Optimum Forensic and Counter-forensic Strategies for Source Identification with Training Data

Mauro Barni [1], Benedetta Tondi [2]

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

***Abstract*—In the attempt to provide a mathematical background to multimedia forensics, we introduce the source identification game with training data. The game models a scenario in which a forensic analyst has to decide whether a test sequence has been drawn from a source X or not. In turn, the adversary takes a sequence generated by a different source a modifies it in such a way to induce a classification error. The source X is known only through one or more training sequences. We derive the asymptotic Nash equilibrium of the game under the assumption that the analyst relies only on first order statistics of the test sequence. A geometric interpretation of the result is given together with a comparison with a similar version of the game with known sources. The comparison between the two versions of the games gives interesting insights into the differences and similarities of the two games.**

I. Introduction

Understanding the fundament limits of multimedia forensics in an adversarial environment is a pressing need to avoid the proliferation of forensic and anti-forensic tools each focused on countering a specific action of the adversary but prone to yet another class of attacks and counter-attacks. The most natural solution to avoid entering this never-ending loop is to cast the forensic problem into a game-theoretic framework and look for the optimum strategies the players of the game (usually a forensic analyst and an adversary) should adopt. Some early attempts in this direction can be found in [1] and [2]. In [1], the authors introduce a game-theoretic framework to evaluate the effectiveness of a given attacking strategy and derive the optimal countermeasures. In [1] the attacker's strategy is fixed and the game-theoretic framework is used only to determine the optimal parameters of the forensic analysis and the attack. A more general approach is adopted in [2], where the source identification game with known statistics, namely the SI_{ks} game, is introduced. According to the framework defined in [2], given a discrete memoryless source (DMS) X with known statistics P_X, it is the goal of the Forensic Analyst (FA) to decide whether a test sequence x^n has been drawn from X or not. In doing so, he has to ensure that the false positive probability, i.e. the probability of deciding that the test sequence has not been generated by X when it actually was, stays below a predefined maximum value. The goal of the adversary (AD) is to take a sequence generated from a different and independent source $Y \simeq P_Y$ and modify it so to let the FA think that the modified sequence has been generated by X. In doing so the AD must satisfy a distortion constraint, i.e. the distance between the original and the modified sequence must be lower than a threshold. The payoff of the AD is the false negative error probability, i.e. the probability that the FA classifies a sequence drawn from Y and further modified by the AD as a sequence drawn from X. The opposite payoff applies to the FA, thus qualifying the SI_{ks} as a zero-sum, competitive game [3]. Under the further assumption that the FA relies only on first order statistics (limited resources assumption) for his analysis and that the sources X and Y are memoryless, the asymptotic Nash equilibrium of the game can be found [2], [4], thus defining the optimum strategies for the FA and the AD when the length of the test sequence tends to infinity. A problem with the analysis carried out in [2] is the assumption that the FA and the AD know the probability mass function (pmf) of the source X. This is not the case in many practical scenarios where sources are known only through one or more training sequences. It is the goal of this paper to reformulate the analysis carried out in [2] to address this new more realistic version of the game. As a main result, we derive the asymptotic Nash equilibrium of the new game, hereafter referred to as the SI_{tr} game, under the same limited resources assumptions used in [2]. In doing so we will discover that the optimal strategies for the FA and the AD deviate from those of the SI_{ks} game. In addition, at least in the case that the training sequences available to the FA and the AD coincide, we can show that passing from the SI_{ks} to the SI_{tr} version of the game is to the AD's advantage.

The paper is organized as follows. In Section II we introduce the notation that will be used throughout the paper. In Section III, we give a rigorous definition of the source identification with training data game. In Section IV, we derive the asymptotic Nash equilibrium of the game. In Section V, we compare the results obtained in this paper with those referring to source identification with known sources. Section VI concludes the paper with some perspective for future research.

II. Notation

In the rest of this work we will use capital letters to indicate discrete memoryless sources (e.g. X). Sequences of length n

WIFS'2012, December, 2-5, 2012, Tenerife, Spain.
978-1-4673-2287-4/12/$31.00 ©2012 IEEE.

drawn from a source will be indicated with the corresponding lowercase letters (e.g. x^n). In the same way, we will indicate with x_i, $i = 1, n$ the $i-$th element of a sequence x^n. The alphabet of an information source will be indicated by the corresponding calligraphic capital letter (e.g. $\mathcal{X}$). Calligraphic letters will also be used to indicate classes of information sources ($\mathcal{C}$). The pmf of a discrete memoryless source X will be denoted by P_X. With a slight abuse of notation, the same symbol will be used to indicate the probability measure ruling the emission of sequences from X, so we will use the expressions $P_X(a)$ and $P_X(x^n)$ to indicate, respectively, the probability of symbol $a \in \mathcal{X}$ and the probability that the source X emits the sequence x^n. Given an event A (be it a subset of $\mathcal{X}$ or $\mathcal{X}^n$), we will use the notation $P_X(A)$ to indicate the probability of the event A under the probability measure P_X.

Our analysis relies heavily on the concepts of type and type class defined as follows (see [5] and [6] for more details). Let x^n be a sequence 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, i.e. $\forall a \in \mathcal{X}, P_{x^n}(a) = \frac{1}{n}\sum_{i=1}^{n}\delta(x_i, a)$. 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.

The Kullback-Leibler (KL) divergence between two distributions P and Q on the same finite alphabet $\mathcal{X}$ is defined as:

$$\mathcal{D}(P||Q) = \sum_{a\in\mathcal{X}} P(a)\log\frac{P(a)}{Q(a)}, \tag{1}$$

where, as usual, $0\log 0 = 0$ and $p\log p/0 = \infty$ if $p > 0$. Empirical distributions can be used to calculate empirical information theoretic quantities, like, for instance, the empirical divergence between two sequences $D(P_{x^n}||P_{y^n})$.

As we said, the goal of this paper is to cast the source identification problem into a game-theoretic framework, wherein identification is seen as a two-player, strategic, zero-sum game. In rigorous terms, a game is defined 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 (actions) 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 $s_{1,i}$ and $s_{2,j}$ is called a profile. In a zero-sum competitive game, the two payoff functions are strictly related to each other since for any profile we have $u_1(s_{1,i}, s_{2,j}) + u_2(s_{1,i}, s_{2,j}) = 0$. A zero-sum game, then reduces to a triplet $G(\mathcal{S}_1, \mathcal{S}_2, u)$, where we have assumed $u = u_1 = -u_2$. Note that in strategic games the players choose their strategies before starting the game so that they have no hints about the strategy actually chosen by the other player. We say that a profile (s_{1,i^*}, s_{2,j^*}) represents a Nash equilibrium if [7], [3]:

$$\begin{aligned} u_1((s_{1,i^*}, s_{2,j^*})) &\geq u_1((s_{1,i}, s_{2,j^*})) \quad \forall s_{1,i} \in \mathcal{S}_1 \\ u_2((s_{1,i^*}, s_{2,j^*})) &\geq u_2((s_{1,i^*}, s_{2,j})) \quad \forall s_{2,j} \in \mathcal{S}_2, \end{aligned} \tag{2}$$

where for a zero-sum game $-u_2 = u_1 = u$.

III. Source identification with training data

Let $\mathcal{C}$ be the class of discrete memoryless sources with alphabet $\mathcal{X}$, and let $X \simeq P_X$ be a source in $\mathcal{C}$. Given a test sequence x^n, the goal of the Forensic Analyst (FA) is to decide whether x^n was drawn from X or not[1]. As opposed to the source identification game with known sources [2], here we assume that the FA does not know P_X, and that he has to base his decision by relying on the knowledge of a training sequence t_{FA}^N drawn from X. On his side, the Adversary (AD) takes a sequence y^n emitted by another source $Y \simeq P_Y$ still belonging to $\mathcal{C}$ and tries to modify it in such a way that the FA thinks that the modified sequence was generated by X. In doing so the AD must satisfy a distortion constraint stating that the distance between the modified sequence, say z^n, and y^n must be lower than a predefined threshold. As the FA, the AD knows P_X through a training sequence t_{AD}^K, that in general may be different than t_{FA}^N. We assume that t_{FA}^N, t_{AD}^K, x^n and y^n are generated independently. With regard to P_Y, we could also assume that it is known through two training sequences, one available to the FA and one to the AD, however we will see that - at least to study the asymptotic behavior of the game - such an assumption is not necessary, and hence we take the simplifying assumption that P_Y is known neither to the FA nor to the AD. As in [2], we define the game by casting the identification problem into a hypothesis decision framework. Let then H_0 be the hypothesis that the test sequence has been generated by X (i.e. the same source that generated t_{FA}^N) and let Λ_0 be the acceptance region for H_0 (similarly we indicate with $\Lambda_1 = \Lambda_0^c$ the rejection region for H_0). We have the following:

Definition 1. *The $SI_{tr,a}(\mathcal{S}_{FA}, \mathcal{S}_{AD}, u)$ game is a zero-sum, strategic, game played by the FA and the AD, defined by the following strategies and payoff.*

- *The set of strategies the FA can choose from is the set of acceptance regions for H_0 for which the maximum false positive probability across all possible $P_X \in \mathcal{C}$ is lower than a certain threshold:*

$$\mathcal{S}_{FA} = \{\Lambda_0 : \max_{P_X\in\mathcal{C}} P_X\{(x^n, t_{FA}^N) \notin \Lambda_0\} \leq P_{fp}\}, \tag{3}$$

  where P_{fp} is a prescribed maximum false positive probability, and where $P_X\{(x^n, t_{FA}^N) \notin \Lambda_0\}$ indicates the probability that two independent sequences generated by X do not belong to Λ_0. Note that the acceptance region is defined as a union of pairs of sequences, and hence $\Lambda_0 \subset R^n \times R^N$.
- *The set of strategies the AD can choose from is formed by all the functions that map a sequence $y^n \in \mathcal{X}^n$ into a new sequence $z^n \in \mathcal{X}^n$ subject to a distortion constraint:*

$$\mathcal{S}_{AD} = \{f(y^n, t_{AD}^K) : d(y^n, f(y^n, t_{AD}^K)) \leq nD\}, \tag{4}$$

[1]With a slight abuse of notation we use the symbol x^n to indicate the test sequence even if strictly speaking it is not known whether the test sequence originated from X or Y.

where $d(\cdot,\cdot)$ is a proper distance function and D is the maximum allowed per-letter distortion. Note that the function $f(\cdot)$ depends on t_{AD}^K, since when performing his attack the AD will exploit the knowledge of the training sequence.

- *The payoff function is defined in terms of the false negative error probability (P_{fn}), namely:*

$$u(\Lambda_0, f) = -P_{fn} = -\sum_{\substack{t_{FA}^N \in \mathcal{X}^N,\ t_{AD}^K \in \mathcal{X}^K \\ y^n:(f(y^n, t_{AD}^K), t_{FA}^N) \in \Lambda_0}} P_Y(y^n) P_X(t_{FA}^N) P_X(t_{AD}^K), \tag{5}$$

where the error probability is averaged across all possible y^n and training sequences and where we have exploited the independence of y^n, t_{FA}^N and t_{AD}^K.

Some explanations are in order with regard to the definition of the payoff function. As a matter of fact, the expression in (5) looks problematic, since its evaluation requires that the pmf's P_X and P_Y are known, however this is not the case in our scenario since we have assumed that P_X is known only through t_{FA}^N and t_{AD}^K, and that P_Y is not known at all. As a consequence it may seem that the players of the game are not able to compute the payoff associated to a given profile and hence have no arguments upon which they can base their choice. While this is indeed a problem in a generic setup, we will show later on in the paper that asymptotically (when n, N and K tend to infinity) the optimum strategies of the FA and the AD are uniformly optimum across all P_X and P_Y and hence the ignorance of P_X and P_Y does not represent a problem. One may wonder why we did not define the payoff under a worst case assumption (from FA's perspective) on P_X and/or P_Y. The reason is that doing so would result in a meaningless game. In fact, given that X and Y are drawn from the same class of sources $\mathcal{C}$, the worst case would always correspond to the trivial case $X = Y$ for which no meaningful forensic analysis is possible[2].

Slightly different versions of the game are obtained by assuming a different relationship between the training sequences. In certain cases we may assume that the FA has a better access to the source X than the AD. In [8], for example, the availability of a number of pictures taken from a camera X and made publicly available is exploited by the AD to take an image produced by a camera Y and modify it in such a way that the fake picture looks as if it were taken by X. The FA, exploits his better access to the source X and the knowledge of the images potentially available to the AD to distinguish the images truly generated by X and the fake images produced by the AD. In our framework, such a scenario can be quite faithfully modeled by assuming that the sequence t_{AD}^K is a subsequence of t_{FA}^N, leading to the following definition.

Definition 2. *The $SI_{tr,b}(\mathcal{S}_{FA}, \mathcal{S}_{AD}, u)$ game is a zero-sum, strategic, game played by the FA and the AD, defined as the $SI_{tr,a}$ game with the only difference that $t_{AD}^K = (t_{FA,l+1}, t_{FA,l+2} \dots t_{FA,l+K})$ with l and K known to the FA.*

Yet another version of the game is obtained by assuming that the training sequence available to the AD corresponds to that available to the FA.

Definition 3. *The $SI_{tr,c}(\mathcal{S}_{FA}, \mathcal{S}_{AD}, u)$ game is a zero-sum, strategic, game played by the FA and the AD, defined as the $SI_{tr,a}$ game with the only difference that $K = N$ and $t_{AD}^K = t_{FA}^N$ (simply indicated as t^N in the following). The set of strategies of the FA and the AD are the same as in the $SI_{tr,a}$ game.*

In the rest of the paper we will focus on the $SI_{tr,c}$ game, leaving the other versions for future research.

IV. Asymptotic equilibrium for the $SI_{tr,c}$ game with limited-resources

Studying the existence of an equilibrium point for the $SI_{tr,c}$ game is a prohibitive task due to the difficulty of determining the optimum strategies for the FA and the AD, hence we consider a simplified version of the game in which the FA can only base his decision on a limited set of statistics computed on the test and training sequences. Specifically, we require that the FA relies only on the relative frequencies with which the symbols in $\mathcal{X}$ appear in x^n and t^N, i.e. P_{x^n} and P_{t^N}. Note that P_{x^n} and P_{t^N} are not sufficient statistics for the FA, since even if Y is also a memoryless source, the AD could introduce some memory within the sequence as a result of the application of $f(\cdot)$. In the same way it could introduce some dependencies between the attacked sequence z^n and t^N. It is then necessary to treat the assumption that the FA relies only on P_{x^n} and P_{t^N} as an explicit - additional - requirement. As in [2], we call this version of the game "source identification with limited-resources", and we refer to it as the $SI_{tr,*}^{lr}$ game. As a consequence of the limited resource assumption, Λ_0 can only be a union of cartesian products of pairs of type classes, i.e. if the pair of sequences (x^n, t^N) belongs to Λ_0, then any pair of sequences belonging to the cartesian product $T(P_{x^n}) \times T(P_{t^N})$ will be contained in Λ_0. Since a type class is univocally defined by the empirical pmf of the sequences contained in it, we can redefine the acceptance region Λ_0 as a union of pairs of types (P, Q) with $P \in \mathcal{P}_n$ and $Q \in \mathcal{P}_N$. In the following, we will use the two interpretations of Λ_0 (as a set of sequences or a set of types) interchangeably, the exact meaning being always clearly recoverable from the context. We are interested in studying the asymptotic behavior of the game when n and N tends to infinity. To avoid the necessity to consider two limits with n and N tending to infinity independently, we decided to express N as a function of n, and study what happens when n tends to infinity. With the above ideas in mind, we can state the following:

Definition 4. *The $SI_{tr,c}^{lr}(\mathcal{S}_{FA}, \mathcal{S}_{AD}, u)$ game is a zero-sum, strategic, game played by the FA and the AD, defined by the*

[2]Alternatively, we could assume that X and Y belong to two disjoint source classes $\mathcal{C}_X$ and $\mathcal{C}_Y$. We leave this analysis for further research.

following strategies and payoff:

$$\mathcal{S}_{FA} = \{\Lambda_0 \subset \mathcal{P}_n \times \mathcal{P}_{N(n)} : \max_{P_X \in \mathcal{C}} P_X\{(x^n, t^{N(n)}) \notin \Lambda_0\} \leq 2^{-\lambda n}\}, \quad (6)$$

$$\mathcal{S}_{AD} = \{f(y^n, t^{N(n)}) : d(y^n, f(y^n, t^{N(n)})) \leq nD\}, \quad (7)$$

$$u(\Lambda_0, f) = -P_{fn} = -\sum_{\substack{t^{N(n)} \in \mathcal{X}^{N(n)} \\ y^n : (f(y^n, t^{N(n)}), t^{N(n)}) \in \Lambda_0}} P_Y(y^n) P_X(t^{N(n)}). \quad (8)$$

Note that we ask that the false positive error probability decay exponentially fast with n, thus opening the way to the asymptotic solution of the game. Similar definitions obviously hold for the a and b versions of the game.

A. Optimum FA strategy

We start the study of the asymptotic equilibrium point of the $SI^{lr}_{tr,c}$ game determining the optimum decision region for the FA. In doing so we will use an analysis similar to that carried out in [9] to analyze a statistical problem with observed statistics (the main difference between our analysis and [9] is the presence of the AD, i.e. the game-theoretic nature of our problem). The derivation of the optimum strategy for the FA passes through the definition of the generalized log-likelihood ratio function $h(x^n, t^{N(n)})$. Given the test and training sequences x^n and $t^{N(n)}$, we define the generalized log-likelihood ratio function as ([9], [10])[3]:

$$h(x^n, t^N) = \mathcal{D}(P_{x^n} || P_{r^{N+n}}) + \frac{N}{n} \mathcal{D}(P_{t^N} || P_{r^{N+n}}), \quad (9)$$

where $P_{r^{N+n}}$ indicates the empirical pmf of the sequence r^{N+n}, obtained by concatenating t^N and x^n, i.e.

$$r^{N+n} = \begin{cases} t_i & i \leq N \\ x_{i-N} & N < i \leq n+N \end{cases}. \quad (10)$$

Observing that $h(x^n, t^N)$ depends on the test and the training sequences only through their empirical pmf, we can use the notation $h(P_{x^n}, P_{t^N})$. The derivation of the Nash equilibrium for the $SI^{lr}_{tr,c}$ game passes through the following lemmas.

Lemma 1. *For any P_X we have:*

$$n\mathcal{D}(P_{x^n} || P_{r^{n+N}}) + N\mathcal{D}(P_{t^N} || P_{r^{n+N}}) \leq n\mathcal{D}(P_{x^n} || P_X) + N\mathcal{D}(P_{t^N} || P_X), \quad (11)$$

with equality holding only if $P_X = P_{r^{n+N}}$.

The proof of Lemma 1 is given in the appendix.

Lemma 2. *Let Λ_0^* be defined as follows:*

$$\Lambda_0^* = \left\{ (P_{x^n}, P_{t^N}) : h(P_{x^n}, P_{t^N}) < \lambda - |\mathcal{X}| \frac{\log(n+1)(N+1)}{n} \right\} \quad (12)$$

with

$$\lim_{n \to \infty} \frac{\log(N(n)+1)}{n} = 0, \quad (13)$$

[3]To simplify the notation sometimes we omit the dependence of N on n.

and let Λ_1^* be the corresponding rejection region. Then:

1) $\max_{P_X} P_X\{(x^n, t^{N(n)}) \notin \Lambda_0^*\} \leq 2^{-n(\lambda - \delta_n)}$*, where δ_n is a sequence of positive numbers such that $\delta_n \to 0$ for $n \to \infty$,*
2) $\forall \Lambda_0 \in \mathcal{S}_{FA}$ *defined as in (6), we have* $\Lambda_1 \subseteq \Lambda_1^*$.

Proof: Being Λ_0^* (and Λ_1^*) a union of pairs of types (i.e. unions of cartesian products of type classes), we have:

$$\begin{aligned} \max_{P_X} P_{fp} &= \max_{P_X \in \mathcal{C}} P_X\{(x^n, t^N) \notin \Lambda_0^*\} \quad (14) \\ &= \max_{P_X \in \mathcal{C}} \sum_{(x^n, t^N) \in \Lambda_1^*} P_X(x^n, t^N) \\ &= \max_{P_X \in \mathcal{C}} \sum_{(P_{x^n}, P_{t^N}) \in \Lambda_1^*} P_X(T(P_{x^n}) \times T(P_{t^N})). \end{aligned}$$

For the class of discrete memoryless sources, the number of types with denominators n and N is bounded by $(n+1)^{|X|}$ and $(N+1)^{|X|}$ respectively [5], so we can write:

$$\begin{aligned} \max_{P_X} P_{fp} &\leq \max_{P_X} \max_{(P_{x^n}, P_{t^N}) \in \Lambda_1^*} \quad (15) \\ &\quad [(n+1)^{|\mathcal{X}|}(N+1)^{|\mathcal{X}|} P_X(T(P_{x^n}) \times T(P_{t^N}))] \\ &\leq (n+1)^{|\mathcal{X}|}(N+1)^{|\mathcal{X}|} \cdot \\ &\quad \max_{P_X} \max_{(P_{x^n}, P_{t^N}) \in \Lambda_1^*} 2^{-n[\mathcal{D}(P_{x^n} || P_X) + \frac{N}{n} \mathcal{D}(P_{t^N} || P_X)]}, \end{aligned}$$

where for the last inequality we have exploited the independence of x^n and t^N and the property of types according to which for any sequence x^n we have $P_X(T(P_{x^n})) \leq 2^{-n\mathcal{D}(P_{x^n} || P_X)}$ ([5]). By exploiting Lemma 1, we can write:

$$\begin{aligned} \max_{P_X} P_{fp} &\leq (n+1)^{|\mathcal{X}|}(N+1)^{|\mathcal{X}|} \quad (16) \\ &\quad \max_{(P_{x^n}, P_{t^N}) \in \Lambda_1^*} 2^{-n[\mathcal{D}(P_{x^n} || P_{r^{N+n}}) + \frac{N}{n} \mathcal{D}(P_{t^N} || P_{r^{N+n}})]} \\ &\leq (n+1)^{|\mathcal{X}|}(N+1)^{|\mathcal{X}|} \ 2^{-n(\lambda - |\mathcal{X}| \frac{\log(n+1)(N+1)}{n})} \\ &= 2^{-n(\lambda - 2|\mathcal{X}| \frac{\log(n+1)(N+1)}{n})}, \end{aligned}$$

where the last inequality derives from the definition of Λ_0^*. Together with (13), equation (16) proves the first part of the lemma with $\delta_n = 2|\mathcal{X}| \frac{\log(n+1)(N+1)}{n}$.

Let now (x^n, t^N) be a generic pair of sequences contained in Λ_1 (with $\Lambda_0 \in \mathcal{S}_{FA}$), due to the limited resources assumption the cartesian product between $T(P_{x^n})$ and $T(P_{t^N})$ will be entirely contained in Λ_1. Then we have:

$$\begin{aligned} 2^{-\lambda n} &\geq \max_{P_X} P_X(\Lambda_1) \quad (17) \\ &\overset{(a)}{\geq} \max_{P_X} P_X(T(P_{x^n}) \times T(P_{t^N})) \\ &\overset{(b)}{\geq} \max_{P_X} \frac{2^{-[\mathcal{D}(P_{x^n} || P_X)] + \frac{N}{n} \mathcal{D}(P_{t^N} || P_X)}}{(n+1)^{|\mathcal{X}|}(N+1)^{|\mathcal{X}|}} \\ &\overset{(c)}{=} \frac{2^{-[\mathcal{D}(P_{x^n} || P_{r^{N+n}})] + \frac{N}{n} \mathcal{D}(P_{t^N} || P_{r^{N+n}})}}{(n+1)^{|\mathcal{X}|}(N+1)^{|\mathcal{X}|}}, \end{aligned}$$

where (a) is due to the limited resources assumption, (b) follows from the independence of x^n and t^N and the lower bound on the probability of a pair of type classes [5], and (c)

derives from Lemma 1. By taking the logarithm of both sides we have that $(x^n, t^N) \in \Lambda_1^*$, thus completing the proof. ■

The first part of the lemma shows that, at least asymptotically, Λ_0^* belongs to $\mathcal{S}_{FA}$, while the second part implies the optimality of Λ_0^*. The most important consequence of Lemma 2 is that the optimum strategy of the FA is univocally determined by the false positive constraint. This solves the apparent problem that we pointed out when defining the payoff of the game, namely that the payoff depends on P_X and P_Y and hence it is not fully known to the FA. Another interesting result is that the optimum strategy of the FA does not depend on the strategy chosen by the AD, thus considerably simplifying the determination of the equilibrium point of the game. As a matter of fact, since the optimum Λ_0^* is fixed, the AD can choose his strategy by relying on the knowledge of Λ_0^*. A last consequence of Lemma 2 is that Λ_0^* is the optimum FA strategy even for versions a and b of the SI_{tr}^{lr} game.

B. Asymptotic Nash equilibrium

To determine the Nash equilibrium of the $SI_{tr,c}^{lr}$ game, we start by deriving the optimum strategy for the AD. This is quite an easy task if we observe that the goal of the AD is to take a sequence y^n drawn from Y and modify it in such a way that:

$$h(z^n, t^N) < \lambda - |\mathcal{X}| \frac{\log(n+1)(N+1)}{n}, \tag{18}$$

with $d(y^n, z^n) \leq nD$. The optimum attacking strategy, then, can be expressed as a minimization problem, i.e.:

$$f^*(y^n, t^N) = \arg \min_{z^n : d(y^n, z^n) \leq nD} h(z^n, t^N). \tag{19}$$

Note that to implement this strategy the AD needs to know t^N, i.e. equation (19) determines the optimum strategy only for version c of the game. Having determined the optimum strategies for the FA and the AD, we can state the fundamental result of this paper, summarized in the following theorem.

Theorem 1. *The profile (Λ_0^*, f^*) defined by Lemma 2 and equation (19) is an asymptotic Nash equilibrium point for the $SI_{tr,c}^{lr}$ game.*

Proof: We have to prove that:

$$u(\Lambda_0^*, f^*) \geq u(\Lambda_0, f^*) \quad \forall \Lambda_0 \in \mathcal{S}_{FA} \tag{20}$$

$$-u(\Lambda_0^*, f^*) \geq u(\Lambda_0^*, f) \quad \forall f \in \mathcal{S}_{AD}. \tag{21}$$

The first relation holds because of Lemma 2, while the second derives from the optimality of f^* when Λ_0^* is fixed, hence proving the theorem. ■

V. Discussion and comparison with the SI_{ks}^{lr} game.

In this section we give an intuitive meaning to the results proved so far. To do so we will compare the optimum strategies of the $SI_{tr,*}^{lr}$ game to those of the SI_{ks}^{lr}, i.e a version of the game in which the FA and the AD know the pmf P_X ruling the emission of symbols from the source X. In [2] it is shown

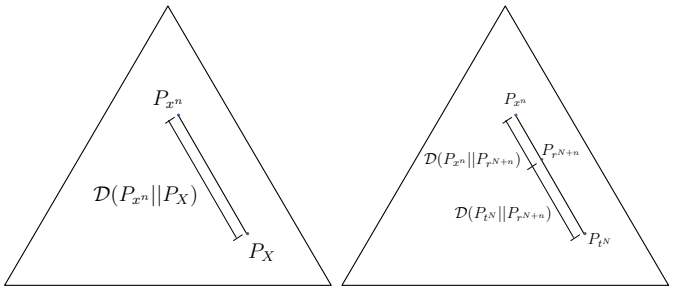

Fig. 1. Geometric interpretation of the optimum FA strategies for the SI_{ks}^{lr} (left) and the $SI_{tr,*}^{lr}$ (right) games.

that the optimum strategy for the FA relies on the divergence between the empirical pmf of the sequence x^n and P_X, i.e.:

$$\Lambda_{0,ks}^* = \left\{ P_{x^n} \in \mathcal{P}_n : \mathcal{D}(P_{x^n}||P_X) < \lambda - |\mathcal{X}| \frac{\log(n+1)}{n} \right\}. \tag{22}$$

One may wonder why the optimum FA strategy for the $SI_{tr,*}^{lr}$ game does not correspond to the comparison of the empirical divergence between x^n and that of the test sequence. The reason for the necessity of adopting the more complicated strategy set by Lemma 2 is that in the current version of the game, the FA must ensure that the false positive probability is below the desired threshold for all possible sources in $\mathcal{C}$. To do so, he has to estimate the pmf that better *explains* the evidence provided by both x^n and t^N. In other words he has to find the pmf under which the probability of observing both the sequences x^n and t^N is maximum. This is exactly the role of $P_{r^{n+N}}$ (see equation (A3)), with the generalized log-likelihood ratio corresponding to the log of the (asymptotic) probability of observing x^n and t^N under $P_{r^{n+N}}$ (a geometrical interpretation of the decision strategies for the two versions of the game is given in Fig. 1).

Another interesting observation regards the optimum strategy of the AD. As a matter of fact, the functions $h(P_{x^n}, P_{t^N})$ and $\mathcal{D}(P_{x^n}||P_{t^N})$ share a similar behavior: both are positive and convex functions with the absolute minimum achieved when $P_{x^n} = P_{t^N}$, so one may be tempted to think that from the AD's point of view minimizing $\mathcal{D}(P_{x^n}||P_{t^N})$ is equivalent to minimizing $h(P_{x^n}, P_{t^N})$. While this is the case in some situations, e.g. for binary sources or when the absolute minimum can be reached, in general the two minimization problems yield different solutions. It is possible, and quite easy in fact, to find two pmf's P'_{x^n} and P''_{x^n} for which $\mathcal{D}(P'_{x^n}||P_{t^N}) > \mathcal{D}(P''_{x^n}||P_{t^N})$, while $h(P'_{x^n}, P_{t^N}) < h(P''_{x^n}, P_{t^N})$.

Our final comment regards the comparison of the payoff at the equilibrium for the $SI_{tr,c}^{lr}$ and the SI_{ks}^{lr} games. Let us consider the two optimal acceptance regions, that for sake of clarity we will indicate with $\Lambda_{0,ks}^*$ and $\Lambda_{0,tr}^*$. The comparison between $\Lambda_{0,ks}^*$ and $\Lambda_{0,tr}^*$ is not straightforward since the former depends only on P_{x^n} (for a given P_X) while the latter depends both on P_{x^n} and P_{t^N}. In order to ease the comparison we assume that $P_X \in \mathcal{P}_n$ and that P_{t^N} is also fixed and equal to P_X. We can show that under this assumption, and

for large n, we have $\Lambda^*_{0,ks} \subseteq \Lambda^*_{0,tr}$. To do so we note that with some algebra the log-likelihood ratio can be rewritten in the following form:

$$h(P_{x^n}, P_{t^N}) = \mathcal{D}(P_{x^n}||P_{t^N}) - \frac{N+n}{n}\mathcal{D}(P_{r^{n+N}}||P_{t^N}). \quad (23)$$

From the above equation we see that $h(P_{x^n}, P_{t^N}) \leq \mathcal{D}(P_{x^n}||P_{t^N})$, hence for $P_{t^N} = P_X$ and n large enough[4], the acceptance region for the game with training data contains that of the game with known sources. As a consequence, it is easier for the AD to bring a sequence y^n generated by a source Y within $\Lambda^*_{0,tr}$ and fool the FA. Version c of the SI^{lr}_{tr} game is then more favorable to the attacker than the SI^{lr}_{ks} game. While, the above argument holds only when $P_{t^N} = P_X$, we argue that this is the case even in a general setting. We leave a rigorous proof of the above property to a subsequent work.

VI. Conclusions

Following the definition of the SI_{ks} game, extensively treated in [2], [4], we took a further step towards the construction of a theoretical background for multimedia forensics. The source identification game with training data, in fact, is significantly closer to real applications than the game with know sources. The solution of version c of the game provided interesting insights into the optimal strategies for the FA and the AD, that somewhat differ from those that one would have obtained by simply extending the optimum strategies of the known sources case. Additional, even more interesting, results are likely to derive from the solution of versions a and b of the SI_{tr} game, which will be the goal of our future work, together with the analysis of the optimal strategies and the resulting payoff for specific cases of particular interest (e.g. for Bernoulli sources). Other interesting directions for future research include the analysis of a version of the game in which the test sequence x^n may have been generated by a (limited) number of sources each known through training sequences. The extensions of the analysis to sources with memory and continuous sources is also worth attention.

Acknowledgment

This work was partially supported by the REWIND Project funded by the Future and Emerging Technologies (FET) program within the 7FP of the European Commission, under FET-Open grant number: 268478.

References

[1] 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 International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, 25-30 March 2012.

[2] M. Barni, "A game theoretic approach to source identification with known statistics," in *ICASSP 2012, IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, 25-30 March 2012.

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

[4] M. Barni and B. Tondi, "The source identification game: an information-theoretic perspective," *submitted to IEEE Transactions on Information Forensics and Security*.

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

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

[7] J. Nash, "Equilibrium points in n-person games," *Proceedings of the National Academy of Sciences*, vol. 36, no. 1, pp. 48–49, 1950.

[8] M. Goljan, J. Fridrich, and M. Chen, "Sensor noise camera identification: countering counter forensics," in *SPIE Conference on Media Forensics and Security, San Jose, CA*, 2010.

[9] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 401–408, March 1989.

[10] M. Kendall and S. Stuart, *The Advanced Theory of Statistics, vol. 2, 4th edition*. New York: MacMillan, 1979.

Appendix

A. Proof of Lemma 1

We start by remembering that for a memoryless source we have [5]:

$$n\mathcal{D}(P_{x^n}||P_X) = -\log(P_X(x^n)) - nH(P_{x^n}). \quad (A1)$$

By applying the above property to the right-hand side of equation (11), we obtain:

$$n\mathcal{D}(P_{x^n}||P_X) + N\mathcal{D}(P_{t^N}||P_X) = \quad (A2)$$
$$- nH(P_{x^n}) - NH(P_{t^N}) - \log P_X(r^{n+N}),$$

where we have used the memoryless nature of P_X due to which $P_X(r^{n+N}) = P_X(t^N) \cdot P_X(x^n)$. For any $P_X \in \mathcal{C}$ we also have[5]:

$$P_X(r^{n+N}) \leq \prod_{a\in\mathcal{X}} P_{r^{n+N}}(a)^{N_{r^{n+N}}(a)}, \quad (A3)$$

where $N_{r^{n+N}}(a)$ indicates the number of times that symbol a appears in r^{n+N}, and where equality holds if $P_X(a) = P_{r^{n+N}}(a)$ for all a. By applying the log function we have:

$$\begin{aligned}\log P_X(r^{n+N}) &\leq \log \prod_{a\in\mathcal{X}} P_{r^{n+N}}(a)^{N_{r^{n+N}}(a)} \quad (A4)\\ &= \log \prod_{a\in\mathcal{X}} P_{r^{n+N}}(a)^{(N_{x^n}(a)+N_{t^N}(a))}\\ &= \sum_{a\in\mathcal{X}} N_{x^n}(a)\log P_{r^{n+N}}(a) +\\ &\quad \sum_{a\in\mathcal{X}} N_{t^N}(a)\log P_{r^{n+N}}(a).\end{aligned}$$

By inserting the above inequality in (A2), and by using the definition of empirical KL divergence we obtain:

$$\begin{aligned}&n\mathcal{D}(P_{x^n}||P_X) + N\mathcal{D}(P_{t^N}||P_X) \quad (A5)\\ &\geq \sum_{a\in\mathcal{X}} N_{x^n}(a)\log\frac{P_{x^n}(a)}{P_{r^{n+N}}(a)} + \sum_{a\in\mathcal{X}} N_{t^N}(a)\log\frac{P_{t^N}(a)}{P_{r^{n+N}}(a)}\\ &= n\mathcal{D}(P_{x^n}||P_{r^{n+N}}) + N\mathcal{D}(P_{t^N}||P_{r^{n+N}}),\end{aligned}$$

where the equality holds if $P_X = P_{r^{n+N}}$, thus completing the proof.

[4] If n is large the terms $\frac{\log(n+1)}{n}$ and $\frac{\log(n+1)(N+1)}{n}$ in $\Lambda^*_{0,ks}$ and $\Lambda^*_{0,tr}$ tend to zero.

[5] Relationship (A3) can be easily proved by resorting to Jensen's inequality.