Multiple-Observation Hypothesis Testing under Adversarial Conditions

Mauro Barni [1], Benedetta Tondi [2]

Department of Information Engineering and Mathematics, University of Siena
Via Roma 56, 53100 - Siena, ITALY
[1] barni@dii.unisi.it
[2] benedettatondi@gmail.com

***Abstract*—We address the problem of binary hypothesis testing based on multiple observations in the presence of an adversary corrupting part or all the observations. We propose a general framework based on game-theory that encompasses a wide variety of situations including distributed detection, data fusion, multimedia forensics, sensor networks. The proposed approach extends the Neyman-Pearson approach to an adversarial setting in which the analyst must ensure that type I error probability stays below a threshold, and the adversary tries to induce a type II error. We derive the equilibrium point of the game in an asymptotic set up, showing that a dominant strategy exists for the analyst. The paper opens the way to further analysis in which the payoff of the game at the equilibrium is analyzed thus permitting to understand the ultimate achievable performance of multiple-observation hypothesis testing under adversarial conditions.**

I. Introduction

Adversarial Signal Processing (Adv-SP), i.e. the study of signal processing techniques explicitly thought to withstand the attacks of one or more adversaries aiming at system failure, is receiving an increasing attention due to its applicability in a wide number of scenarios, including multimedia forensics, biometrics, digital watermarking, steganography and steganalysis, network intrusion detection, traffic monitoring, video-surveillance, just to mention a few [1]. Adversary-aware hypothesis testing (or binary decision) is undoubtedly one of the most common problems in Adv-SP, due to its importance in several application scenarios. In multimedia forensics, for instance, the analyst has to decide whether a document has been generated by a given source (a specific camera or a camera model), or has undergone a given processing. In spam filtering, e-mail messages have to be classified either as spam or authentic messages. In 1-bit watermarking, the detector has to decide whether a document is watermarked or not, while it is the goal of steganalysis to distinguish between cover and stego-images. In yet other situations, the security of a system relies on the capability of distinguishing malevolent users from fair ones (again a binary classification problem). Even if specific solutions have been proposed for each of the above problems, the need for a general theoretical framework that models the interplay between the analyst and the attacker is becoming evident [1], [2]. In [3], [4], a game-theoretic framework has been introduced to analyze the hypothesis testing problem under adversarial conditions. By assuming that the analyst can rely only on first order statistics of the observables and that the attacker has to satisfy a distortion constraint, the asymptotic equilibrium point of the game is derived when the length of the observed sequence tends to infinity.

In this paper, we extend the framework introduced in [3] to deal with binary hypothesis testing under multiple observations. This is a relevant scenarios in several applications, including multimedia forensics [5], data fusion, distributed hypothesis testing and detection [6], sensor networks [7] and cognitive radio networks [8]. In all these cases, a fusion center has to take a binary decision about the status of a system by relying on a number of observations made available by different sensors (as in [6]) or a number of traces detected by different investigation tools (as in [5]). In many situations, it is possible, actually probable, that an attacker (or more attackers) corrupts the observations or deliberately provide misleading data to induce a decision error at the fusion center. It is the goal of this paper to introduce a general information-theoretic framework to analyze the above situations and devise the optimal strategies for both the analyst and the attacker in a game-theoretic sense, that is by determining the equilibrium point of the game. We will do so for several versions of the game, thus encompassing a large number of scenarios addressing many diverse applications. The rest of this paper is organized as follows. In Section II, we introduce the Multiple-Observation Hypothesis Testing setup. In Section III, we adopt the point of view of the analyst and derive the optimum decision strategy. In Section IV, we consider the optimum attacking strategy. The main results of the paper are summarized and discussed in Section V.

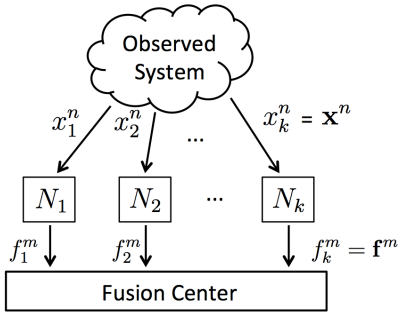

Fig. 1. The multiple-observation hypothesis testing setup.

II. The setup

A sketch of the Multiple-Observation Hypothesis Testing (MO-HT) setup considered in this paper is reported in Fig. 1. The status of a system is observed by k nodes which gather k observation sequences, $x_1^n, x_2^n \dots x_k^n$, each of which consists of n samples, i.e., $x_l^n = (x_{l,1}, x_{l,2} \dots x_{l,n}), l = 1 \dots k$. The nodes summarize their observations into k feature sequences of length m ($m \leq n$), $f_1^m, f_2^m \dots f_k^m$, with $f_l^m = (f_{l,1}, f_{l,2} \dots f_{l,m}), l = 1 \dots k$. The summaries are sent to a fusion center which has to either accept or reject a certain hypothesis H_0 about the status of the system. This is a very general setup that can be used to model a wide variety of situations. The most obvious application regards distributed hypothesis testing [6]. As an example, the nodes may be part of a sensor network and the observed sequences $x_1^n \dots x_k^n$ may describe the physical state of the system over time. e.g., the temperature,

WIFS'2013, November 18-21, 2013, Guangzhou, China.
ISBN 978-1-4673-5593-3

measured at different locations. In more complex situations, the observed sequences may correspond to complex signals like a video or an audio sequence. As to the summaries, in the simplest case they coincide with the observed sequences. More often, they are obtained by extracting a number of features from the observed sequences, or by taking a local decision on the system status. In the latter case, $m = 1$ and $f_l^m = 0$ or 1 depending on the local decision on the validity of hypothesis H_0 taken by node l.

A less obvious instantiation of the setup reported in Fig. 1 regards the use of data fusion techniques for multimedia forensics. In this case, the observed system is a document, for instance an image or a video, which is analyzed by means of different tools (identified here by $N_1, N_2 \dots N_k$). Each tool analyzes a different aspect of the document. In the case of still images, for instance, the tools may analyze different color bands, or different frequency coefficients, in the case of video, the observables may refer to the audio and video tracks and so on. The tools extract a number of features and send them to a data fusion center, that is in charge of taking the final decision on a certain aspect of the analyzed document (e.g. its origin). As in the distributed hypothesis testing scenario, two extreme cases are obtained when the features correspond to the entire set of observables, and when each tool takes a local decision and the fusion is carried out at the decision level.

When MO-HT is framed in an adversarial setting, we must take into account the possibility that an adversary corrupts part of the system so to induce a decision error. In this paper we consider two main possibilities. As a first case, we assume that the attacker corrupts h out of k summaries. This is possible if the attacker seizes h nodes or if he controls h links between the nodes and the fusion center (see for instance [9]). Two sub-cases are possible depending on whether the attacker can choose which nodes he is going to attack or not. For the rest, we do not put any further limitation on the attacker's actions. In the following, we will refer to this setting as *MO-HT with (chosen) corrupted nodes*[1]. In a second scenario, the nodes and the links between the nodes and the fusion center are under the full control of the analyst and hence the attacker can only modify h out of k observed sequences. This is typically the case in applications wherein the system is analyzed from different points of view by using different analysis tools and the decision on system status is taken by fusing the output of the tools. As an example, we mention data fusion for multimedia forensics analysis, in which an analyst studies various aspects of the document at hand, and takes a decision on the provenance or integrity of the document by fusing the results of the different analyzes. The attacker, on his side, modifies the document so to hide its true origin or its previous history. In these cases, it makes sense to require that the amount of modification the attacker can introduce into the document is limited. In the following, we will refer to this scenario as *MO-HT with corrupted observations*. A graphical representation of the two kinds of attacks is given in Fig. 2.

Several versions of the two general settings described above are obtained depending on the actions allowed to the attacker and the analyst, their specific goals, the knowledge they have about the system, including its status and its statistical characterization, the knowledge that the attacker has on the links and nodes that he does not control and so on. In the next sections, we will analyze some

[1]In principle we should distinguish between an adversary that takes full control of the nodes and an adversary that controls only the links between the nodes and the fusion center, since in the former case the attacker can observe the sequences x_l^n of the corrupted nodes, thus acquiring information about the system status. In this paper we consider an omniscient attacker, hence making the distinction between the two cases irrelevant.

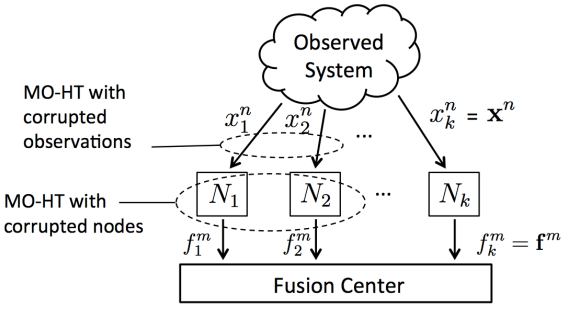

Fig. 2. Multiple-observation hypothesis testing under adversarial conditions.

of these variants, by framing them into a rigorous game-theoretic setting. As we will see, game-theory provides a natural and flexible way to take into account all the above information and to study the optimal strategies of the two players in terms of game equilibrium and achievable payoff.

III. Dominant fusion strategies for the defender

As anticipated, we use game-theory to give a formal definition of the MO-HT problems outlined in the previous section. In this section we adopt the perspective of the analyst, hereafter referred to as the Defender (D), defining his goals, his possible actions and deriving the optimum fusion strategies under some general assumptions.

A. Game-theory in a nutshell.

A 2-player game can be seen as a 4-uple $G(\mathcal{S}_1, \mathcal{S}_2, u_1, u_2)$, where $\mathcal{S}_1 = \{s_{1,1} \dots s_{1,n_1}\}$ and $\mathcal{S}_2 = \{s_{2,1} \dots s_{2,n_2}\}$ are the set of strategies the first and the second player can choose from, and $u_l(s_{1,i}, s_{2,j}), l = 1, 2$, is the payoff of the game for player l, when the first player chooses the strategy $s_{1,i}$ and the second chooses $s_{2,j}$. A pair of strategies $(s_{1,i}, s_{2,j})$ is called a profile. When $u_1(s_{s1,i}, s_{2,j}) + u_2(s_{1,i}, s_{2,j}) = 0$, the win of a player is equal to the loss of the other and the game is said to be a zero-sum game. In the set-up adopted in this paper, $\mathcal{S}_1$, $\mathcal{S}_2$ and the payoff functions are assumed to be known to the two players. In addition, we assume that the players choose their strategies before starting the game without knowing the strategy chosen by the other player (strategic game).

A common goal in game theory is to determine the existence of equilibrium points, i.e. profiles that in *some sense* represent a *satisfactory* choice for both players [10]. The most famous equilibrium notion is due to Nash. Intuitively, a profile is a Nash equilibrium if each player does not have any interest in changing its choice assuming the other does not change its strategy. Despite its popularity, the practical meaning of Nash equilibrium is doubtful, since there is no guarantee that the players will end up playing at the equilibrium. A notion with a more practical meaning is that of dominant equilibrium. A strategy is said to be strictly dominant for one player if it is the best strategy for the player, no matter how the other player decides to play. In many cases dominant strategies do not exist, however when one such strategy exists for one of the players, he will surely adopt it (at least under the assumption of rational behavior). The other player, in turn, will choose his strategy anticipating that the first player will play the dominant strategy. It is then easy to see that when a dominant strategy exists, the players have only one rational choice called the only rationalizable equilibrium of the game [11]. Games with the above property are called *dominance solvable* games.

In the rest of this section we consider three versions of the MO-HT game, by focusing on the strategy and payoff of the defender. As we will see, in our setup a dominant strategy exists for the defender (i.e., the games are dominance solvable), hence opening the way to

the derivation of the equilibrium point of the game (see Section IV). Such results are summarized in Theorems 1 through 3 in the sections below. Due to lack of space we report only the proof of Theorem 3, since in our opinion this is the most original proof among the three. The reader may get a feeling about the way the other proofs work by referring to the proofs of the Theorems in [3].

B. Notation and definitions

In our framework the system is modeled by a vector of discrete[2] random variables $\mathbf{X} = X_1, X_2 \dots X_k$ taking values in the same alphabet $\mathcal{X}$. Being related to the same system, the random variables are not independent and hence they are described by means of the joint probability mass function (pmf), say $P_{\mathbf{X}}(x_1, x_2 \dots x_k) = P_{\mathbf{X}}(\mathbf{x})$.

Our analysis relies on the concepts of type and type class defined as follows (see [12] and [13] for more details). Given a sequence x^n with elements belonging to an alphabet $\mathcal{X}$, the type P_{x^n} of x^n is the empirical pmf induced by the sequence x^n. In the following, we indicate with $\mathcal{P}_n$ the set of types with denominator n, i.e. the set of types induced by sequences of length n. Given $P \in \mathcal{P}_n$, we indicate with $T(P)$ the type class of P, i.e. the set of all the sequences in $\mathcal{X}^n$ having type P. Being interested in vector sequences, we will also use the vector extension of the above definitions. By considering, for instance, the observation vectors, we indicate by $\mathbf{x}_i = (x_{1,i}, x_{2,i} \dots x_{k,i})$ the vector with the observations of all the nodes at the time instant i, and with $\mathbf{x}^n = \mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x_n}$ the sequence with all the observed vectors $\mathbf{x}_i$. We then use the notation $P_{\mathbf{x}^n}$ to indicate the empirical joint pmf (the type) induced by the sequence $\mathbf{x}^n$ and with $T(P)$ the type class with all the vector sequences having the empirical pmf equal to P. Finally, we indicate with $\boldsymbol{\mathcal{P}}_n$ all the types for vector sequence of size k and length n.

Throughout the paper, we adopt a Neyman-Pearson perspective according to which D is interested to accept or reject the hypothesis H_0 that the state is in a safe or normal condition characterized by a pmf $P_{\mathbf{X}}$. In doing so D must ensure that the false positive error probability (P_{fp}) of rejecting H_0 when H_0 holds stays below a threshold. On his side, the attacker aims at inducing a type II error, i.e. to hide the fact that the system exited its normal status. We indicate by $P_{\mathbf{Y}}$ the pmf when H_0 does not hold (H_1). As in [3], we consider an asymptotic version of the problem (by letting n go to infinity) and require that P_{fp} decays exponentially fast with error exponent at least equal to λ. In addition, we force D to rely on first order statistics only, i.e. to neglect the possible dependence between consecutive observations (this assumption is sometimes referred to as limited resources assumption [3]).

C. MO-HT with full knowledge

As a first scenario, we consider a simplified case in which the nodes take the observed sequences and pass them to the data fusion center as they are, i.e., $f_l^m = x_l^n, \forall l$. Even if the above condition is rarely verified in practice, this scenario represents a kind of most favorable case for the defender since he can base his decision on all the available information. In addition, the analysis is rather simple since it is a straightforward extension of the game considered in [3]. In the following, we will refer to this scenario as the MO-HT game with full knowledge. Let us, then, define the strategies and payoff of the defender. Mimicking the Neyman-Pearson approach to hypothesis testing, the possible strategies for D are all the possible acceptance regions ensuring a given false positive error probability. In formula:

$$\mathcal{S}_D = \{\Lambda_0 \in 2^{\boldsymbol{\mathcal{P}}_n} \text{ s.t. } P_{fp} \leq 2^{-\lambda n}\}, \tag{1}$$

[2]Rigorously speaking our analysis is valid only for discrete random variables, the case of continuous variables, however, can be treated by quantizing the continuous alphabet at a resolution which is fine enough.

where Λ_0 is seen as a union of types (a subset of the power set of $\boldsymbol{\mathcal{P}}_n$) due to the limited resources assumption. Thanks to this assumption, in fact, if a vector sequence stays in Λ_0, all the other sequences in the same type class must belong to Λ_0, hence permitting to define Λ_0 as a union of type classes and hence a union of types.

As to the payoff, the defender wishes to minimize the type II error probability, i.e.

$$u_D = -P_{fn} = -\sum_{\mathbf{y}^n : P_{\mathbf{y}^n} \in \Lambda_0} P_{\mathbf{Y}}(\mathbf{y}^n), \tag{2}$$

where with a light abuse of notation $P_{\mathbf{Y}}(\mathbf{y}^n)$ indicates the probability that Y emits the vector sequence $\mathbf{y}^n$. Our main result regarding the MO-HT game with perfect knowledge is the following.

Theorem 1. *The strategy*

$$\Lambda_0^* = \left\{ P \in \boldsymbol{\mathcal{P}}_n : \mathcal{D}(P||P_{\mathbf{X}}) < \lambda - |\mathcal{X}|^k \frac{\log(n+1)}{n} \right\} \tag{3}$$

where $\mathcal{D}(P||P_{\mathbf{X}})$ indicates the divergence [12] between P and $P_{\mathbf{X}}$, is a dominant strategy for D.

Proof: The proof is virtually identical to the proof of Lemma 1 in [3] and is omitted. ∎

In practice the fusion center gathers all the observations and verifies if their joint empirical pmf is in accordance with the expected statistics of $\mathbf{X}$ when H_0 holds.

D. Marginal-based MO-HT

As a second scenario we consider a situation in which the nodes summarize their observations by passing to the fusion center the first order statistics of the observed sequences. In other words, we assume that $m = |\mathcal{X}|$ and $f_l^{|\mathcal{X}|} = P_{x_l^n}$. As an example in which such a scenario applies, we may consider the case of a sensor network in which the nodes observe the system but their link to the fusion center has a very low transmission rate (hypothetically tending to 0). The nodes, then, transmit only the empirical pmf of the observed sequences, i.e. the number of times that each symbol of $\mathcal{X}$ appears in x_l^n. The number of necessary bits to transmit such an information is upper bounded by $|\mathcal{X}| \times \log_2 n$, since each symbol of X may appear in x_l^n at most n times. The rate necessary to code this information is hence $\frac{|\mathcal{X}| \times \log_2 n}{n}$, which tends to 0 when $n \to \infty$. Another possible justification for this scenario is the practical difficulty of getting a reliable estimate of the empirical joint pmf. It makes sense, then, for the defender to rely only on the empirical marginal pmf's, but still exploit the knowledge he has on the joint pmf of $\mathbf{X}$.

Given that decision fusion is carried out by considering only the empirical marginal distribution of the vector of observations $\mathbf{x}^n$, the defender is forced to choose a region for H_0 which is a subset of the Cartesian product among the marginal types, i.e. $\mathcal{P}_n^k = \mathcal{P}_n \times \mathcal{P}_n \dots \mathcal{P}_n$. More precisely we have:

$$\mathcal{S}_D = \{\Lambda_0 \in 2^{\mathcal{P}_n^k} \text{ s.t. } P_{fp} \leq 2^{-\lambda n}\}. \tag{4}$$

As to the payoff, D still aims at minimizing P_{fn} (equation (2)). Finding the optimal acceptance region requires that we compute the probability that a source with a joint pmf $P_{\mathbf{X}}$ emits a sequence having certain marginals. This can be done by considering the probability, under $P_{\mathbf{X}}$, of all the joint type classes having the desired marginals. To elaborate, let us indicate by $\mathcal{A}_n(P_1, P_2 \dots P_k)$ the set with all joint types with marginals $P_1, P_2 \dots P_k$, that is:

$$\mathcal{A}_n(P_1 \dots P_k) = \{P \in \boldsymbol{\mathcal{P}}_n : \sum_{-i} P(x_1 \dots x_k) = P_i \ \forall i\}, \tag{5}$$

where $\sum_{-i}$ indicates summation over all variables x_j but x_i. Given that the probability of a generic type class Q under $P_{\mathbf{X}}$ decays exponentially fast with exponent $\mathcal{D}(Q||P_{\mathbf{X}})$ and given that the number of types increases polynomially with n, we can proceed as in Lemma 1 in [3] to prove the following theorem.

Theorem 2. *The strategy*

$$\Lambda_0^* = \Bigg\{(P_1 \dots P_k) \in \mathcal{P}_n^k : \min_{P \in \mathcal{A}_n(P_1 \dots P_k)} \mathcal{D}(P||P_{\mathbf{X}}) < \lambda - |\mathcal{X}|^k \frac{\log(n+1)}{n}\Bigg\} \quad (6)$$

is a dominant strategy for D.

Proof: The proof is omitted for sake of brevity. ∎

One may wonder how the above result changes when the defender does not know $P_{\mathbf{X}}$ but only its marginals. This is the case, for instance, of JPEG forensic tools that analyze separately the DCT coefficients of an image without considering the dependencies between them. In this case it makes sense to adopt a worse case perspective and require that $P_{fp} \leq 2^{-\lambda n}$ for all joint pmf's with assigned marginals. The dominant strategy then includes a double minimization as follows:

$$\Lambda_0^* = \Bigg\{(P_1 \dots P_k) \in \mathcal{P}_n^k : \min_{P_{\mathbf{X}} \in \mathcal{A}(P_{X_1} \dots P_{X_k})} \min_{P \in \mathcal{A}_n(P_1 \dots P_k)} \mathcal{D}(P||P_{\mathbf{X}}) < \lambda - |\mathcal{X}|^k \frac{\log(n+1)}{n}\Bigg\}. \quad (7)$$

E. MO-HT based on local decisions

The last scenario we are going to consider assumes that the nodes can send to the fusion center only one bit of information ($m = 1$ and $f_l^1 \in \{0, 1\}$). This is a common situation, occurring, for instance but not only, when the nodes take their own decision about the state of the system and data fusion is carried out at the decision level. This scenario also models a multimedia forensic analysis in which the analyst applies several tools each of which provides a binary output regarding the origin or the authenticity of the analyzed document. It is the task of the fusion center to take a final decision by considering the output of all the tools. In principle we would like to derive the optimal decision strategies at the nodes and the optimal fusion strategy. This is a complex task, so we make the simplifying assumption that D adopts an AND fusion strategy, that is H_0 is accepted only if all the nodes accept it. Assuming an AND-based decision rule is equivalent to imposing that the the overall acceptance region is the Cartesian product of the acceptance regions adopted by the nodes, i.e., $\Lambda_0 = \Lambda_{0,1} \times \Lambda_{0,2} \dots \Lambda_{0,k}$. As in the previous sections, we assume that the nodes can rely only on the first order statistics of the observed sequences.

According to the above scenario, the space of strategies of the defender consists of all k-uple of local acceptance regions, that is:

$$\mathcal{S}_D = \{(\Lambda_{0,1} \dots \Lambda_{0,k}) : \Lambda_{0,i} \in 2^{\mathcal{P}_n} \text{ and } P_{fp} \leq 2^{-\lambda n}\}. \quad (8)$$

The payoff function is again the false negative error probability. We now prove the following theorem.

Theorem 3. *The strategy*

$$\Lambda_{0,i}^* = \Bigg\{P_i \in \mathcal{P}_n : \mathcal{D}(P_i||P_{X_i}) < \lambda - |\mathcal{X}|\frac{\log(n+1)}{n}\Bigg\} \ \forall i \quad (9)$$

is a dominant strategy for D.

Proof: The proof consists of two steps. First, we prove that the acceptance region Λ_0^* resulting from the local decision rules defined in (9) is an asymptotically admissible choice for D (i.e. it satisfies the constraint on type I error probability). Then we show that, under the assumption that D adopts an AND fusion rule, the local acceptance regions in (9) minimize the overall type II error probability. Let $\Lambda_{0,i}^{*,c}$ be the rejection region of H_0 at node i. We have:

$$\begin{aligned} P_{fp} &= P_{\mathbf{X}}(\mathbf{x}^n \in \Lambda_0^{*,c}) \\ &= P_{\mathbf{X}}(x_1^n \in \Lambda_{0,1}^{*,c} \text{ OR } x_2^n \in \Lambda_{0,2}^{*,c} \text{ OR } \dots \text{ OR } x_k^n \in \Lambda_{0,k}^{*,c}) \\ &\leq \sum_{i=1}^{k} P_{X_i}(x_i^n \in \Lambda_{0,i}^{*,c}). \end{aligned} \quad (10)$$

Due to the first-order assumption, the acceptance region at each node is a union of type classes (or equivalently a union of types with denominator n), hence we can write:

$$\begin{aligned} P_{fp} &\leq \sum_{i=1}^{k} \sum_{P \in \Lambda_{0,i}^{*,c}} P_{X_i}(T(P)) \\ &\overset{a}{\leq} \sum_{i=1}^{k} (n+1)^{|\mathcal{X}|} \max_{P \in \Lambda_{0,i}^{*,c}} P_{X_i}(T(P)) \\ &\overset{b}{\leq} \sum_{i=1}^{k} (n+1)^{|\mathcal{X}|} 2^{-n \min_{P \in \Lambda_{0,i}^{*,c}} \mathcal{D}(P||P_{X_i})} \\ &\overset{c}{\leq} k(n+1)^{|\mathcal{X}|} 2^{-n\left(\lambda - |\mathcal{X}|\frac{\log(n+1)}{n}\right)} \end{aligned} \quad (11)$$

where a and b derive from known upper bound on the number of types with denominator n and on the probability of a type class under a probability measure P_{X_i} [12], and c is a consequence of (9). We have thus shown that $P_{fp} \leq 2^{-n(\lambda - \delta_n)}$ with $\delta_n \to 0$ for $n \to \infty$, and hence Λ_0^* asymptotically satisfies the constraint on P_{fp}.

We now pass to the second part of the proof to show that the strategy in (9) is indeed optimal. Let Λ_0 be an AND-based acceptance region resulting from any other set of local regions $\Lambda_{0,i}$ satisfying the constraint on false positive error probability. Finally, let $\mathbf{x}^{n,*}$ belong to Λ_0^c. This means that $x_i^{n,*} \in \Lambda_{0,i}^c$ for at least one i, say j. We have:

$$\begin{aligned} 2^{-n\lambda} &\geq P_{\mathbf{X}}(x_i^n \in \Lambda_{0,i}^c, \text{ for some } i) \\ &\overset{a}{\geq} P_{X_j}(x_j^n \in \Lambda_{0,j}^c) \\ &= \sum_{P \in \Lambda_{0,j}^c} P_{X_j}(T(P)) \\ &\overset{b}{\geq} P_{X_j}(T(P_{x_j^{n,*}})) \overset{c}{\geq} \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-n\mathcal{D}(P_{x_j^{n,*}}||P_{X_j})}, \end{aligned} \quad (12)$$

where a is obtained by observing that the probability of a union of events is always larger than the probability of one such event, b holds since we have assumed that $\Lambda_{0,j}^c$ contains at least $x_j^{n,*}$ (and the corresponding type class), and c derives from a known lower bound on the probability of a type class [12]. By considering the first and the last term in (12), we see that $\mathbf{x}^{n,*} \in \Lambda_0^{*,c}$ and hence $\Lambda_0^* \subseteq \Lambda_0$. This shows that any other acceptance region Λ_0 satisfying the false positive constraint results in a higher false negative probability, thus proving the optimality of Λ_0^*. ∎

In practice, according to Theorem 3, H_0 is accepted only if the empirical marginals of the sequences observed by the nodes are in accordance with the system model under H_0. Moreover, somewhat expectedly, D does not exploit the knowledge of the joint pmf $P_{\mathbf{X}}$, the optimum decision rule depending only on P_{X_i}.

A unifying, and very important, characteristic of all the scenarios considered in this section, is that the requirement that P_{fp} tends to zero exponentially fast with decay exponent λ and the adoption of a decision rule based on first order statistics already define the optimum

defender's strategy regardless of the strategy chosen by attacker, thus resulting in the existence of a dominant strategy for D. Moreover, the dominant strategy does not depend on $P_{\mathbf{Y}}$, that is the statistical characterization of the system when H_0 does not hold, making such a knowledge un-necessary.

IV. Optimal attacker's strategies

Having derived the optimal strategies for the defender, we now adopt the perspective of the attacker (hereafter referred to as A). The existence of a dominant strategy for D makes it possible to study the optimal attacker's strategy by knowing that the acceptance region adopted by D is equal to Λ_0^*. Together with Λ_0^*, A's optimum strategy defines the equilibrium point of the game, which, being a dominant equilibrium, is also the only rationalizable equilibrium of the game.

A. Strategy space of the attacker

As a first step, we must define the space of strategies A can choose from and the information he has access to. As detailed in Section II, A acts only when H_0 does not hold with the aim of inducing a type II error. In order to do so, he corrupts either the observation sequences (MO-HT with corrupted observations), or the summaries sent by the nodes to the fusion center (MO-HT with corrupted nodes). In the former case, A must satisfy a distortion constraint specifying to which extent the sequences $x_1^n \dots x_k^n$ can be modified. In both cases, A may be allowed to attack all the sequences or only h of them. In the following, we indicate with y_l^n the observed sequences when H_1 holds and with v_l^m the corresponding feature sequences. The action of the attacker corresponds to applying a function $g(\cdot)$ either to y_l^n or v_l^m to produce k attacked sequences z_l^n (w_l^m in the case of corrupted nodes).

1) $\mathcal{S}_A$ for MO-HT with corrupted observations: The set of strategies available to A for the MO-HT game with corrupted observations is given by:

$$\mathcal{S}_A = \{g(\cdot) : d(\mathbf{z}^n, \mathbf{y}^n) \leq nD_{max}\}, \tag{13}$$

where D_{max} is the maximum allowed average per letter distortion. Alternatively, we can impose independent constraints on the distortion introduced in each of the observed sequences:

$$\mathcal{S}_A = \{g(\cdot) : d(z_i^n, y_i^n) \leq nD_{i,max}\ \forall i\}. \tag{14}$$

Similar definitions hold when A can corrupt up to h sequences.

2) $\mathcal{S}_A$ for MO-HT with corrupted nodes: In the case of corrupted nodes the attacker has much more freedom, since in this case he can work directly on the feature sequences v_l^m. All the more that, due to the absence of the distortion constraint, he can replace the feature sequences of the attacked nodes at will. The only applicable constraint is that he can substitute up to h sequences. In the case of chosen corrupted nodes, the space of strategies includes also the choice of the to-be attacked nodes.

Having defined $\mathcal{S}_A$, we must specify the information available to A. To do so, we adopt a worse case assumption and consider an omniscient attacker, who knows the system status (this is implicit in the Neaman-Pearson setup) and can observe all observation and feature sequences, even those that he is not allowed to modify.

B. Optimum attack for MO-HT with full knowledge

Let us consider the case of corrupted observations first. Given the optimal defender's strategy in (3), it is easy to realize that the optimum strategy for A is to modify the observed sequences so that the divergence between their empirical joint pmf and $P_{\mathbf{X}}$ is as small as possible while satisfying the distortion constraint, that is:

$$g^*(\mathbf{y}^n) = \arg \min_{\mathbf{z}^n : d(\mathbf{z}^n, \mathbf{y}^n) \leq nD_{max}} \mathcal{D}(P_{\mathbf{z}^n} || P_{\mathbf{X}}). \tag{15}$$

This result is analogous to Theorem 1 in [3] (see equation (16) therein), the only difference being that vector sources are involved instead of scalar ones. We point out that, in principle, A could reach the same goal by using a lower $\mathcal{D}$, stopping as soon as the pmf gets inside the acceptance region. Given our definition of the game, however, such a situation would not result in a higher payoff. This is the way to save as much distortion as possible which however, in our case, is unnecessary. A similar result holds when the distortion constraint applies to each observed sequence separately. Note that, even if theoretically simple, solving the minimization in (15) may be computationally very expensive, as already pointed out in [3] for the scalar case.

In the case of MO-HT with corrupted nodes, the situation is by far more favorable to the attacker, since he has to solve the minimization problem without any constraint. It is obvious, then, that A can pass to the fusion center completely fake sequences for which the divergence between the empirical joint pmf and $P_{\mathbf{X}}$ is arbitrarily small. Such sequences will pass the test in (3), thus always resulting in a false negative error.

The situation is different when A can attack only h out of k nodes. Even in the most favorable case of corrupted nodes, A can not control the empirical marginals of the non-attacked nodes and the joint pmf between them. If such marginals, or joint pmf, under H_1 are different from those under H_0, it may still be possible for the defender to reliably distinguish between the two hypothesis (though with a higher P_{fn}). It is also evident that, in the case of chosen corrupted nodes, A will attack the nodes for which the pmf's of the observations under H_0 and H_1 differ most in terms of divergence.

C. Optimum attack for Marginal-based MO-HT

Even in this case the optimal attacking strategy follows directly from the knowledge of D's dominant strategy. In fact, for the case of corrupted observations, from equation (6), it follows that:

$$g^*(\mathbf{y}^n) = \arg \min_{\mathbf{z}^n : d(\mathbf{z}^n, \mathbf{y}^n) \leq nD_{max}} \min_{P \in \mathcal{A}_n(P_{z_1^n} \dots P_{z_k^n})} \mathcal{D}(P||P_{\mathbf{X}}). \tag{16}$$

A similar result holds when equation (7) applies instead of (6). The situation is more favorable when the attacker can corrupt the output of the nodes, since in this case he can choose directly the pmf's $P_1 \dots P_k$ that minimize $\min_{P \in \mathcal{A}_n(P_1 \dots P_k)} \mathcal{D}(P||P_{\mathbf{X}})$. In fact, by letting $w_i^{|\mathcal{X}|,*} = P_i = P_{X_i}$ for all i, we have a perfect attack, since in this case $\min_{P \in \mathcal{A}_n(P_1 \dots P_k)} \mathcal{D}(P||P_{\mathbf{X}})$ is equal to 0. Of course, this is not possible when the attacker controls only h nodes, in which case the optimum attack boils down to the following minimization (w.l.o.g. we assume that A attacks the first h nodes):

$$P^* = \arg \min_{P \in \mathcal{A}_n(\dots, P_{y_{h+i}^n} \dots P_{y_k^n})} \mathcal{D}(P||P_{\mathbf{X}}), \tag{17}$$

where $\mathcal{A}_n(\dots, P_{y_{h+i}^n} \dots P_{y_k^n})$ denotes the set with all joint pmf's with only the last $n-h$ marginals fixed. Once the minimization is solved, A sets $w_i^{|\mathcal{X}|,*} = P_i^*$, $\forall i = 1 \dots h$.

Finally, when the attacker chooses which nodes to attack, a further minimization is required to minimize (17) over all possible subsets of attacked nodes.

D. Optimum attack for MO-HT based on local decisions

Once again the optimum attacker's strategy follows directly from the knowledge of the dominant strategy of the defender. By considering Theorem 3, in fact, is easy to conclude that the optimum strategy for A in the case of corrupted observations is:

$$g^*(\mathbf{y}^n) = \arg \min_{\mathbf{z}^n : d(\mathbf{z}^n, \mathbf{y}^n) \leq nD_{max}} \max_i \mathcal{D}(P_{z_i^n} || P_{X_i}). \tag{18}$$

As before the derivation of the optimum attack may be computationally expensive due to the presence of the distance constraint. If the squared Euclidean distance is adopted, a kind of waterfilling approach can be applied. The attacker, in fact, can operate as follows: choose i such that $\mathcal{D}(P_{y_i^n}||P_{X_i})$ is maximum, and compute z_i^n such that $\mathcal{D}(P_{z_i^n}||P_{X_i}) = \lambda - |\mathcal{X}|\log(n+1)/n - \varepsilon$ (with ε arbitrarily small), and the squared Euclidean distance between z_i^n and y_i^n is minimum. If the distortion is lower than nD_{max}, go on with the next i such that $\mathcal{D}(P_{y_i^n}||P_{X_i})$ is maximum, and iterate the above procedure until all $\mathcal{D}(P_{y_i^n}||P_{X_i})$ are lower than the decision threshold or when the maximum distortion is reached.

A considerably simpler situation is obtained when separate distortion constraints apply to the different sequences. In this case in fact, the attacker has to solve at most k independent scalar minimizations.

To conclude, we consider the case of corrupted nodes. In this case the optimum attack is trivial, since the attacker needs only to set the output of all the nodes under his control to 0, namely $w_i^{1,*} = 0, \forall i = 1 \dots h$. Note however that, if A does not control all the nodes, this may not be enough to make the final decision fail, since the fusion center accepts H_0 only if all the nodes accept it.

In the case of chosen attacked nodes, A will attack the nodes for which the marginals under H_1 differ most (in terms of divergence) from those under H_0.

We point out that this scenario is somewhat different from the usual case of decision fusion in the presence of Byzantines [9]. In that case, in fact, the byzantine nodes do not have a full knowledge of system status (which they know only through the observation of $\mathbf{x}^n$) and flip the output of the local decisions with a certain probability. In addition they usually act both when H_0 holds and when it doesn't.

V. Discussion and conclusions

Having derived the equilibrium point of several versions of the MO-HT game, we are ready to derive some conclusions and summarize the main lessons that we learnt from our analysis. At a first sight, in fact, our analysis may look rather theoretical making difficult distilling some practical conclusions.

To start with, we observe that the theoretical framework with the taxonomy of several kinds of scenarios referring to different practical applications, is by itself a fundamental step towards the comprehension of the addressed problems and the development of practical strategies for both the attacker and the defender.

With regard to the specific results we have proven, the most interesting result regards the existence of a dominant strategy for the defender. What Theorems 1 through 3 say, in fact, is that the defender may choose its strategy without caring about the attacker. For instance, he would get no advantage from the knowledge of the attacked nodes, let alone from any attempt to discover them. This marks an important difference with respect to previous works in which the defender tries to distinguish between honest and malicious nodes (for some examples of such an approach see [14], [15]). In hindsight, the reason for such an apparently strange behavior, is the adoption of a Neyman-Pearson setup wherein the attacker acts only when H_0 does not hold, while the defender is asked to satisfy a requirement on P_{fp}, i.e., by assuming that H_0 holds. Coupled with the adoption of an asymptotic setup, this results in the existence of a dominant strategy for D that does not need to know whether a node (or an observation) is controlled by the adversary or not. It goes without saying that is some applications the assumptions we made may not be reasonable, thus opening the way to different formulations of the MO-HT game.

Having determined the equilibrium point of the various games, the next step would require that the payoff at the equilibrium is evaluated so to know who is going to *win* the game. In other words, given the pmf's under H_0 and H_1 (res. $P_{\mathbf{X}}$ and $P_{\mathbf{Y}}$), and a distortion constraint D_{max} (in the corrupted observations setup), we would like to know whether the probability of a type II error ultimately tends to 0 or 1 when $n \to \infty$. Doing so for $\lambda \to 0$ would finally permit us to decide whether the two hypothesis H_0 and H_1 are ultimately distinguishable or not, when the attacker is allowed to attack h observation sequences (or nodes) with a maximum per letter distortion D_{max} (see [16] for a preliminary analysis in this sense for a single-observation test).

Acknowledgment

This work was partially supported by the European Office of Aerospace Research and Development under Grant FA8655-12-1-2138: AMULET - A multi-clue approach to image forensics, and by the REWIND Project, funded within the FET (Future and Emerging Technologies) program by EC under grant 268478.

References

[1] M. Barni and F. Pérez-González, "Coping with the enemy: advances in adversary-aware signal processing," in *ICASSP 2013, IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada, May 2013.

[2] M. C. Stamm, W. S. Lin, and K. J. R. Liu, "Forensics vs anti-forensics: a decision and game theoretic framework," in *ICASSP 2012, IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Kyoto, Japan, 25-30 March 2012.

[3] M. Barni and B. Tondi, "The source identification game: an information-theoretic perspective," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 3, pp. 450–463, March 2013.

[4] ——, "Binary hypothesis testing game with training data," *Submitted to IEEE Transactions on Information Theory*, arXiv preprint arXiv:1304.2172, 2013.

[5] M. Fontani, T. Bianchi, A. De Rosa, A. Piva, and M. Barni, "A framework for decision fusion in image forensics based on Dempster-Shafer theory of evidence," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 4, pp. 593–607, April 2013.

[6] P. K. Varshney, *Distributed Detection and Data Fusion*. Springer-Verlag, 1997.

[7] J.-F. Chamberland and V. V. Veeravalli, "Decentralized detection in sensor networks," *IEEE Transactions on Signal Processing*, vol. 51, no. 2, pp. 407–416, February 2003.

[8] I. F. Akyildiz, B. F. Lo, and R. Balakrishnan, "Cooperative spectrum sensing in cognitive radio networks: A survey," *Physical Communication*, vol. 2011, no. 4, pp. 40–62, 2011.

[9] B. Kailkhura, S. Brahma, Y. S. Han, and P. K. Varshney, "Optimal distributed detection in the presence of byzantines," in *ICASSP 2013, IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vancouver, Canada, 27-31 May 2013.

[10] M. J. Osborne and A. Rubinstein, *A Course in Game Theory*. MIT Press, 1994.

[11] Y. C. Chen, N. Van Long, and X. Luo, "Iterated strict dominance in general games," *Games and Economic Behavior*, vol. 61, no. 2, pp. 299–315, November 2007.

[12] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley Interscience, 1991.

[13] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems. 2nd edition*. Cambridge University Press, 2011.

[14] A. S. Rawat, P. Anand, C. Hao, and P. K. Varshney., "Collaborative spectrum sensing in the presence of byzantine attacks in cognitive radio networks," *IEEE Transactions on Signal Processing*, vol. 59, no. 2, pp. 774–786, 2011.

[15] Y. Liu and Y. L. Sun, "Anomaly detection in feedback-based reputation systems through temporal and correlation analysis," in *Proc. of 2nd IEEE Int. Conf. on Social Computing*, August 2010.

[16] M. Barni and B. Tondi, "The security margin: a measure of source distinguishability under adversarial conditions," in *Submitted to IEEE Global Conference on Signal and Image Processing*, Dec. 2013.