COPING WITH THE ENEMY: ADVANCES IN ADVERSARY-AWARE SIGNAL PROCESSING

Mauro Barni

University of Siena. Dept. of Information Eng.
Via Roma 56, 53100, Siena - ITALY
barni@dii.unisi.it

Fernando Pérez-González

University of Vigo. ES Telecomunicacion
Campus Universitario, 36310 Vigo- SPAIN
fperez@gts.uvigo.es

ABSTRACT

This paper is a first attempt to provide a unified framework for studying signal processing problems where designers have to cope with the presence of an adversary, including media forensics, watermarking, adversarial machine learning, biometric spoofing, etc. We focus on the binary decision problem and discuss which strategies the adversary can use to flip the decision output at minimal cost, including blind sensitivity attacks and hill-climbing attacks. As the defender can also play smarter by considering the presence of a rational adversary, we introduce a game-theoretic approach where some advances have been recently made. We conclude by discussing some trends raised by this game-theoretic formulation.

Index Terms— Game-theory, adversary, watermarking, forensics, biometric spoofing, reputation, binary decision.

1. INTRODUCTION

Security-oriented applications of signal processing are receiving increasing attention. Multimedia forensics, biometrics, digital watermarking, steganography and steganalysis, network intrusion detection, traffic monitoring, video-surveillance, are just a few examples of such an interest. Despite enormous differences, a unique feature characterizes all these fields: the presence of one or more adversaries aiming at making the system fail.

So far, research in these disciplines has been carried out by different communities with no or very few interactions among them. It is no surprise, then, that similar solutions are re-invented several times, and that the same problems are faced with again and again by ignoring that satisfactory solutions have already been discovered in contiguous fields. Similar errors are also repeated, e.g., security requirements are misunderstood. In watermarking, for instance, it took several years to recognize that robustness and security are contrasting requirements calling for the adoption of different countermeasures. In a similar way, security issues in biometric research are often neglected, privileging pattern recognition issues more related to robustness than security. Similar concerns apply to several other fields. As a natural consequence, resources are wasted and advances proceed at a much lower pace than it would be desirable. Even worse, the lack of a unifying view does not permit to grasp the essence of the addressed problems and work out truly effective and general solutions. Times are ripe to go beyond this limited view and lay the basis of a general theory that takes into account the impact that the presence of an adversary has on the design of effective signal processing tools, i.e. a theory of *adversarial signal processing* (Adv-SP).

Some scattered steps in the right direction have already been taken: watermarking security is now clearly distinguished from watermark robustness [1]; multimedia forensics in the presence of an adversary has raised the attention of researchers and counter-forensics (and even counter-counter forensics) techniques are now studied [2, 3]; the study of anti-spoofing techniques is an active research field [4]; adversary-aware machine learning is investigated with applications to spam filtering, network intrusion detection and malware detection[5, 6]. It is the aim of this paper to review the most recent advances in the fields where signal processing designers have to cope with the presence of an adversary, highlighting the similarities between the existing approaches, and provide a unitary view for one of the most commonly encountered problems, namely binary decision.

2. BROWSING PRIOR ART

The signal processing fields wherein the presence of an adversary can not be neglected include many diverse disciplines: Multimedia Forensics (MF), adversarial machine learning [5], watermarking, steganography and steganalysis, traffic analysis [7] and intrusion detection [8] are among the most popular ones. Other examples include biometric spoofing [4], security of reputation systems [9], cognitive radio [10], fingerprinting and traitor tracing, content based information retrieval [11], and many others. In some cases, researchers are well aware

Research supported by the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission under project REWIND (FET-Open grant number: 268478)

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

of the challenges set by the presence of an adversary and have started addressing them. In other cases, such awareness has still to be fully developed and the presence of an adversary is treated only in some scattered works.

While each of the above fields has its own peculiarities, there are several common problems, whose solution under a unified framework would speed up the understanding of the associated security problems and the development of effective solutions. Doubtlessly, the most studied problem is hypothesis testing, or binary decision. In MF, for instance, a common problem is to decide whether a given document (e.g. a still image) has been generated by a given source (a specific camera or a camera model), or if a certain signal has undergone a given processing or not. A very popular example in machine learning is spam filtering, i.e. the classification of e-mails 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 many other cases, it is important that malevolent users are distinguished from fair ones (again a binary classification problem); this is the case, for instance, of reputation systems where malevolent feedbacks have to be discarded, and cognitive radio, where primary users must be distinguished from cheating secondary users. Despite the similarity of the addressed problems, the most common approach to face with them is by far the development of ad-hoc solutions specifically tailored to fit a given application. Yet, a closer look reveals that a similar rationale exists behind some of the most popular techniques developed so far. When the answer of the detector is available, the adversary can exert his attack by querying an oracle. This is the case of the sensitivity and BNSA attacks in watermarking [12, 13], the hill climbing attack in biometric masquerade attacks [4], or the ACRE attack in machine learning [6]. Countermeasures also rely on similar approaches, starting from classical security by obscurity mechanisms, in which the access to the detector is denied to the attacker, to more sophisticated approaches like detector randomization [14, 15] or the adoption of complicated detection regions [15, 16].

By referring to MF, some early attempts to define a general framework for adversarial hypothesis testing have been made, including [2, 17], where game-theory and information theory are used to derive the Nash equilibrium point of the source identification game, and [3] where the Kullback-Leibler distance is used to measure the validity of an attack regardless of the adopted countermeasures. Similar attempts have been carried out in the field of adversarial machine learning (see [5] for a review).

A possible generalization of the basic binary decision problem regards the number of involved players. In attacks against reputation systems, for instance, several attackers may pool to degrade the performance of the system [9], leading to a multiple-player game. A similar situation is encountered in traitor tracing systems, with the noticeable difference that in this case active techniques like fingerprinting may be used to improve the performance of the system. Attacks against reputation systems introduce yet another perspective into the picture: the collaborative nature of the to-be-performed tasks and the attacks. In addition to the presence of multiple players, this requires that proper solutions are adopted to either encourage fair behaviors, e.g. through the definition of a suitable pay-off function, or to allow cross-checking between users, as commonly done in sentiment tagging applications. In these cases, the presence of a large number of independent users, with a vast majority of fair users, ensures the proper behavior of the system.

Adversarial machine learning enriches the Adv-SP picture with a new original twist, due to the possibility for the adversary to attack the system during the learning phase [5]. This leads to a significant complication of the game-theoretic framework making it very difficult to determine the equilibrium point of the game. In most applications, the players act in a multi-round sequential fashion, during which they can adapt their strategies according to the moves of the adversary, thus leading to a natural formulation of the problem as a sequential game. In [18], the possibility of adopting methods typical of robust statistics is advanced, so to minimize the impact that the injection of a limited amount of fake training data has on the accuracy of the system. Still in [5], the resort to data fusion is seen as a possible way to minimize the influence of targeted attacks, as in [19], where spatial and temporal features are fused to aid the detection of fraudulent feedbacks.

While the above review is by no means intended to be an exhaustive one, it clearly shows how similar problems are encountered in several fields, thus confirming the advisability of developing a general theory that encompasses all of them.

3. LOOKING FOR A GENERAL FRAMEWORK: THE CASE OF ADVERSARIAL BINARY DECISION

After the general view we gave in the previous section, we now delve into one specific problem, namely adversarial binary decision, showing how the advances made in various fields actually fit into a unique framework

We decided to focus on the binary decision problem because this is by far the most studied scenario and most of the lessons learned here can be extended to the case of multiple decisions. Moreover, from the adversarial point of view, some multiple-decision problems can be simplified to binary ones. We consider a binary decision function ϕ that takes a feature vector $\mathbf{x}$ and gives a binary output, i.e., $\phi(\mathbf{x}) \in \{0, 1\}$. The set of possible feature vectors is denoted by $\mathcal{X} \subset \mathbb{R}^N$. The feature vector is usually the result of a dimensionality reduction function that takes a signal in the original space and maps it into a space with fewer dimensions. This function may be proprietary, as often occurs in biometric recognition systems, although even in this case many details can be learned or reversed-engineered. We let $\mathcal{R}_j \doteq \{\mathbf{x} \in \mathcal{X} : \phi(\mathbf{x}) = j\}$, for $j = 0, 1$.

The decision function is sometimes designed to optimize a certain objective. For instance, in the binary hypothesis testing problem, ϕ is generally chosen according to the Neyman-Pearson criterion, which minimizes the miss probability subject to a bound on the false-alarm probability. In binary classification problems, ϕ minimizes the empirical risk measured over the samples in the training set. However, in many cases, ϕ is designed ad-hoc to simply test for a certain property that members in, say, $\mathcal{R}_0$ hold whereas members in $\mathcal{R}_1$ lack. The decision function can depend on some secret parameters; for example, in watermarking, ϕ (and, in consequence, $\mathcal{R}_0$ and $\mathcal{R}_1$) depend on a key that is unknown to the adversary.

Consider now the role of the adversary. Frequently, e.g. in watermarking, forensics, spam filtering, the adversary has an object $\mathbf{y} \in \mathcal{X}$ with some assigned value and is interested in modifying it the least possible so that the object retains its value while fooling the decision procedure. Formally, given $\mathbf{y} \in \mathcal{R}_0$ and a certain real-valued distortion measure $d(\cdot, \cdot)$, the adversary aims at finding some $\mathbf{y}' \in \mathcal{R}_1$ such that $d(\mathbf{y}, \mathbf{y}') \leq \tau$, where τ is an acceptability threshold.[1] Examples of distortion measures can be the Euclidean or Hamming distances. A variant of the previous adversarial target is to find $\mathbf{y}^* \in \mathcal{R}_1$ such that $d(\mathbf{y}, \mathbf{y}^*)$ is minimal. Due to space constraints, here we focus on this variant, which we call the *closest point problem*. We notice, however, that there are other scenarios (e.g., biometric spoofing) where $\mathbf{y}$ is not available and the sole purpose of the adversary is to find a valid $\mathbf{y}'$ in $\mathcal{R}_1$.

If the distortion function is convex and we assume that $\mathcal{R}_0$ is an open set, then the solution $\mathbf{y}^*$ must lie on the boundary $\delta\mathcal{R}$. If the decision function is known to the adversary, then the solution can be either obtained in closed-form or numerically. When ϕ is not fully known, the adversary may try to solve the problem by querying the decisor to learn as much as possible about ϕ or, better yet, $\delta\mathcal{R}$, to later generate $\mathbf{y}^*$. Of course, the feasibility of this solution depends on the number of queries that can be made, as in some cases the system to be attacked will stop accepting them after a number of trials. In the sequel, we will assume an unbounded number of queries, however bearing in mind the complexity issue. Due to its dependence on a secret key, the case of unknown ϕ has been extensively studied in the field of watermarking. We summarize next some of the achievements made therein.

The original *sensitivity attack* [20] is suitable when ϕ is a hyperplane. It starts with a vector $\mathbf{y} \in \mathcal{R}_0$ and modifies it to $\mathbf{z} \in \mathcal{R}_0$ near the boundary $\delta\mathcal{R}$ (this can be achieved by a binary search as long as one point in $\mathcal{R}_1$ is known). Then, it works by changing one component of $\mathbf{z}$ at a time and observing the output of the decision function to learn the normal vector that represents the hyperplane. For more complicated decision boundaries, [12] proposes an iterative approach which moves along the hyperplane tangent to the decision boundary at $\mathbf{z}$. In [21] the normal vector is obtained similarly to the sensitivity attack; from this it is immediate to obtain the approximate gradient vector at $\mathbf{z}$. Knowledge of this vector suffices to obtain a good local estimate of the decision function ϕ, as long as the adversary knows its form. A similar approach for the case of linear boundaries and l_1-norm distances has been proposed in the context of machine learning for spam filtering in the so-called *Adversarial Classifier Reverse Engineering* (ACRE) method [6].

In [13] a powerful variant of the sensitivity attack which implements Newton's descent algorithm is presented to iteratively find $\mathbf{y}^*$. The algorithm is completely blind, in the sense that no knowledge of the decision function is assumed; the first and second order local derivatives information required by the iterative algorithm are estimated by querying the decisor. The algorithm, termed Blind Newton Sensitivity Attack (BNSA), has been proven very effective in removing the watermark for a number of existing schemes; moreover, it has been used in the winning strategy in the popular BOWS contest organized by the watermarking community to measure the effectiveness of oracle-based attacks [22].

All the previous attacks assume the availability of some vector $\mathbf{t}$ in $\mathcal{R}_1$ such that given $\mathbf{y} \in \mathcal{R}_0$, there is some $\lambda \in (0, 1)$ for which $\lambda\mathbf{t} + (1 - \lambda)\mathbf{y} \in \delta\mathcal{R}$. This value is found through a binary search. When $\mathbf{t}$ is not available, as it occurs in attacks to biometric authentication systems, a *hill-climbing* search is generally performed [23],[24]. This search benefits from the fact that ϕ outputs a non-binary matching score, and starts with some $\mathbf{y}$ and modifies it along a number of arbitrary (but meaningful) directions to find a new point that is closer to $\mathcal{R}_1$ and which serves as the basis for the next iteration. Quantization of the output values of ϕ is a relatively effective countermeasure at the expense of performance [23]; in fact, had ϕ a binary output, hill-climbing would not work better than a brute-force search.

In view of the previous attacks, it seems reasonable to make the function ϕ more complicated. Several works have proposed solutions along this line: in [25] the decision boundary is 'fractalized' in an attempt to hamper the use of learning algorithms; in [12] the boundary is 'randomized' so that for points $\mathbf{z}$ close to the boundary, $\phi(\mathbf{z})$ is 0 with a certain probability; both countermeasures can be easily overcome by an adversary respectively using the 'envelope' of the fractal boundary or averaging out the boundary randomness. The decision function can be even implemented in zero-knowledge, so that the adversary cannot learn anything but the binary output by querying the decisor. Strikingly, this minimum disclosure of information (at most one bit per query) is enough for BNSA to work, especially as most existing proposals, with the exception of [26], use simple decision boundaries.

The previous paragraphs hint at the existence of a game between the decision function designer and the adversary. The most natural way to model the interplay between the dif-

[1] Actually, in most cases the distortion is measured in the original space. We use the feature space here to keep the discussion simple.

ferent goals and constraints of the decision function designer, a.k.a. the *defender*, and the adversary, is through game theory. In the following we use the approach outlined in [2] to show how game theory can be used to formulate a general binary decision game, whose analysis can shed new light on the limits and achievable performance of binary decision under adversarial conditions.

A general, 2-player game is 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 $s_{2,j}$. In a zero-sum game, the win of a player is equal to the loss of the other, so we have $u_1(s_{1,i}, s_{2,j}) + u_2(s_{1,i}, s_{2,j}) = 0$. In this case, it is enough to specify the payoff of the first player (generally indicated by u). A quite general version of the binary decision game is obtained if we assume that the feature vector can be generated by sources X and Y with known pdf's P_X and P_Y, and that the decision corresponds to determining whether $\mathbf{x}$ was generated by X or Y (with $\mathcal{R}_0$ corresponding to X and $\mathcal{R}_1$ to Y). In this probabilistic framework, the zero-sum binary decision game is defined as follows:

The set of **defender's strategies** $\mathcal{S}_D$ is the subset of $\mathcal{R}_0$ for which the false positive probability (i.e. the probability that an $\mathbf{x}$ generated by X falls in $\mathcal{R}_1$) is below a certain threshold, i.e., $\mathcal{S}_D = \{\mathcal{R}_0 : P_X(\mathbf{x} \notin \mathcal{R}_0) < P_{fp}\}$, where P_{fp} is a prescribed maximum false positive probability.

The set of **attacker's strategies** $\mathcal{S}_A$ is formed by all the functions that map a vector $\mathbf{y}$ produced by Y into a new vector $\mathbf{y}'$ subject to a distortion constraint, i.e., $\mathcal{S}_A = \{f(\mathbf{y}) : d(\mathbf{y}, \mathbf{y}') \leq D_{max}\}$, where $d(\cdot, \cdot)$ is a proper distance function and D_{max} is the maximum allowed distortion.

The **payoff function** is defined in terms of the false negative error probability (P_{fn}), namely, $u(\mathcal{R}_0, f) = -P_{fn} = -\sum_{\mathbf{y}: f(\mathbf{y}) \in \mathcal{R}_0} P_Y(\mathbf{y})$.

4. A GAME-THEORETIC OUTLOOK

A first advantage of the above game-theoretic formulation is that it permits to cast under a unique umbrella all the similar versions of the binary decision problem encountered in different applications, highlighting how most of the approaches used so far fail to recognize the existence of this game. For instance, in the closest-point approach the attacker optimizes a certain payoff assuming a fixed strategy on the defender. Then, it makes sense to ask which is the best possible move for a defender who knows that his strategy will be attacked. In fact, assuming this kind of 'worst-case' threat significantly changes the design constraints for the decision function. To ellaborate, assume a gradient descent based algorithm to solve the closest-point problem is used; then, the designer will be interested in creating a decision boundary that leads to local minima, where the search can get stuck. This, in turn, will imply that the adversary achieves a smaller payoff. Notice that the previous requirement is not equivalent to a convoluted decision boundary, as depending on the available information, the adversary may be focusing on a small region where the global minimum can be easily found.

An even more difficult problem appears when the two parties must choose their strategies beforehand without knowing anything about the other player's move. This usually corresponds to determining the Nash equilibrium(s) of the game [27], a problem that can be solved only in some very specific cases. Doing so, however would permit to: i) determine the optimum strategies of both the defender and the attacker, ii) compute the payoff at the equilibrium, iii) understand how far are the available practical solutions from the best achievable performance. Among the few works where such a problem is faced with and solved we mention [2], in which the asymptotic[2] Nash equilibrium of a game very similar to the one outlined in the previous section is derived, and [28], where a version of the game in which the sources are known through training sequences is considered. The latter case is particularly interesting since it opens the way to two very interesting generalizations. First, it could allow the analysis of some classes of adversarial learning games, where the attacker has the possibility to modify the training sequence the defender relies on. Secondly, if we assume that the training sequence used by the defender is not known to the attacker, it introduces into the picture the possibility of making the decision regions partially depend on a secret, thus getting closer to the typical scenario in watermarking.

One of the drawbacks of assuming rational adversaries is that the solutions generally lead to very conservative designs of the decision function which in turn yield a bad performance even for non-malicious users. A promising approach would be to distinguish those adversaries trying to solve the closest point problem from normal users and block access to the oracle in the former case. A striking result, applied to the detection of abnormal behavior in social or sensor networks in [29], relies on Afriat's theorem, which can be used to understand if the adversary is trying to maximize a certain payoff *without even knowing* such payoff. This is done by probing the response of the adversaries; in our case probing could be achieved by slightly changing the decision function at every query to measure whether the adversary reacts to the change. Of course, such scenario poses a challenge to the adversary who is then interested in solving the closest-point problem without being noticed.

To conclude, we expect that in the near future a general framework for Adv-SP will be developed, by combining elements of game, detection, machine learning, optimization and complexity theories. We envisage that Adv-SP will become a stimulating and challenging field whose developments will immediately find a vast number of applications.

[2] The length of the vector $\mathbf{x}$ tends to ∞.

5. REFERENCES

[1] L. Pérez-Freire, P. Comesana, J. Troncoso-Pastoriza, and F. Pérez-González, "Watermarking security: a survey," *Transactions on Data Hiding and Multimedia Security I*, pp. 41–72, 2006.

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

[3] R. Böhme and M. Kirchner, "Counter-forensics: Attacking image forensics," in *Digital Image Forensics*, H. T. Sencar and N. Memon, Eds. Springer Berlin / Heidelberg, 2012.

[4] A. K. Jain, A. Ross, and U. Uludag, "Biometric template security: challenges and solutions," in *Proc. of 13th European Signal Processing Conf.*, 2005.

[5] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," *Machine Learning*, vol. 81, pp. 121–148, 2010.

[6] D. Lowd and C. Meek, "Adversarial learning," in *Proc. of 13th ACM SIGKKD Int. Conf. Knowledge Discovery and Data Mining European*, 2005, pp. 641–647.

[7] J. Deng, R. Han, and S. Mishra, "Countermeasures against traffic analysis attacks in wireless sensor networks," in *Int. Conf. Security and Privacy for Emerging Areas in Comm. Networks*, 2005, pp. 113–126.

[8] K. M. C. Tan, K. S. Killourhy, and R. A. Maxion, "Undermining an anomaly-based intrusion detection system using common exploits," in *In Recent advances in intrusion detection (RAID)*, 2002, pp. 54–73.

[9] Y. Yang, Y. L. Sun, S. Kay, and Q. Yang, "Defending online reputation systems against collaborative unfair raters through signal modeling and trust," in *Proc. of ACM Symp. on Applied Computing*, March 2009.

[10] W. Wang, H. Li, Y. Sun, and Z. Han, "Securing collaborative spectrum sensing against untrustworthy secondary users in cognitive radio networks," *EURASIP J. on Advances in Signal Processing*, vol. 2010, 2010.

[11] T-T. Do, E. Kijak, T. Furon, and L. Amsaleg, "Deluding image recognition in SIFT-based CBIR systems," in *Proc. MiFor '10 ACM workshop on Multimedia in forensics, security and intelligence*, 2010.

[12] J-P. M. G. Linnartz and M. Van Dijk, "Analysis of the sensitivity attack against electronic watermarks in images," in *Proc. Information Hiding, LNCS Vol. 1525*, 1998, pp. 258–272.

[13] P. Comesaña, L. Pérez-Freire, and F. Pérez-González, "Blind Newton sensitivity attack," in *IEE Proc. on Information Security*. IET, 2006, vol. 153, pp. 115–125.

[14] R. Venkatesan and M. H. Jakubowski, "Randomized detection for spread-spectrum watermarking: defending against sensitivity and other attacks," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, 2005.

[15] K. Wang, J. J. Parekh, and S. J. Stolfo, "Anagram: A content anomaly detector resistant to mimicry attack," in *In Recent advances in intrusion detection (RAID)*, 2006.

[16] T. Furon and P. Duhamel, "An asymmetric watermarking method," *IEEE Trans. on Signal Processing*, vol. 5, no. 4, pp. 981–995, 2003.

[17] 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.

[18] D. Wagner, "Resilient aggregation in sensor networks," in *Proc. of the ACM workshop on security of Ad Hoc and sensor networks (SASN)*, 2004, pp. 78–87.

[19] Y. Liu and Y. 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.

[20] I. J. Cox and J. P. M. G. Linnartz, "Public watermarks and resistance to tampering," in *IEEE International Conference on Image Processing ICIP'97*, Santa Barbara, California, USA, October 1997, vol. 3, pp. 3–6.

[21] M. El Choubassi and P. Moulin, "Noniterative algorithms for sensitivity analysis attacks," *IEEE Trans. on Information Forensics and Security*, vol. 2, no. 2, pp. 113–126, 2007.

[22] P. Comesaña and F. Pérez-González, "Breaking the bows watermarking system: key guessing and sensitivity attacks," *EURASIP Journal on Information Security*, vol. 2007, 2007.

[23] C. Soutar, R. Gilroy, and A. Stoianov, "Biometric system performance and security," in *Conf. IEEE Auto. Identification Advanced Technol*, 1999.

[24] A. Adler, "Biometric system security," in *Handbook of biometrics*, A.K. Jain, P. Flynn, and A.A. Ross, Eds., pp. 381–402. Springer, 2008.

[25] M. F. Mansour and A. H. Tewfik, "LMS-based attack on watermark public detectors," in *IEEE International Conference on Image Processing, ICIP'02*, September 2002, vol. 3, pp. 649–652.

[26] J.R. Troncoso-Pastoriza and F. Pérez-González, "Zero-knowledge watermark detector robust to sensitivity attacks," in *Proceedings of the 8th workshop on Multimedia and security*. ACM, 2006, pp. 97–107.

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

[28] M. Barni and B. Tondi, "Optimum forensic and counter-forensic strategies for source identification with training data," in *WIFS 2013, IEEE Int. Work. on Information Forensics and Security*, Tenerife, Spain, 2013.

[29] V. Krishnamurthy and W. Hoiles, "Afriat's test for detecting malicious agents," *IEEE Signal Processing Letters*, vol. 19, no. 12, pp. 801–804, 2012.