# Distributed Inference in the Presence of Byzantine Sensors

*(Invited Paper)*

Stefano Marano, Vincenzo Matta
Dept. of Information & Electrical Engineering (DIIIE),
University of Salerno,
via Ponte don Melillo I-84084, Fisciano (SA), Italy.
E-mails: {marano, vmatta}@unisa.it

Lang Tong
Dept. of Electrical & Computer Engineering (ECE),
Cornell University,
Ithaca, NY 14853 USA.
E-mail: ltong@ece.cornell.edu

***Abstract*— A wireless sensor network designed for distributed detection undergoes a Byzantine attack in which a fraction of sensors cooperatively transmit fictitious signals to impair the detection capability of the fusion center. The optimal attacking distributions are derived and the fundamental tradeoff between detection power (best asymptotic exponent of the miss detection error probability) and the attacking power (fraction of compromised sensors) is characterized.**

**Also considered is a hierarchical network made of $m$-sensors clusters. For large $m$ the optimal miss detection error exponent is found to be a binary divergence, and the asymptotic performance is shown to scale with the number $n$ of clusters but, remarkably, not with the cluster size $m$. The optimal test, in this case, reduces to a kind of infected-cluster counting.**

*Index Terms*—**Distributed detection, Sensor networks, Byzantine attack, Network defense.**

## I. Introduction

We consider a large Wireless Sensor Network (WSN) engaged in the task of distributed binary detection. The network consists of $n$ nodes or sensors, each making an independent and identically distributed (iid) observation about the State of the Nature (say $\mathcal{H}_1$ or $\mathcal{H}_0$). These observations are successively delivered to a common fusion center (parallel architecture) for the final decision on the underlying statistical hypothesis. Actually the network is under attack: a clique of traitorous sensors cooperatively works against the network. These sensors, referred to as the Byzantines (and the kind of attack described is then called Byzantine attack [1]), deliver data according to certain fictitious distributions properly designed in order to impair the detection capability of the fusion center. The Byzantines are assumed to *know* the true underlying hypothesis; the uninfected and FC, obviously, do not. The fusion center, however, is aware of the presence of the Byzantines. Specifically, it knows that a fraction $\alpha$ of the sensors are traitorous and will deliver data drawn according to the optimal (from the Byzantine viewpoint) attacking distributions. As a consequence, the decision rule implemented at the fusion center is a Neyman-Pearson test that do account for the fraction of fictitious data. The addressed problem is schematically depicted in Fig. 1. Note that the described strategy is more sophisticated than the naivest *black-hole* attack, in which the intruder simply destroys the owned fraction of sensors.

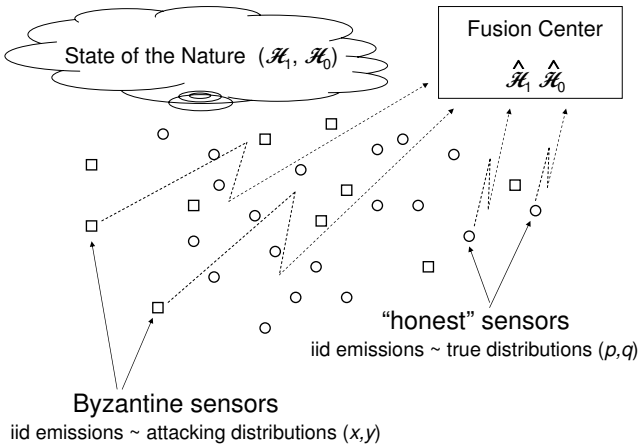


Fig. 1. Conceptual scheme of the addressed problem. A sensor network designed for binary hypothesis testing undergoes a Byzantine attack where a fraction $\alpha$ of nodes cooperatively conspire with the aim of impairing the detection capabilities. The Byzantine sensor deliver to the fusion center samples drawn iid from certain attacking distributions, in place of the observations ruled by the underlying State of the Nature.

In the above we implicitly referred to a network where each node makes a single observation about the State of the Nature. Also considered in this work is a hierarchical network in which $m$ individual sensors form a cluster, and there are $n$ different clusters. In this scenario, either all the $m$ sensors of a cluster are Byzantine or all are honest, and the sensors shown in Fig. 1 become clusters of sensors, with each cluster delivering to the fusion center a vector of $m$ samples.

In the two described system architectures we address the following basic questions. What are the optimal attacking distributions that the Byzantine will employ? What is the resulting test performance? What about the minimum fraction

of traitorous sensors/clusters such that the network becomes useless?

General network security is widely considered in the literature, see *e.g.*, [2], while less investigated is the the topic of secure sensor networks for distributed detection and data fusion [3]–[5]. Relevant to our approach are also [6]–[8]. For an entry point to the notion of the Byzantine general problem see [1]. A related information theoretical view of Byzantine attacks in wireless sensor networks is provided in [9], where the focus is on the capacity of collaborative fusion. The presence of misinformed nodes is instead dealt with in [10].

The paper is organized as follows. In the next section the problem is formalized and answers to the stated questions are provided. Extension to the hierarchical architecture is dealt with in Sect. III, while final comments are provided in Sect. IV.

## II. Problem statement & Solution

Let $n$ be the number of sensors in the network and assume that a fraction $\alpha$ of these are cooperatively traitorous, *i.e.*, Byzantine sensors. The statistical hypothesis test can be formulated as

$$\begin{array}{rcl} \mathcal{H}_0 & : & \Pr\{K^{(j)}=k\} = z_k := (1-\alpha)q_k + \alpha y_k, \\ \mathcal{H}_1 & : & \Pr\{K^{(j)}=k\} = w_k := (1-\alpha)p_k + \alpha x_k, \end{array} \tag{1}$$

where $j$ spans the sensor class, $K^{(j)}$ is the scalar observation made at the $j^{th}$ node (observations are assumed iid under both the hypotheses), $k \in \mathcal{K} := \{0,1,2,\ldots,|\mathcal{K}|-1\}$, and $p$ $q$, $x$ and $y$ are probability mass functions defined over $\mathcal{K}$.

The Byzantine sensors deliver to the FC fictitious samples drawn iid from suitably chosen attacking distributions $x$ and $y$ in the attempt of worsening the network performance. As our focus is on *large* sensor networks, this latter is here measured in terms of the Kullback-Leibler divergence $D(z||w)$; we also use the symbol $d(y;x)$ to denote such a divergence, where the dependence on the attacking distributions is emphasized. The fusion center collects the nodes observations, with no possibility of distinguishing between fair and fictitious, and finally implements a Neyman-Pearson test between the actual pmfs $w$ and $z$, being aware of the presence of the Byzantines.

The final exponent of the test (asymptotic miss detection error rate) will be the minimum of the divergence between the two hypotheses

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

where the pmfs $(x,y)$ attaining such minimum are the attacking distributions employed by the intruder, as characterized in the next theorem.

Let us define the *blinding power*

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

where $(c)^+$ stands for $\max\{0,c\}$. The following result can be proven.

**Theorem 1**

$(i)$ *For $\alpha \geqslant \alpha_b$, $\Delta(\alpha) = 0$.*

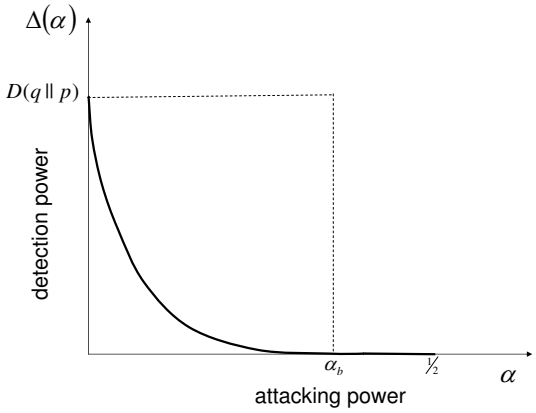


Fig. 2. The fundamental tradeoff between the detection power of the network $\Delta(\alpha)$ and the attacking power of the intruder $\alpha$. For $\alpha = 0$ no Byzantine sensors exist. By increasing $\alpha$, the fraction of traitorous nodes, the conspiracy takes place and the detection capabilities of the network are accordingly reduced. At $\alpha = \alpha_b$ the intruder blinds the system, and $\Delta(\alpha)$ falls to zero.

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

$$\left\{ \begin{array}{rcl} x_k & = & \frac{1-\alpha}{\alpha}(q_k - p_k)^+ , \\ y_k & = & \frac{1-\alpha}{\alpha}(p_k - q_k)^+ . \end{array} \right. \tag{3}$$

*If $\alpha > \alpha_b$, there exist infinitely many solutions $(x,y)$ that nullifies the test exponent. These are obtained by starting with $x_k = \frac{1-\alpha}{\alpha}(q_k - p_k)^+$ (which is not a pmf) and increasing arbitrarily some of the $x_k$'s until $x$ becomes a pmf; the corresponding $y$ is then obtained as $y_k = x_k + \frac{1-\alpha}{\alpha}(p_k - q_k)$.*

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

$$\left\{ \begin{array}{l} x_k = \dfrac{1-\alpha}{\alpha}\ (\gamma_x q_k - p_k)^+ , \\ y_k = \dfrac{1-\alpha}{\alpha}\ (\gamma_y p_k - q_k)^+ , \end{array} \right. \tag{4}$$

*where $0 < \gamma_x, \gamma_y \leqslant 1$ 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 $\Delta(0) = D(q||p)$ and $\lim_{\alpha \to \alpha_b} \Delta(\alpha) = 0$.* △

*Proof:* Provided in [11].

Part $(i)$ of the above theorem, along with the condition $\alpha_b \leqslant 1/2$, implies that the divergence can be always nullified, provided that the fraction of Byzantine sensors exceeds $50\%$. In this case no meaningful inference about the surrounding hypothesis can be made, and we say that the attack has completely blinded the network. The system is useless.

Solution (3) in part $(ii)$ is actually a special case of that in eqs. (4), obtained with $\gamma_x = \gamma_y = 1$.

As to statement $(iii)$, it can be shown that the likelihood ratio test corresponding to solutions (4) reduces to a censored

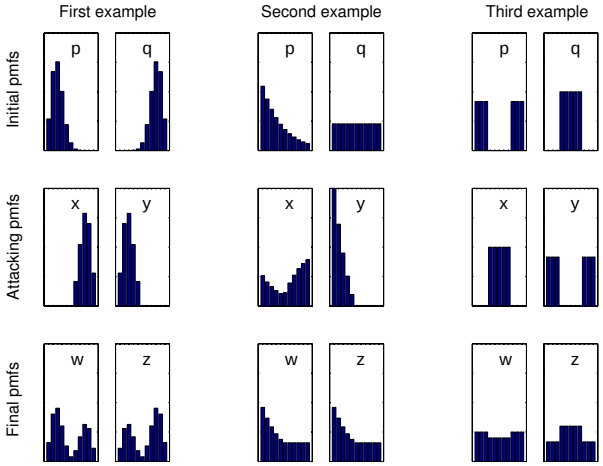


Fig. 3. Examples of application of the theory. See the main text for details.

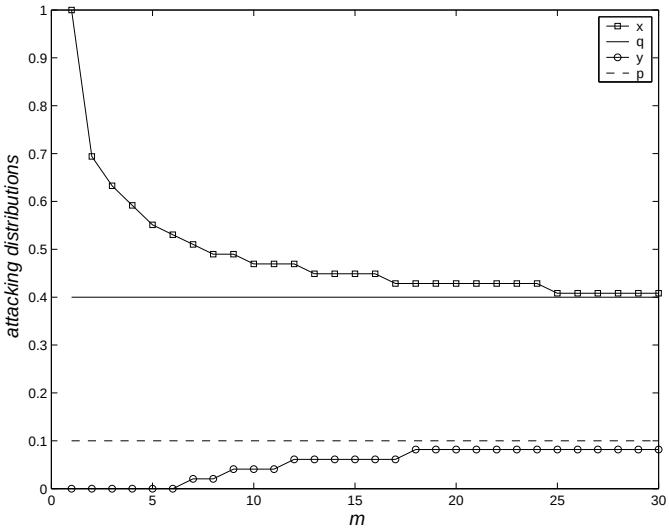


Fig. 4. Binary attacking distributions versus the cluster size $m$, in the case that $p = [0.1, 0.9]$, $q = [0.4, 0.6]$, and $\alpha \approx 0.19 < \alpha_b$. The shown curves refer to $x_0$, $y_0$, $q_0$, and $p_0$. We know that at $m = 1$, $x_0 = 1$ and $y_0 = 0$ (from eq. (4)). We also know that, at large $m$, $x_0 \to q_0$ and $y_0 \to p_0$ (hypothesis-reversed attack, see Theorem 2).

version of the original test between $p$ and $q$. This is reminiscent of some statistical tests arising in the context of robust detection, and in fact $w$ and $z$ are nothing but $\alpha$-contaminated mixtures, and $w$ and $z$, accounting for eqs. (4), are nothing but the least favorable distributions of the test [12]–[14]. This also implies that a Byzantine change of the attacking distributions when the network is not aware of this (*i.e.*, when it still implements the censored test), cannot provide any advantage from the intruder's viewpoint.

According to part $(iv)$ of the theorem, a typical shape of $\Delta(\alpha)$ is depicted in Fig. 2. Note that the straight line $(1-\alpha)D(q||p)$ corresponding to a *black-hole* attack, lies always above $\Delta(\alpha)$, confirming the higher power of the Byzantine conspiracy.

Figure 3 gives three examples of applications. We are given the initial distributions $p$ and $q$ and the attack power $\alpha$. Then, the attacking distributions $x$ and $y$, can be computed as indicated by Theorem 1. These are depicted in Fig. 3; also depicted are the final distributions, used at the fusion center to implement the Neyman-Pearson test. In the first example $p$ and $q$ are such that $\alpha_b \approx 0.49$. If we assume that the intruder power is $\alpha = 0.4$, the attacking distributions and the final pmfs are as shown in the figure. The divergence between the hypotheses is $D(q||p) \approx 8.3178$ nats while, as consequence of the attack, the final divergence between $z$ and $w$ reduces to $\Delta(\alpha) \approx 6.4\ 10^{-2}$ nats. In the second example the divergence can be nullified. There, in fact, we set $\alpha = 0.3$, while $\alpha_b \approx 0.22$. We start from $D(q||p) \approx 0.237$ and the result of the Byzantine attack is $\Delta(\alpha) = 0$. The last examples deals with a singular detection problem ($D(q||p) = \infty$) and the attack leads to $\Delta(\alpha) \approx 8.1\ \ 10^{-2}$ nats. In this case $0.4 = \alpha < \alpha_b = 1/2$.

## III. Byzantine attacks to clustered networks

In some applications sensor networks may be clustered, in the sense that several individual nodes are somehow aggregated (*e.g.*, closely deployed) to form a supernode, or cluster. We now assume that the sensors of the network are grouped $m$ by $m$ to form a clustered architecture of neighborhood nodes. One implications of the spatial proximity is that either all the $m$ nodes of a cluster are traitorous or none is. At the same time, sensors' closeness is not enough to produce significant correlation among the observations made at the $m$ nodes of the same cluster.

Let $\boldsymbol{K}^{(j)}$ be the $m$-dimensional vector of per-cluster observations, and let $W$ and $Z$ be its statistical distributions under $\mathcal{H}_1$ and $\mathcal{H}_0$, respectively. Define also $Q(\boldsymbol{k}) = \prod_{i=1}^m q_{k_i}$ $Y(\boldsymbol{k}) = \prod_{i=1}^m y_{k_i}$, $P(\boldsymbol{k}) = \prod_{i=1}^m p_{k_i}$ and $X(\boldsymbol{k}) = \prod_{i=1}^m x_{k_i}$.

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

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

Let us denote with $d_m(y;x)$ the Kullback-Leibler distance between $Z$ and $W$, thus emphasizing the dependence upon the *marginal* distributions $x$ and $y$. Also, let $\Delta_m(\alpha) := \min_{x,y} d_m(y;x)$ be the worst test exponent forced by the Byzantine sensors when the emissions of these latter are regulated by the optimal attacking distributions. In what follows we always assume $m > 2$, and some attention is payed to the behaviors of $d_\infty(y;x)$ and $\Delta_\infty(\alpha)$ describing the asymptotic case of $m \to \infty$.

We have the following result, in which $h(\alpha)$ is the KL distance between the binary pmfs $[1-\alpha, \alpha]$ and $[\alpha, 1-\alpha]$.

**Theorem 2**

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

$(ii)$ *For any $\alpha < 1/2$, $\Delta_m(\alpha)$ 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.* $\triangle$

*Proof:* Provided in [11].

Note that part $(ii)$ of the above theorem implies that, differently from the scenario considered in Theorem 1, an attack based on less than $50\%$ of the nodes can never impair the system.

On the other hand, part $(iii)$ tells us that infinitely many observations (per cluster) do not imply error-free decision. At first glance, perhaps one would expect $d_m(y; x)$ to scale linearly with the number of local observations $m$, and therefore $d_\infty(y; x) = \infty$ (error-free, singular test). Instead, there exist attacking distributions $x$ and $y$ such that $d_\infty(y; x) < \infty$. The traitorous Byzantines, obviously, will choose that.

This admits an intuitive explanation. Consider the ensemble of $m$ samples collected by a generic cluster and made available to the FC. Assume further that $m$ is so large that the FC can *exactly* know whether these samples come from $p$ or from $q$. For instance, if they are known to be drawn from $p$ then, in view of the reversed emission strategy, this lead to the conclusion that *either* the true hypothesis is $\mathcal{H}_1$ and the cluster is honest (this happens with probability $1 - \alpha$) *or* the true hypothesis is $\mathcal{H}_0$ and the cluster is infected (with probability $\alpha$). Similarly, if data comes from $q$, then we either have a honest cluster and $\mathcal{H}_0$ is true (probability $\alpha$) or the cluster is infected and $\mathcal{H}_1$ is true (probability $1 - \alpha$). Exploiting the many clusters of the network, the FC implements a test whose asymptotic miss detection error exponent is just $h(\alpha)$.

Comparing the results of Theorems 1 and 2 in the simple scenario of binary alphabets, we see that by increasing $m$ the optimal attacking distributions move from a deterministic delivering $x_0 = 1$, $y_0 = 0$ (which is the best for the case of $m = 1$, as a simple derivation from eq. (4) reveals) to the asymptotically optimum hypothesis-reversed emission strategy, *i.e.*, $x_0 = q_0$, $y_0 = p_0$. This is illustrated in Fig. 4.

## IV. Conclusions

A sensor network designed for distributed detection and subject to a Byzantine attack has been considered, with two system architectures addressed: a network made of individual sensors, and a hierarchical structure with groups of $m$ sensors tied together to form a cluster.

For the former architecture we show that if more than $50\%$ of the nodes are Byzantine, the attack can always destroy any detection capability: the network become useless. Actually this turns out to be true provided that the fraction of traitorous nodes exceeds a quantity that we call the blinding power $\alpha_b$ (which is always smaller than or equal to $1/2$).

The fundamental tradeoff between detection capability and attacking power is characterized, the optimal attacking probability laws are derived, and the decision rule implemented at the fusion center turns out to be a censored likelihood ratio test. This bears similarities to Huber's robust statistics.

In the clustered network, this analogy breaks down and different behaviors arise. In fact, attackers owning less than one half of the total sensors cannot completely impair the system. They can, however, severely degrade the performance of the network, and the optimal attacking distributions that achieve this goal are "hypothesis-reversed": the Byzantine emissions are drawn from the distribution corresponding to the *false* State of the Nature. A remarkable fact is that the asymptotic detection probability does not scale exponentially with the cluster size. Actually, it does not scale at all with $m$. The practical consequence is a saturation effect: increasing the number of per-cluster sensors beyond a certain amount does not provide any significant performance improvement. On the other hand, the expected scaling law is instead preserved with respect to the total number of clusters.

## References


[1] L. Lamport, R. Shostak, and M. Pease, "The Byzantine generals problem," *ACM Transactions on Programming Languages and Systems*, vol. 4, no. 3, pp. 382–401, July 1982.

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

[3] F. Ye, H. Luo, S. Lu, and L. Zhang, "Statistical en-route filtering of injected false data in sensor networks," *IEEE J. Select. Areas Commun.*, vol. 23, no. 4, pp. 839–850, Apr. 2005.

[4] X. Luo, M. Dong, and Y. Huang, "On distributed fault-tolerant detection in wireless sensor networks," *IEEE Trans. Comput.*, vol. 55, no. 1, pp. 58–70, Jan. 2006.

[5] T. Clouqueur, K. K. Saluja, and P. Ramanathan, "Fault tolerance in collaborative sensor networks for target detection," *IEEE Trans. Comput.*, vol. 53, no. 3, pp. 320–333, Mar. 2004.

[6] 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.

[7] A. Wood and J. Stankovic, "Denial of service in sensor networks," *IEEE Computer*, pp. 54–62, 2002.

[8] C. Karlof and D. Wagner, "Secure routing in wireless sensor networks: Attacks and countermeasures," in *Proc. IEEE Intl. Workshop on Sensor Network Protocols and Applications*, May 2003.

[9] O. Kosut and L. Tong, "Capacity of cooperative fusion in the presence of Byzantine sensors," in *Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, September 27-29 2006.

[10] Z. Yang and L. Tong, "Cooperative sensor networks with misinformed nodes," *IEEE Trans. Inform. Theory*, vol. IT-51, no. 12, pp. 4118–4133, Dec. 2005.

[11] S. Marano, V. Matta, and L. Tong, "Distributed detection by large wireless sensor networks in the presence of Byzantine attacks," in preparation.

[12] S. A. Kassam and H. V. Poor, "Robust techniques for signal processing: A survey," *Proc. IEEE*, vol. 73, no. 3, pp. 433–481, Mar. 1985.

[13] P. J. Huber, "A robust version of the probability ratio test," *Ann. Math. Statist.*, vol. 36, pp. 1753–1758, Dec. 1965.

[14] ——, *Robust Statistics*. New York: John Wiley & Sons, 1981.