ and $y=p$ (reversed emission)	$(1-\alpha)D(q\|\|p)+\alpha D(y\|\|p)$ minimized by $y=p$

observations m and therefore $d_\infty(y;x)=\infty$. For instance, let us consider the two solutions in the last column of Table I. Here, for large m, the intruder can achieve a minimum of $d_m(y;x)\sim m(1-\alpha)D(q\|\|p)$ by selecting, $x=y$ or $y=p$, respectively. This reaches the same asymptotic performance of a black-hole attack, wherein the exponent of the test is reduced by $(1-\alpha)$, the fraction of uninfected nodes.

As said, when $\lim_{m\to\infty} d_m(y;x)/m$ is positive, the unnormalized limit $d_\infty(y;x)$ diverges. However, there exist two possibilities to avoid that. These are those in Table I that nullify $\lim_{m\to\infty} d_m(y;x)/m$, namely, $x=y=q$, and $x=q$, $y=p$ (i.e., the hypothesis reversed). In the proof of the theorem it is shown that the latter is better for the intruder, leading to a lower value of $d_\infty(y;x)$. The conclusion is that the hypothesis-reversed distributions are optimal for the attack; the pertinent divergence is then $d_m(p;q)$ that does not scale with m.

The results of the theorem admit the following interpretation. At a first glance one can expect that, with infinitely many observations per sensor, the FC can reliably recognize which sensors are honest and which ones are Byzantine. The exponent $d_m(y;x)$ should accordingly scale at least as $m(1-\alpha)D(q\|\|p)$, corresponding to having discarded the fraction of infected nodes (see also Table I).

However, a little thought reveals that the above argument fails when the attacking distributions are exactly the original pmfs p and q used in reversed order. With this choice the FC is still able to recognize whether the samples from a certain sensor come from p or from q, but this does not enable to identify the Byzantine sensors. In fact, if the samples appear to be drawn from p, in view of the reversed emission strategy, this lead to the conclusion that: *either* the true hypothesis is $\mathcal{H}_1$ and the sensor is honest (this happens with probability $1-\alpha$) *or* the true hypothesis is $\mathcal{H}_0$ and the node is infected (with probability α). Similar considerations apply if data appear to come from q. It is now clear that the FC can only count the proportion of these two possibilities and, consistently, the asymptotic miss detection error exponent is just $h(\alpha)$.

It is worth noting that the log-sum inequality implies[7] $\Delta_m(\alpha)\le\Delta_{m+1}(\alpha)$ $(<\Delta_\infty(\alpha)=h(\alpha))$. As expected, from the network viewpoint, increasing the number of observations per sensor improves the detection performance. As consequence of Theorem 2, we also have that $\Delta_\infty(\alpha)$ is continuous and convex (such being $h(\alpha)$).

Finally, wherever only an upper bound $\bar{\alpha}$ is available at the fusion center, in the limit of $m\to\infty$ the Byzantine emissions do not change since the attacking distributions $x=q$ and $y=p$ are obviously independent of α, and the strategy of the fusion center still amounts to a sort of counting.

[7]In fact, $\Delta_{m+1}(\alpha)$ is a divergence encompassing the optimal choice of the $(m+1)$-dimensional attacking distributions. This is simply shown to be larger than the m-dimensional divergence that one obtains by marginalizing with respect to the $(m+1)$th dimension which, in turn, is larger than the m-dimensional divergence that one obtains by replacing x and y with those minimizing the m-dimensional problem. This latter, by definition, is $\Delta_m(\alpha)$.

V. NUMERICAL EXPERIMENTS

A. Fusion of Hard Decisions and Check of Convergence

The general results provided in Section III and Section IV can be specialized to the binary case, which can model networks where local sensors make their own decisions about the state of the nature, and deliver such binary data (also called hard decisions) to the FC. This latter implements an optimal fusion rule (likelihood ratio test) in order to end up with the final decision.

Let us first consider the single-observation case, i.e., $m=1$. Assume that $\mathcal{K}=\{0,1\}$, with the interpretation that $K^{(j)}=1$ whenever the local decision is for $\mathcal{H}_1$, and $K^{(j)}=0$ otherwise. Assume also, without loss of generality, that the two initial distributions p and q are such that $p_0\le q_0$, so that necessarily $p_1\ge q_1$ (if inequalities were reversed, in what follows one should simply exchange the roles of 0 and 1). From (7), we get $\alpha_b=(q_0-p_0)/(1+q_0-p_0)$. According to Theorem 1, we know that if $\alpha=\alpha_b$ there exists a single pair of distributions nullifying the divergence. Such pair is, see (8) (*deterministic emission*)

$$x_0=1,\quad x_1=0,\quad y_0=0,\quad y_1=1. \tag{13}$$

Equation (9), specialized to the binary case, reveals that the above solution is exactly the same attaining $d(y;x)=\Delta(\alpha)$ in the case where $\alpha<\alpha_b$, and the divergence cannot be brought to zero. On the other hand, if $\alpha>\alpha_b$, the infinitely many pairs of binary pmfs that make $\Delta(\alpha)=0$ can be put in the form $x_0=[(1-\alpha)/\alpha][\alpha_b/(1-\alpha_b)]+\zeta$, $y_0=\zeta$ and $x_1=1-[(1-\alpha)/\alpha][\alpha_b/(1-\alpha_b)]-\zeta$, $y_1=1-\zeta$, where ζ can be arbitrarily chosen between zero and $1-[(1-\alpha)\alpha_b/(\alpha(1-\alpha_b)]$.

As α approaches α_b from above, the admissible values of ζ tend to the single value $\zeta=0$. At that point we get (13). Further decreasing of α does not modify the solution but obviously impacts on the resulting (nonzero) $\Delta(\alpha)$. In this regime, from (13) we see that the strategy of the intruder is to compensate the original proportion of zeros and ones delivered by the sensors, by enforcing the nodes under its influence to send always the symbol 1, if $\mathcal{H}_0$ is in force, and always the symbol 0 when $\mathcal{H}_1$ is the underlying state of the nature. This *deterministic* Byzantine emissions is exactly what one expects.

Let us now switch to the case of many observations, i.e., $m>1$. We compute numerically the minimum achievable divergence $\Delta_m(\alpha)$ and the corresponding optimal pair (x,y), with the aim of checking the effectiveness of the asymptotic results given in Theorem 2.

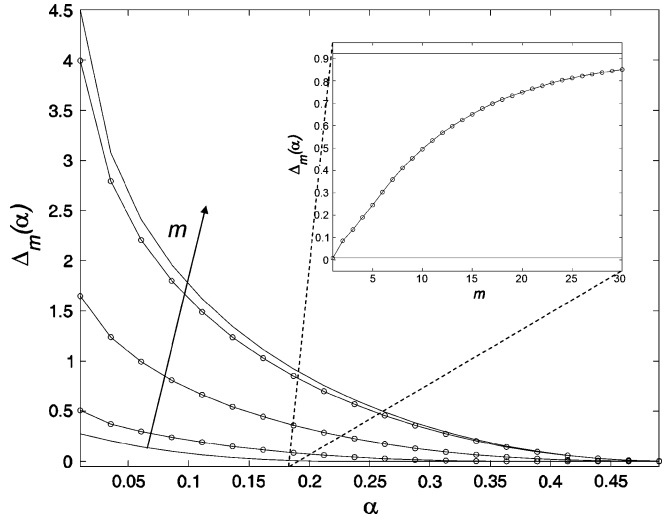

Fig. 5. Attacking exponent $\Delta_m(\alpha)$ versus the attacking power α, for different values of the vector size $m = 2, 7, 30$. The original distributions are $p = [0.1,\ 0.9]$ and $q = [0.4,\ 0.6]$. The lowermost curve is $\Delta_1(\alpha) \equiv \Delta(\alpha)$ considered in Theorem 1 and depicted in Fig. 1. The uppermost curve is $\Delta_\infty(\alpha) = h(\alpha)$ and refers to the case of infinitely many observations per sensor. This latter, clearly, does not depend upon the considered p and q. The inner plot refers to $\alpha = 0.1868$ and shows the behavior of $\Delta_m(\alpha)$ versus the vector size m. A monotonic growth is observed from $\Delta(\alpha)$ (lower horizontal line) to $h(\alpha)$ (upper), as expected. The rate of convergence varies from case to case; here m in the order of a few tens may be enough.

In Fig. 5, the curve $\Delta_m(\alpha)$ is plotted as a function of the intruder's power α, for several values of m, and for a particular choice of the binary pmfs p and q. Also the (theoretical) limiting curve $\Delta_\infty(\alpha) = h(\alpha)$ is depicted. The emerging trend is that, as predicted by the theory, the curves $\Delta_m(\alpha)$ monotonically approach the limiting curve $\Delta_\infty(\alpha)$, as the number of samples per sensor grows.

An alternative view of the convergence is given in the inner plot of Fig. 5, where α is held fixed, and $\Delta_m(\alpha)$ is plotted as a function of m. For comparison purposes, we also display the straight lines corresponding to the values $\Delta(\alpha)$ and $\Delta_\infty(\alpha)$, which are exactly computed thanks to Theorems 1 and 2, respectively. The convergence is more clearly highlighted, and it is seen that a relatively moderate number of observations per sensor is sufficient to reach the asymptote.

Inner plot of Fig. 5 also suggests a further speculation. Suppose that one knows the exact value of $\Delta_m(\alpha)$ for a given m. Then, thanks to the monotone behavior of $\Delta_m(\alpha)$ with m, the residual error that one would make in assuming true the asymptotic value $\Delta_\infty(\alpha)$, for vector sizes larger than m, cannot exceed $\Delta_\infty(\alpha) - \Delta_m(\alpha)$.

B. Actual Detection Probability

In this section, we report the results of several Monte Carlo (MC) simulations, aimed at verifying the effectiveness of our asymptotic analysis, in scenarios of practical relevance. Specifically, we numerically estimate the detection probability pertaining to different attacking strategies, chosen among those previously encountered.

As a case study, we consider an original (intrusion-free) hypothesis test involving two binomial pmfs with different parameters $q = Bi(L, \pi_0)$, $p = Bi(L, \pi_1)$, where $Bi(L, c)$ stems for a pmf pertaining to a binomial experiment with L trials and probability of success c. We implement likelihood ratio tests for different choices of the intruder's pmfs x and y, and for both the cases of single and multiple observations. Specifically, we address the following.

1) The case $m = 1$ with the pmfs $x = x^*$ and $y = y^*$ which are optimal for the case of a single observation per sensor (i.e., found by the water-filling approach).
2) The case $m > 1$, with the hypothesis-reversed emission strategy $x = q$, $y = p$, which optimizes the multiple-observations scenario.
3) The case $m > 1$, with the strategy $x = y = q$, which is not optimal but, according to what previously discussed in Section IV, gives a KL distance that does not scale with m.
4) The case $m > 1$, adopting the attacking strategy $x = x^*$, $y = y^*$, that is, using the pmfs optimal for the case of $m = 1$. Oppositely to previous items, this attack is expected to give a KL distance growing with m.

In Fig. 6, left panel, we plot the (MC-estimated) detection probability as a function of α, for the above four cases, for fixed false alarm probability ≈ 0.1. The actual performances, estimated by 10^4 MC iterations, are compared with those computed using the KL distances according to Stein's lemma. Two relevant features emerge from the inspection of the figure.

First, consider the behavior of the optimized solutions (items 1 and 2 above). In the leftmost curve, the detection probability pertaining to the case $m = 1$ is reasonably close to the asymptotic approximation $1 - e^{-n\Delta(\alpha)}$; similarly, the case $m = 10$ (rightmost curve) exhibits a good match between the actual detection probability and the asymptotic value $1 - e^{-nh(\alpha)}$. The only marked difference arises when $\alpha > \alpha_b$ in the case that $m = 1$, and when α is near $1/2$, for $m > 1$. This can be simply understood. Theorems 1 and 2, part i), establish that in these ranges the test reduces to an independent coin flip. The simulated curves behave accordingly. The inaccuracy of the theoretical curves are to be ascribed to the asymptotic nature of Stein's lemma, from which these curves are derived.

As to items 3 and 4, the simulations evidence how these non-optimal attacking distributions are definitely disadvantageous for the intruder. Due to finite MC-samples effect, the estimated detection probabilities are all too close to unity, and an increase of the Monte Carlo runs would be necessary to compare the results each other. This is symptomatic of the strong sub-optimality of the intruder's choice.

A more quantitative analysis is given in right panel of Fig. 6 that focuses on the estimated KL distances corresponding to the same simulations of the left panel, and compares these values to the asymptotic values given in Table I. It is seen that the case of $m = 1$ is the least favorable for the network. Progressive performance improvements are observed for the cases $m > 1$, in the expected ordering. In particular, the most pronounced intruder's performance loss (gain for the network) appears in the case $x = x^*$, $y = y^*$, in that, as already noticed, this choice let the divergence scale with m. In addition, starting from a certain value of α, this last curve behaves somehow irregularly. Such a value of α is recognized to be exactly $\alpha_b = 0.297$ and the strange shape of the KL distance is a consequence of the arbi-

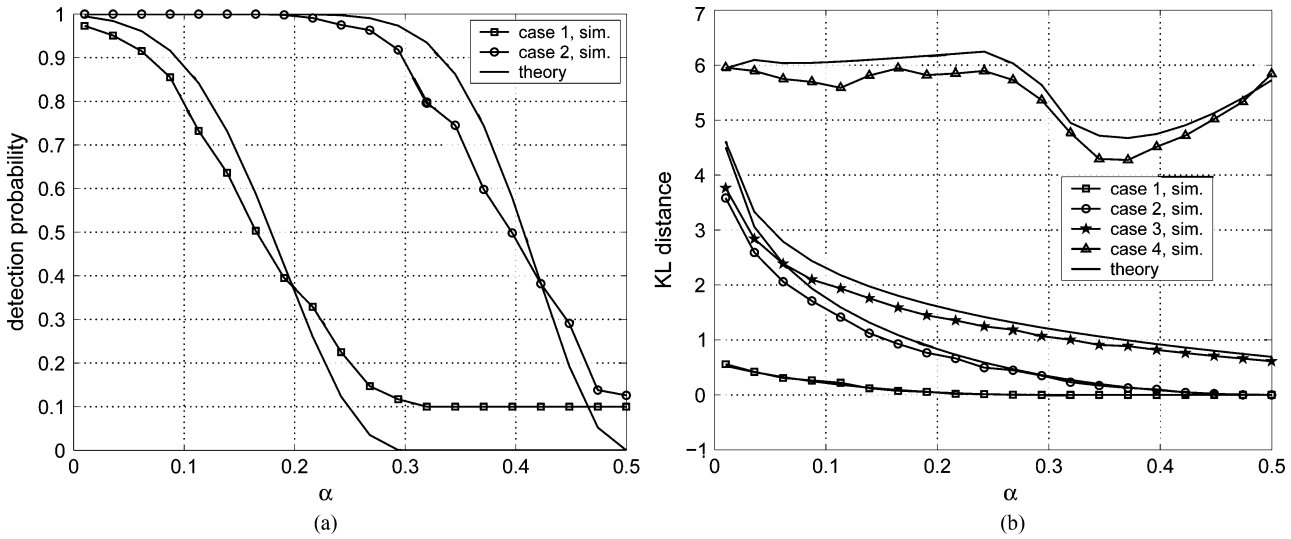

Fig. 6. (a) With reference to the four attacking distributions as detailed in the list of Sections V-B, we estimate the detection probability by 10^4 Monte Carlo runs. Here the false alarm probability is $\approx 10^{-1}$, the number of sensors is $n = 10$, and p and q are two binomial pmfs with $L = 5$, $\pi_1 = 0.3$, and $\pi_0 = 0.1$. For comparison, the detection probabilities obtained via Stein's lemma exploiting the theoretical exponents are also shown. The optimized (from the intruder's viewpoint) attacking distributions (cases 1 and 2) lead to the shown curves; cases 3 and 4 are not shown because the pertinent curves all collapse to 1, as symptomatic of a high performing network. (b) For the same attacking distributions we depict the error exponent of the statistical tests. This allows us to show all the four cases described in the list of Sections V-B.

trary choice of the distributions, which is allowed in that range [see Theorem 2, part ii)].

C. Two Practical Issues

In this section we discuss some practical aspects of assuming: i) knowing an upper bound $\bar{\alpha}$ instead of the actual attacking power α; ii) the presence of some statistical correlation among the sensors' delivering. We limit the analysis to a couple of illustrative examples.

We consider the same scenario as in Fig. 6, left panel, for $m = 1$, again with $n = 10$ sensors. Different from the previous simulations, here we assume that an upper bound $\bar{\alpha} = 0.25$ is available, while the actual α is allowed to vary. The Byzantines select the attacking distributions x and y corresponding to the true α, while, as previously said, the network opts for a (conservative) design exploiting the likelihood ratio test for the worst-case $\bar{\alpha}$.

The resulting detection probability is displayed in the uppermost plot of Fig. 7, as a function of α. For $\alpha = \bar{\alpha}$, we obtain the same value shown in Fig. 6. On the other hand, according to the prescriptions of robust statistics, a decrease of the actual attacking power results in an improvement of the detection probability and of the false alarm rate (not shown here for the sake of brevity), which can be substantially smaller than the nominal value 0.1.

As a comparison, we also report the detection performance of a network having access to the exact value of α (see curve labelled as "known α"), for a false alarm rate corresponding to the nominal requirement of 0.1. Such comparison, clearly, is not made at the same false alarm level. To complete the analysis, we tune the threshold of the latter system in order to let both detectors share the same false alarm probability (see curve labelled as "known α, tuned threshold"). As it must be, the detector that knows α always outperforms that knowing only the bound.

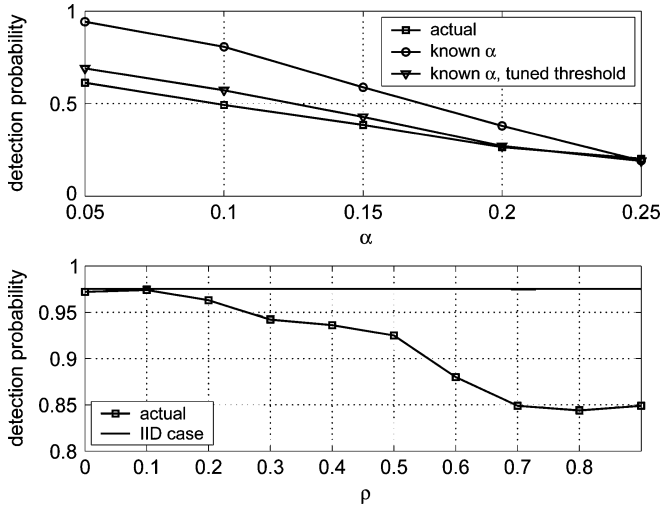

Fig. 7. Detection probability in the presence of non perfect knowledge of the attacking power (upper plot), and in the presence of non i.i.d. data (lower plot). See the main text for details.

We now consider the effects of measurement correlation. To impose a certain degree of statistical dependence among the honest sensors' observations, we simply use the classical model of correlated Gaussian observations, with correlation coefficient ρ and unit variance. We consider a shift-in-mean problem, with zero-mean variables under $\mathcal{H}_0$, and mean μ under $\mathcal{H}_1$. The observations are uniformly quantized before sensors' delivering. In this case, both the network and the Byzantines follow the operational mode designed under the i.i.d. situation. In the lowermost plot of Fig. 7, the actual detection probability is displayed as a function of the correlation coefficient ρ. As it can be seen, the theoretical predictions are accurate enough for moderate degrees of correlation.

VI. Conclusion

In this paper, we have considered distributed detection in the presence of Byzantine sensors created by an intruder, and characterized the power of attack analytically. We are able to provide closed-form expressions for the worst detection error exponent of an optimized NP detector at the fusion center, and for the corresponding attacking distributions. We also give expressions of the minimum attacking power α_b above which the ability to detect is completely destroyed.

As to the case of vector observations, we find that an intruder infecting less than 50% of the nodes cannot completely impair the system, regardless of the distributions of the sensors' observations. It can, however, severely degrade the performance of the network, and the attacking distributions that achieve this goal are "hypothesis-reversed": when the true state of the nature is $\mathcal{H}_1$, infected sensors deliver data according to the distribution that actually pertains to hypothesis $\mathcal{H}_0$, and vice versa. A notable fact is that the asymptotic detection probability does not scale exponentially with the vector size m. Actually, it does not scale at all. The practical consequence is a saturation effect: increasing the number of per-sensor observations beyond a certain amount does not provide any significant improvement.

This scaling behavior may appear rather counterintuitive if one considers that each (honest/Byzantine) sensor contributes to the detection statistics with m independent observations. On the other hand, the expected scaling law is instead preserved with respect to the total number of (independent) nodes n. Indeed, we find that the asymptotic detection probability (for a given false alarm level) is approximately $1-\exp[-n\,h(\alpha)]$, where $h(\alpha)$ represents a binary divergence. This has the nice interpretation that, once saturated, the best test that the network can implement is a sort of honest/Byzantine-appearing sensor counting.

We should also point out several limitations of our results. Note that we have assumed very strong Byzantine sensors that actually know the true hypothesis. This model is overly conservative in practice. Thus one should view the results presented here are a form of worst case assessment of the risk of Byzantine attack. The conditional i.i.d. assumption is also limiting.

As a future line of research it may be of interest to consider a Bayesian formulation with *a priori* probabilities assigned to the hypotheses. In this setting the asymptotic performance can be measured in terms of the Chernoff information [19]. It might be also assumed that the intruder not only knows the state of the nature, but it may also controls it by tuning, to some extent, the probability of occurrence of the hypotheses. Finally, note that the same tools used in this paper are certainly exploitable for studying the attacks of a less dangerous intruder that does *not* know the true state of the nature.

Appendix I
Proof of Theorem 1

Preliminary, let us note that $\alpha_b \le 1/2$ immediately follows from $\sum(q_k-p_k)^+ = \sum_{k:\,q_k>p_k} q_k - \sum_{k:\,q_k>p_k} p_k \le 1$.

Conditions for Definitely Impairing the Network—Parts i) and ii) in Theorem 1: Clearly, $d(y;x)$ in (4) is zero if and only if $w=z$, i.e., if the following set of conditions hold: $\forall k \in \mathcal{K}$

$$\begin{cases} y_k = x_k + \frac{1-\alpha}{\alpha}(p_k-q_k) \\ \sum x_k = \sum y_k = 1 \\ x_k, y_k \ge 0. \end{cases} \Rightarrow \begin{cases} y_k = x_k + \frac{1-\alpha}{\alpha}(p_k-q_k) \\ \sum x_k = 1 \\ x_k \ge \frac{1-\alpha}{\alpha}\,(q_k-p_k)^+. \end{cases}$$

Let us set $\xi_k = 1-\alpha/\alpha\,(q_k-p_k)^+$. We have only three possibilities.

1) $\sum \xi_k = 1$. In this case ξ is a distribution vector. From the definition of α_b in (7), it follows $\alpha = \alpha_b$, and the *unique* solution of the above system is readily obtained assuming equality in the last equation, i.e., $x_k = \xi_k$, and $y_k = \xi_k + 1 - \alpha/\alpha(p_k - q_k)$, yielding (8).
2) $\sum \xi_k < 1$. In this regime $\alpha > \alpha_b$. Let us tentatively set $x = \xi$ as in the solution (8). As $\sum \xi_k < 1$, x is *not* a pmf, but we can clearly arbitrarily increase values of some of the x_k's until x becomes a pmf. After doing that, setting y as in the first equation of the above system, verifies the conditions for $d(y;x)=0$.
3) $\sum \xi_k > 1$. Here $\alpha < \alpha_b$. The divergence cannot be nullified because the elements of x are lower bounded by those in ξ. This case is dealt with in the following.

Minimizing the Achievable Exponent—Parts iii) and iv) in Theorem 1: We can formally state the problem as the following constrained convex optimization

$$\begin{cases} \min_{x,y} d(y;x) \\ x_k, y_k \ge 0, \qquad \forall k \in \mathcal{K} \\ \sum x_k = \sum y_k = 1. \end{cases} \tag{14}$$

The objective function $d(y;x)$ is continuously differentiable and it is easily recognized to be convex $\cup$ in the pair (x,y). These properties are obviously also true for the inequality constraints, $g_{x,k}(x,y) := -x_k \le 0$ and $g_{y,k}(x,y) := -y_k \le 0$, as well as for the equality constraints, $h_x(x,y) := \sum x_k - 1 = 0$ and $h_y(x,y) := \sum y_k - 1 = 0$.

Accordingly, Karush-Kuhn-Tucker equations lead to the following result [22]. Necessary and sufficient conditions for the point (x,y) to be a minimizer of $d(\cdot;\cdot)$ are that there must exist constants $\mu_{x,k}, \mu_{y,k} \ge 0$, $k \in \mathcal{K}$, and Ω_x, Ω_y, such that (see the equation at the bottom of the page).On accounting for the

$$\begin{cases} \nabla d(y;x) - \sum \mu_{x,k}\,\nabla g_{x,k}(x,y) - \sum \mu_{y,k}\,\nabla g_{y,k}(x,y) \\ \qquad -\Omega_x\,\nabla h_x(x,y) - \Omega_y\,\nabla h_y(x,y) = 0 cr \mu_{x,k} \quad x_k = 0,\, k \in \mathcal{K} \\ \mu_{y,k}\,y_k = 0, \quad k \in \mathcal{K}. \end{cases}$$

definitions of $g_{x,k}$, $g_{y,k}$, h_x and h_y, the above equations can be manipulated to work out the pertinent derivatives, yielding

$$\begin{cases} \frac{(1-\alpha)q_k+\alpha y_k}{(1-\alpha)p_k+\alpha x_k} = \frac{1}{\gamma_x} & \text{all } k \text{ s.t. } x_k > 0 \\ \frac{(1-\alpha)q_k+\alpha y_k}{(1-\alpha)p_k} \leq \frac{1}{\gamma_x} & \text{all } k \text{ s.t. } x_k = 0 \\ \frac{(1-\alpha)q_k+\alpha y_k}{(1-\alpha)p_k+\alpha x_k} = \gamma_y & \text{all } k \text{ s.t. } y_k > 0 \\ \frac{(1-\alpha)q_k}{(1-\alpha)p_k+\alpha x_k} \geq \gamma_y & \text{all } k \text{ s.t. } y_k = 0 \end{cases} \tag{15}$$

where we have set for simplicity $\gamma_x = -\alpha/\Omega_x$, and $\gamma_y = \exp[(\Omega_y - \alpha)/\alpha]$.

As it is easily recognized that γ_x, $\gamma_y > 0$, combining the first and the last two formulas yields

$$x_k = \left[\gamma_x \left(\frac{1-\alpha}{\alpha} q_k + y_k\right) - \frac{1-\alpha}{\alpha} p_k\right]^+, \tag{16}$$

$$y_k = \left[\gamma_y \left(\frac{1-\alpha}{\alpha} p_k + x_k\right) - \frac{1-\alpha}{\alpha} q_k\right]^+. \tag{17}$$

In the above, γ_x and γ_y are to be set to guarantee that $\sum x_k = \sum y_k = 1$, in order to get a valid pair (x, y). The procedure for doing that is known as water-filling and has been described in the main text.

The solution provided by (16) and (17) can be greatly simplified by showing that, for any $k \in \mathcal{K}$, x_k and y_k cannot be both positive, i.e., $x_k\, y_k = 0, \forall k \in \mathcal{K}$. To this aim, we first show that γ_x and γ_y are both in $(0, 1]$. In fact, from the first two equations in (15), we get $[(1-\alpha)q_k + \alpha y_k]/[(1-\alpha)p_k + \alpha x_k] \leq 1/\gamma_x$, or equivalently, $\gamma_x\,[(1-\alpha)q_k + \alpha y_k] \leq (1-\alpha)p_k + \alpha x_k$, which, summed over $k \in \mathcal{K}$, yields $\gamma_x \leq 1$. In the same way, by elaborating on the last two equations in (15), we have $\gamma_y \leq 1$.

Assume now that x_k and y_k are both positive for some k. Then, first and third equation in (15) immediately imply $\gamma_x\gamma_y = 1$ that, accounting for $0 < \gamma_x,\ \gamma_y \leq 1$, reduces to $\gamma_x = \gamma_y = 1$. On the other hand, if $\gamma_x = \gamma_y = 1$, the same first and third equation in (15) yield

$$(1-\alpha)q_k + \alpha y_k = (1-\alpha)p_k + \alpha x_k \tag{18}$$

for all k such that either x_k, or y_k, or both, are strictly positive. For the remaining indexes k's, namely those such that $x_k = y_k = 0$, second and fourth equations of (15) tell us that $q_k/p_k \leq 1$ and $q_k/p_k \geq 1$, namely $p_k = q_k$; this implies that (18) still holds true. We have thus shown that if there exists an index k such that $x_k > 0$ and $y_k > 0$, then $\gamma_x\gamma_y = 1$ and (18) is true for all $k \in \mathcal{K}$. The immediate implication is that $d(y; x) = 0$, which contradicts our assumption of $\alpha < \alpha_b$.

Capitalizing on the property that $x_k\, y_k = 0,\ \forall k \in \mathcal{K}$, the solution expressed by (16) and (17) can be simplified to the expressions given in (9) of the theorem. The existence of suitable constants γ_x and γ_y is guaranteed by the fact that $\sum x_k$ is either zero, or a strictly increasing continuous function of $\gamma_x \in (0, 1]$: by increasing γ_x there certainly exists an unique value such that the sum is exactly 1. Obviously, similar arguments apply to γ_y. This concludes the proof of part iii).

Consider now the properties of $\Delta(\alpha)$. Accounting for the water-filling procedure (see discussion following Theorem 1), it is easily seen that both the constants γ_x and γ_y vary continuously with α. Then, from solution (9) it follows that x_k and y_k are continuous in α, and the same holds for w and z at optimality. The continuity of the divergence with respect to the involved distributions immediately implies the continuity of $\Delta(\alpha)$.

As α increases, the minimization of $D(z\|w)$ is performed over larger and larger sets and this ensures that $\Delta(\alpha)$ is non-increasing. Thus, the function $\Delta(\alpha)$ goes with continuity from $D(q\|p)$, when $\alpha = 0$, to 0, when $\alpha = \alpha_b$. We next show that $\Delta(\alpha)$ is convex $\cup$ so that all the claims of the theorem follow.

Let us pose $\alpha = \lambda\alpha' + (1-\lambda)\alpha''$, and further define y', x' as the pmfs which minimize $d(y; x)$ at the point α', and y'', x'' as the pmfs which minimize $d(y; x)$ at the point α''. We can accordingly write

$$\begin{aligned} &\lambda\Delta(\alpha') + (1-\lambda)\Delta(\alpha'') \\ &\quad = \lambda D(z'\|w') + (1-\lambda)D(z''\|w'') \\ &\quad \geq D\left(\lambda z' + (1-\lambda)z''\|\lambda w' + (1-\lambda)w''\right) \end{aligned} \tag{19}$$

where we have set $z' = (1-\alpha')q + \alpha' y'$, $z'' = (1-\alpha'')q + \alpha'' y''$, $w' = (1-\alpha')p + \alpha' x'$, $w'' = (1-\alpha'')p + \alpha'' x''$, and the last inequality follows by the convexity of the divergence. Straightforward algebra yields

$$\begin{aligned} \lambda z' + (1-\lambda)z'' &= \left[1 - (\lambda\alpha' + (1-\lambda)\alpha'')\right] q \\ &\quad + \lambda\alpha' y' + (1-\lambda)\alpha'' y'' \\ &= (1-\alpha)q + \alpha\bar{y} \end{aligned} \tag{20}$$

where $\bar{y} = [\lambda\alpha' y' + (1-\lambda)\alpha'' y'']/\alpha$, which can be readily verified to possess all the requirements for being a pmf. With the same arguments, it is easy to obtain

$$\lambda w' + (1-\lambda)w'' = (1-\alpha)p + \alpha\bar{x} \tag{21}$$

with $\bar{x} = [\lambda\alpha' x' + (1-\lambda)\alpha'' x'']/\alpha$. Using (20) and (21), the LHS of (19) is $\geq D((1-\alpha)q + \alpha\bar{y}\|(1-\alpha)p + \alpha\bar{x}) \geq \Delta(\alpha)$, where $\bar{y}$ and $\bar{x}$ are not necessarily the minimizing pmfs corresponding to α. The proof is now complete. •

Appendix II
Proof of Theorem 2

Vectors of Finite Size—Parts i) and ii) in Theorem 2: Actually, proof of statement i) is simple and is provided in the comments following Theorem 2. Consider hence part ii). Assuming $\alpha < 1/2$, to completely impair the network the intruder must nullify the KL distance in (12). This happens if and only if $(1-\alpha)\prod_{i=1}^m q_{k_i} + \alpha\prod_{i=1}^m y_{k_i} = (1-\alpha)\prod_{i=1}^m p_{k_i} + \alpha\prod_{i=1}^m x_{k_i}$. Differently from the scalar case, we now show that assuming $m \geq 3$, the above equation cannot be ever satisfied. In fact, to fulfill the above equation for the joint m-dimensional pmfs, it is clearly required that all the joint pmfs of lower order must verify analogous equalities. Assuming $m = 3$ this implies, $\forall\ k_1$, k_2, $k_3 \in \mathcal{K}$:

$$\begin{cases} y_{k_1} = x_{k_1} + \frac{1-\alpha}{\alpha}(p_{k_1} - q_{k_1}) \\ y_{k_1}y_{k_2} = x_{k_1}x_{k_2} + \frac{1-\alpha}{\alpha}(p_{k_1}p_{k_2} - q_{k_1}q_{k_2}) \\ y_{k_1}y_{k_2}y_{k_3} = x_{k_1}x_{k_2}x_{k_3} + \frac{1-\alpha}{\alpha}(p_{k_1}p_{k_2}p_{k_3} - q_{k_1}q_{k_2}q_{k_3}). \end{cases}$$

By restricting to the special case of $k = k_1 = k_2 = k_3$, we have

$$\begin{cases} y_k = x_k + \frac{1-\alpha}{\alpha}(p_k - q_k) \\ y_k^2 = x_k^2 + \frac{1-\alpha}{\alpha}(p_k^2 - q_k^2) \\ y_k^3 = x_k^3 + \frac{1-\alpha}{\alpha}(p_k^3 - q_k^3) \end{cases}$$

that can be easily shown to admit a solution in the pair (x, y) only for $p = q$, provided that α is strictly lower than 1/2.

Vectors of Infinitely Many Samples—Part iii) in Theorem 2: For notational convenience, let us set

$$\Lambda(\boldsymbol{k}) = \log \frac{(1-\alpha)Q(\boldsymbol{k}) + \alpha Y(\boldsymbol{k})}{(1-\alpha)P(\boldsymbol{k}) + \alpha X(\boldsymbol{k})}. \tag{22}$$

We start with a couple of Lemmas which will turn out to be useful for the proof of the theorem.

Lemma 1: A positive constant μ exists such that $|\Lambda(\boldsymbol{k})| \leq \log 1/\alpha + m\mu$. ■

Proof: We rule out the cases leading to infinite divergences, in that we are (or better the intruder is) interested in *minimizing* the achievable exponent. Elaborating on the numerator of (22), we have

$$\begin{aligned}(1-\alpha)Q(\boldsymbol{k}) + \alpha Y(\boldsymbol{k}) &\leq (1-\alpha)q_{\max}^m + \alpha y_{\max}^m \\ &\leq [\max(q_{\max}, y_{\max})]^m\end{aligned}$$

where $q_{\max} = \max_k q_k$ and $y_{\max} = \max_k y_k$. On the other hand, for the denominator

$$\begin{aligned}&\min_{\boldsymbol{k}}[(1-\alpha)P(\boldsymbol{k}) + \alpha X(\boldsymbol{k})] \\ &= \min\left\{ \min_{\boldsymbol{k}:P(\boldsymbol{k})\neq 0}[(1-\alpha)P(\boldsymbol{k}) + \alpha X(\boldsymbol{k})] \right. \\ &\qquad \left. \min_{\boldsymbol{k}:X(\boldsymbol{k})\neq 0}[(1-\alpha)P(\boldsymbol{k}) + \alpha X(\boldsymbol{k})] \right\}\end{aligned} \tag{23}$$

because, in view of the assumption that the denominator is never zero, $\{\boldsymbol{k} : P(\boldsymbol{k}) = 0\} \subseteq \{\boldsymbol{k} : X(\boldsymbol{k}) \neq 0\}$. Equation (23) leads to

$$\begin{aligned}&\min_{\boldsymbol{k}}[(1-\alpha)P(\boldsymbol{k}) + \alpha X(\boldsymbol{k})] \\ &\geq \min\left[(1-\alpha)p_{\min}^m, \alpha x_{\min}^m\right] \\ &\geq \alpha[\min(x_{\min}, p_{\min})]^m\end{aligned}$$

where $p_{\min} = \min_{k:p_k\neq 0} p_k$ and $x_{\min} = \min_{k:x_k\neq 0} x_k$. Putting pieces together, the simple bound

$$\begin{aligned}\Lambda(\boldsymbol{k}) &\leq \log\frac{1}{\alpha} + m\log\frac{\max(q_{\max}, y_{\max})}{\min(x_{\min}, p_{\min})} \\ &= \log\frac{1}{\alpha} + m\mu_u\end{aligned}$$

is derived. By symmetry arguments, it is easily shown that

$$\begin{aligned}\Lambda(\boldsymbol{k}) &\geq \log\alpha + m\log\frac{\min(q_{\min}, y_{\min})}{\max(x_{\max}, p_{\max})} \\ &= \log\alpha + m\mu_l\end{aligned}$$

with $q_{\min} = \min_{k:q_k\neq 0} q_k$, $y_{\min} = \min_{k:y_k\neq 0} y_k$. By posing $\mu = \max(\mu_u, -\mu_l)$ the statement of the lemma follows. •

For the special case that $y = p$ and $x = q$, a bound independent on m can be obtained. This is contained in the following result.

Lemma 2: For the hypothesis-reversed strategy $x = q$, $y = p$, it holds true that $|\Lambda(\boldsymbol{k})| \leq \log(1 - \alpha/\alpha)$. ■

Proof: For the reversed strategy the log-likelihood becomes

$$\Lambda(\boldsymbol{k}) = \log \frac{(1-\alpha)Q(\boldsymbol{k}) + \alpha P(\boldsymbol{k})}{(1-\alpha)P(\boldsymbol{k}) + \alpha Q(\boldsymbol{k})}. \tag{24}$$

Assuming all probability terms strictly greater than zero, the log-sum inequality implies

$$\begin{aligned}&[(1-\alpha)Q(\boldsymbol{k}) + \alpha P(\boldsymbol{k})]\Lambda(\boldsymbol{k}) \\ &\leq (1-\alpha)Q(\boldsymbol{k})\log\frac{(1-\alpha)}{\alpha} \\ &\quad + \alpha P(\boldsymbol{k})\log\frac{\alpha}{(1-\alpha)}.\end{aligned} \tag{25}$$

On the other hand, when, for instance, $Q(\boldsymbol{k}) = 0$ (recall that we do not allow both Q and P to be zero), the above inequality is in fact an equality. Equation (25) can be rephrased as $\Lambda(\boldsymbol{k}) \leq \log(1 - \alpha/\alpha)$. Similarly, by simply exchanging P with Q, one gets $\Lambda(\boldsymbol{k}) \geq \log(\alpha/1-\alpha)$. Combining the above bounds proves the lemma. •

We are ready now to prove Theorem 2. In the proof we first separate the class of $\boldsymbol{k}$-sequences in two classes, by defining suitable *typical* sets (broadly speaking). Then, the asymptotic behavior of the relevant KL distance is investigated separately in these subsets.

Let z and w a generic pair of (one-dimensional) pmfs. We define

$$\mathcal{A}_\epsilon^{(m)}(z, w) := \left\{ \boldsymbol{k} : \left| \frac{1}{m}\sum_{i=1}^{m} \log\frac{z_{k_i}}{w_{k_i}} - D(z\|w) \right| < \epsilon \right\} \tag{26}$$

and

$$\mathcal{B}_q := \mathcal{A}_\epsilon^{(m)}(q, p) \cap \mathcal{A}_\epsilon^{(m)}(q, x) \cap \mathcal{A}_\epsilon^{(m)}(q, y). \tag{27}$$

Note that, if $\boldsymbol{K}$ is a random vector drawn i.i.d. according to q, then the probability that it belongs to $\mathcal{B}_q$, say $\Pr\{\boldsymbol{K} \in \mathcal{B}_q | q\}$, tends to 1 when $m \longrightarrow \infty$, as direct consequence of the law of large numbers that makes the probability of each $\mathcal{A}_\epsilon^{(m)}$ vanishingly small, $\forall \epsilon > 0$.

Now, from definition (13), we can write

$$d_m(y; x) = (1-\alpha)\sum_{\boldsymbol{k}} Q(\boldsymbol{k})\Lambda(\boldsymbol{k}) + \alpha\sum_{\boldsymbol{k}} Y(\boldsymbol{k})\Lambda(\boldsymbol{k}). \tag{28}$$

Let us first focus on the first summation, which can be split as

$$\sum_{\boldsymbol{k}} Q(\boldsymbol{k})\Lambda(\boldsymbol{k}) = \sum_{\boldsymbol{k}\notin\mathcal{B}_q} Q(\boldsymbol{k})\Lambda(\boldsymbol{k}) + \sum_{\boldsymbol{k}\in\mathcal{B}_q} Q(\boldsymbol{k})\Lambda(\boldsymbol{k}). \tag{29}$$

Lemma 1 implies

$$\frac{1}{m}\left|\sum_{\boldsymbol{k}\notin\mathcal{B}_q} Q(\boldsymbol{k})\Lambda(\boldsymbol{k})\right| \le \left(\frac{1}{m}\log\frac{1}{\alpha}+\mu\right)\sum_{\boldsymbol{k}\notin\mathcal{B}_q} Q(\boldsymbol{k}) = \left(\frac{1}{m}\log\frac{1}{\alpha}+\mu\right)\Pr\{\boldsymbol{K}\notin\mathcal{B}_q|q\}$$

which vanishes as m goes to infinity.

As to the second sum in (29), the log-likelihood $\Lambda(\boldsymbol{k})$ can be rewritten as

$$\log\frac{(1-\alpha)+\alpha\exp\left[-\sum_{m=1}^{m}\log\frac{q_{k_i}}{y_{k_i}}\right]}{(1-\alpha)\exp\left[-\sum_{m=1}^{m}\log\frac{q_{k_i}}{p_{k_i}}\right]+\alpha\exp\left[\sum_{m=1}^{m}\log\frac{q_{k_i}}{x_{k_i}}\right]}$$

where we assume that none of the involved divergences is infinite. These singular cases can be then addressed by continuity arguments, and, as expected, are of no interest.

In the set $\mathcal{B}_q$,

$$\log\frac{(1-\alpha)+\alpha e^{-m[D(q\|y)+\epsilon]}}{(1-\alpha)e^{-m[D(q\|p)-\epsilon]}+\alpha e^{-m[D(q\|x)-\epsilon]}} \le \Lambda(\boldsymbol{k}) \le \log\frac{(1-\alpha)+\alpha e^{-m[D(q\|y)-\epsilon]}}{(1-\alpha)e^{-m[D(q\|p)+\epsilon]}+\alpha e^{-m[D(q\|x)+\epsilon]}}$$

yielding

$$\begin{aligned}\Pr[\boldsymbol{K}\in\mathcal{B}_q|q]\ &\frac{1}{m}\log\frac{(1-\alpha)+\alpha e^{-m[D(q\|y)+\epsilon]}}{(1-\alpha)e^{-m[D(q\|p)-\epsilon]}+\alpha e^{-m[D(q\|x)-\epsilon]}}\\ \le&\frac{1}{m}\sum_{\boldsymbol{k}\in\mathcal{B}_q} Q(\boldsymbol{k})\Lambda(\boldsymbol{k})\\ \le&\frac{1}{m}\log\frac{(1-\alpha)+\alpha e^{-m[D(q\|y)-\epsilon]}}{(1-\alpha)e^{-m[D(q\|p)+\epsilon]}+\alpha e^{-m[D(q\|x)+\epsilon]}}.\end{aligned}\tag{30}$$

Observe now that

$$\lim_{m\longrightarrow\infty}\frac{1}{m}\log\frac{1}{(1-\alpha)e^{-m[D(q\|p)\pm\epsilon]}+\alpha e^{-m[D(q\|x)\pm\epsilon]}} = \begin{cases} D(q\|x)\pm\epsilon & D(q\|p)>D(q\|x)\\ D(q\|p)\pm\epsilon & D(q\|p)\le D(q\|x)\end{cases}\tag{31}$$

and

$$\lim_{m\longrightarrow\infty}\frac{1}{m}\log\left\{(1-\alpha)+\alpha e^{-m[D(q\|y)-\epsilon]}\right\}\le\epsilon.\tag{32}$$

Using (31) and (32) in the bounds of (30) yields

$$\frac{1}{m}\sum_{\boldsymbol{k}} Q(\boldsymbol{k})\Lambda(\boldsymbol{k})\longrightarrow\begin{cases} D(q\|x), & D(q\|p)>D(q\|x)\\ D(q\|p), & D(q\|p)\le D(q\|x).\end{cases}\tag{33}$$

The second addend at RHS of (28) can be managed by similar arguments, essentially exchanging the roles of Q and Y and introducing the typical set $\mathcal{B}_y := \mathcal{A}_\epsilon^{(m)}(y,x)\cap\mathcal{A}_\epsilon^{(m)}(y,p)\cap\mathcal{A}_\epsilon^{(m)}(y,q)$, in place of $\mathcal{B}_q$. The final result is

$$\frac{1}{m}\sum_{\boldsymbol{k}} Y(\boldsymbol{k})\Lambda(\boldsymbol{k})\longrightarrow\begin{cases} D(y\|x), & D(y\|p)>D(y\|x)\\ D(y\|p), & D(y\|p)\le D(y\|x).\end{cases}\tag{34}$$

It remains to combine the results (33) and (34) to compute $\lim_{m\longrightarrow\infty}(1/m)d_m(y;x)$, i.e.,

$$\lim_{m\longrightarrow\infty}\frac{1}{m}\left[(1-\alpha)\sum_{\boldsymbol{k}} Q(\boldsymbol{k})\Lambda(\boldsymbol{k})+\alpha\sum_{\boldsymbol{k}} Y(\boldsymbol{k})\Lambda(\boldsymbol{k})\right].$$

The value of this limit depends upon the particular combinations of the possible cases in (33) and (34), and is summarized in Table I (see main text).

From that table, we see that two possibilities exist for nullify $\lim_{m\longrightarrow\infty}(1/m)d_m(y;x)$ thus ensuring that $d_\infty(y;x)<\infty$. These are i) $x=y=q$, and ii) $x=q$, $y=p$. The best from the intruder's viewpoint is to select the one yielding a smaller value of $d_\infty(y;x)$. We now show that the optimal choice is the reversed emission strategy $x=q$, $y=p$. Indeed, the (finite m) divergence for this latter is

$$d_m(p;q)=\sum[(1-\alpha)Q(\boldsymbol{k})+\alpha P(\boldsymbol{k})]\log\frac{(1-\alpha)Q(\boldsymbol{k})+\alpha P(\boldsymbol{k})}{(1-\alpha)P(\boldsymbol{k})+\alpha Q(\boldsymbol{k})}$$

while, assuming $x=y=q$ gives

$$d_m(q;q)=\sum Q(\boldsymbol{k})\log\frac{Q(\boldsymbol{k})}{(1-\alpha)P(\boldsymbol{k})+\alpha Q(\boldsymbol{k})}.$$

It is expedient to define the auxiliary function

$$g(t)=\sum[(1-\alpha t)Q(\boldsymbol{k})+\alpha t P(\boldsymbol{k})]\log\frac{(1-\alpha t)Q(\boldsymbol{k})+\alpha t P(\boldsymbol{k})}{(1-\alpha)P(\boldsymbol{k})+\alpha Q(\boldsymbol{k})}$$

such that one can write $d_m(q;q)-d_m(p;q)=g(0)-g(1)$, and $d_m(q;q)>d_m(p;q)$ if $g(t)$ is a decreasing function for $t\in[0,1]$. This is in fact the case, indeed

$$\frac{dg}{dt}=\alpha\sum[P(\boldsymbol{k})-Q(\boldsymbol{k})]\log\frac{(1-\alpha t)Q(\boldsymbol{k})+\alpha t P(\boldsymbol{k})}{(1-\alpha)P(\boldsymbol{k})+\alpha Q(\boldsymbol{k})}.$$

Straightforward calculation reveals that, $\forall t\in[0,1]$ and $\alpha<1/2$, $(1-\alpha t)Q(\boldsymbol{k})+\alpha t P(\boldsymbol{k})>(1-\alpha)P(\boldsymbol{k})+\alpha Q(\boldsymbol{k})\Leftrightarrow Q(\boldsymbol{k})>P(\boldsymbol{k})$, thus $dg/dt<0$.

Having discovered that $d_m(p;q)$ is the lower exponent for any m, the last part of the proof amounts to compute its value in the limit of $m\longrightarrow\infty$. To this aim, we can redo basically the same steps from (28) on, specializing to the case that the likelihood ratio is as in (24). Furthermore, more or less obviously, the pertinent typical sets can be defined simply as $\mathcal{B}=\mathcal{B}_q=\mathcal{B}_y=\mathcal{A}_\epsilon^{(m)}(q,p)$.

As to the first term in (24), Lemma 2 implies

$$\left|\sum_{\boldsymbol{k}\notin\mathcal{B}} Q(\boldsymbol{k})\Lambda(\boldsymbol{k})\right|\le\Pr\{\boldsymbol{K}\notin\mathcal{B}|q\}\ \log\frac{1-\alpha}{\alpha}\longrightarrow 0$$

when $m\longrightarrow\infty$. In the set $\mathcal{B}$

$$\log\frac{(1-\alpha)+\alpha e^{-m[D(q\|p)+\epsilon]}}{(1-\alpha)e^{-m[D(q\|p)-\epsilon]}+\alpha}\le\Lambda(\boldsymbol{k})\le\log\frac{(1-\alpha)+\alpha e^{-m[D(q\|p)-\epsilon]}}{(1-\alpha)e^{-m[D(q\|p)+\epsilon]}+\alpha}$$

implying that

$$\Pr[\boldsymbol{K} \in \mathcal{B}|q] \ \log \frac{(1-\alpha)+\alpha e^{-m[D(q||p)+\epsilon]}}{(1-\alpha)e^{-m[D(q||p)-\epsilon]}+\alpha} \leq \sum_{\boldsymbol{k}\in\mathcal{B}} Q(\boldsymbol{k})\Lambda(\boldsymbol{k}) \leq \log \frac{(1-\alpha)+\alpha e^{-m[D(q||p)+\epsilon]}}{(1-\alpha)e^{-m[D(q||p)+\epsilon]}+\alpha}$$

[differently from (30) we do not divide by m]. This yields $\lim_{m\to\infty}\sum_{\boldsymbol{k}\in\mathcal{B}} Q(\boldsymbol{k})\Lambda(\boldsymbol{k}) = \log(1-\alpha/\alpha)$. Repeating for the term $\sum_{\boldsymbol{k}} P(\boldsymbol{k})\Lambda(\boldsymbol{k})$ is straightforward, and yields $d_\infty(p;q) = (1-\alpha)\log(1-\alpha/\alpha)+\alpha\log(\alpha/1-\alpha) = h(\alpha)$, which concludes the proof. •

REFERENCES

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

[2] P. Willett, P. Swaszek, and R. Blum, "The good, bad and ugly: Distributed detection of a known signal in dependent Gaussian noise," *IEEE Trans. Signal Process.*, vol. 48, pp. 3266–3279, Dec. 2000.

[3] P. K. Varshney, *Distributed Detection and Data Fusion*. New York: Springer, 1997.

[4] R. Viswanathan and P. K. Varshney, "Distributed detection with multiple sensors: Part I—Fundamentals," *Proc. IEEE*, vol. 85, pp. 54–63, Jan. 1997.

[5] R. S. Blum, A. Kassam, and H. V. Poor, "Distributed detection with multiple sensors: Part II—Advanced topics," *Proc. IEEE*, vol. 85, pp. 64–79, Jan. 1997.

[6] Z.-Q. Luo, M. Gastpar, J. Liu, and A. Swami, "Distributed signal processing in sensor networks," *IEEE Signal Process. Mag.*, vol. 23, no. 4, pp. 14–15, Jul. 2006.

[7] E. Shi and A. Perrig, "Designing secure sensor networks," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 38–41, Dec. 2004.

[8] F. Ye, H. Luo, S. Lu, and L. Zhang, "Statistical en-route filtering of injected false data in sensor networks," *IEEE J. Sel. Areas Commun.*, vol. 23, pp. 839–850, Apr. 2005.

[9] X. Luo, M. Dong, and Y. Huang, "On distributed fault-tolerant detection in wireless sensor networks," *IEEE Trans. Comput.*, vol. 55, pp. 58–70, Jan. 2006.

[10] T. Clouqueur, K. K. Saluja, and P. Ramanathan, "Fault tolerance in collaborative sensor networks for target detection," *IEEE Trans. Comput.*, vol. 53, pp. 320–333, Mar. 2004.

[11] W. Du, J. Deng, Y. Han, and P. Varshney, "A witness-based approach for data fusion assurance in wireless sensor networks," in *Proc. GLOBECOM*, 2003, pp. 1435–1439.

[12] L. Lamport, R. Shostak, and M. Pease, "The Byzantine generals problem," *ACM Trans. Program. Languages Syst.*, vol. 4, pp. 382–401, Jul. 1982.

[13] D. Dolev, "The Byzantine generals strike again," *J. Algorithms*, vol. 3, no. 1, pp. 14–30, 1982.

[14] B. Pfitzmann and M. Waidner, Information Theoretic Pseudosignatures and Byzantine Agreement for $t \geq n/3$ 1996, IBM Research Report, Tech. Rep. RZ2882.

[15] T. Ho, B. Leong, R. Koetter, M. Médard, M. Effrons, and D. Karger, "Byzantine modification detection in multicast networks using randomized network coding," in *IEEE Proc. Int. Symp. Inf. Theory*, Jun. 2, 2004, pp. 143–143.

[16] O. Kosut and L. Tong, "Distributed source coding with Byzantine sensors," *IEEE Trans. Inf. Theory*, vol. 54, 2008.

[17] S. A. Kassam and H. V. Poor, "Robust techniques for signal processing: A survey," *Proc. IEEE*, vol. 73, pp. 433–481, Mar. 1985.

[18] P. J. Huber, "A robust version of the probability ratio test," *Ann. Math. Statist.*, vol. 36, no. 6, pp. 1753–1758, Dec. 1965.

[19] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[20] P. J. Huber, *Robust Statistics*. New York: Wiley, 1981.

[21] I. Csiszár and C. Tusnady, "Information geometry and alternating minimization procedures," in *Statistics Decisions, Suppl. Issue 1*, 1984, pp. 205–237.

[22] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K. : Cambridge Univ. Press, 2004.

Stefano Marano received the Laurea degree in electronic engineering (*cum laude*) and the Ph.D. degree in electronic engineering and computer science both from the University of Naples, Naples, Italy, in 1993 and 1997, respectively.

Currently, he is a Professor with the University of Salerno, Fisciano, Italy, where he formerly served as Assistant Professor. His areas of interest include statistical signal processing with emphasis on distributed inference, sensor networks, and information theory. He published about 80 papers on these and related topics, including several invited, mainly in top international journals/transactions and proceedings of international conferences. He has also given several invited talks in the area of statistical signal processing.

Prof. Marano was awarded the IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION 1999 Best Paper Award (jointly with G. Franceschetti and F. Palmieri) for his work on stochastic modeling of electromagnetic propagation in urban areas. As a reviewer, he handled hundreds of papers, mainly for the IEEE TRANSACTIONS, and was selected as Appreciated Reviewer by the IEEE TRANSACTIONS ON SIGNAL PROCESSING, in 2007. He has been on the Organizing Committee of some top international conferences in the field of signal processing and data fusion, as well as in the Technical Program Committee of many international symposia.

Vincenzo Matta received the Laurea degree in electronic engineering and the Ph.D. degree in information engineering from the University of Salerno, Fisciano, Italy, in 2001 and 2005, respectively.

He is currently an Assistant Professor with the University of Salerno. His main research interests include detection and estimation theory, signal processing, wireless communications, multiterminal inference, and sensor networks.

Lang Tong (F'05), photograph and biography not available at the time of publication.