Threat Models and Games for Adversarial Multimedia Forensics

Mauro Barni
University of Siena
CNIT: National Inter-university Consortium for Telecommunications
via Roma 56, Siena, Italy 53100
barni@dii.unisi.it

Benedetta Tondi
University of Siena
CNIT: National Inter-university Consortium for Telecommunications
via Roma 56, Siena, Italy 53100
benedettatondi@gmail.com

ABSTRACT

We define a number of threat models to describe the goals, the available information and the actions characterising the behaviour of a possible attacker in multimedia forensic scenarios. We distinguish between an investigative scenario, wherein the forensic analysis is used to guide the investigative action and a use-in-court scenario, wherein forensic evidence must be defended during a lawsuit. We argue that the goals and actions of the attacker in these two cases are very different, thus exposing the forensic analyst to different challenges. Distinction is also made between model-based techniques and techniques based on machine learning, showing how in the latter case the necessity of defining a proper training set enriches the set of actions available to the attacker. By leveraging on the previous analysis, we then introduce some game-theoretic models to describe the interaction between the forensic analyst and the attacker in the investigative and use-in-court scenarios.

KEYWORDS

Multimedia forensics; adversarial signal processing; threat models; game theory; simultaneous and sequential games.

ACM Reference format:
Mauro Barni and Benedetta Tondi. 2017. Threat Models and Games for Adversarial Multimedia Forensics. In *Proceedings of , Bucharest, Romania, (MFSec'17, June 6, 2017),* 5 pages.
DOI: http://dx.doi.org/10.1145/3078897.3080533

1 INTRODUCTION

Multimedia forensics in adversarial conditions has received increasing attention due to the ease with which multimedia forensic tools can be fooled by an adversary (often referred to as *attacker*), aiming at deleting the traces left within a document during the acquisition phase or as a consequence of subsequent processing [1, 9, 10, 14, 18, 22]. At the beginning, anti-counter-forensic research has been carried out in a scattered way, with the development of tools to detect the traces left by specific counter-forensic techniques [6, 11, 16, 24]. More recently, some efforts have been made to ground the race of arms between the forensic analyst and the attacker on solid theoretical bases. The most promising approach consists in modelling the interplay between the analyst an the attacker as a competitive game, in which the two players have opposite goals [3, 4, 21, 23]. In [3], for instance, the forensic analysis consists of a binary hypothesis test, where the analyst is interested in minimising the false negative error probability for a fixed value of the false positive probability, while the attacker pursues the opposite goal of maximising the false negative probability. In [21], game theory is used to model the anti-counter-forensic problem in the framework of video tampering through frame deletion. Despite these efforts, no attempt has been made to categorise the goals that an attacker may pursue, and the means and information available to him to reach his goals in different scenarios. In this paper, we take a step in this direction, by proposing a number of threat models describing the behaviour of the attacker in some of the most common forensic scenarios. The threat models can be used by researchers to clearly identify the scenario addressed in their works, and to develop suitable models, possibly assuming the form of a multimedia forensic game, so to derive the optimum strategies that the forensic analyst and the attacker should adopt in a given set-up. Expanding the horizon beyond multimedia forensics, our analysis parallels a similar effort made in the context of machine learning classification [5, 8] and biometrics [12, 20]. We are not going to provide a thorough taxonomy covering all possible applications. On the contrary, we focus on one of the most common forensic problems, that is binary decision, or hypothesis testing [2]. Our analysis considers two forensic scenarios: i) investigative operations, and ii) use of forensic evidence in court. As we will see, the security models characterising these two scenarios are quite different and pose different challenges. We will also distinguish two basic categories of forensic techniques depending on the approach used by the analyst to solve the binary decision problem, namely model-based techniques and techniques based on machine learning. In fact, the importance of and a certain arbitrariness in the choice of the training set, play an important role in the definition of the threat models for this kind of techniques. We complete our analysis by introducing a number of games, in which the opposite goals of the analyst and the attacker and the interplay between their choices are casted in a game-theoretic framework.

1.1 Notation

Throughout the paper, we assume that the forensic analysis corresponds to looking for the traces, somewhat called footprints, left by a particular event in the history of a document. As an example, the analysis may aim at detecting the presence of the traces left by a specific camera, or the cameras belonging to a certain class, within a digital image. Alternatively, the forensic analysis may look for the traces left within an image by a certain class of processing tools, like lossy compression, filtering or contrast-enhancement. Without loss of generality, we let H_1 correspond to the presence of

MFSec'17, June 6, 2017, Bucharest, Romania
© 2017 ACM. 978-1-4503-5034-1/17/06…$15.00
DOI: http://dx.doi.org/10.1145/3078897.3080533

the trace, and H_0 to its absence. Given the above, the work of the forensic analyst corresponds to the application of a binary decision function ϕ that takes a feature vector $\mathbf{x}$ as input and gives a binary output, that is $\phi(\mathbf{x}) \in \{0, 1\}$, corresponding to the choice of one hypothesis between H_0 and H_1. The feature vector is usually the result of a dimensionality reduction process that takes a signal $\mathbf{s}$ in the original space and maps it into a space with fewer dimensions. The function ϕ splits the feature space (and hence the signal space) into two regions Λ_0 and Λ_1 corresponding to the feature vectors for which ϕ is in favour of H_0 and H_1, respectively. A decision error occurs either when H_1 holds and $\phi(\mathbf{x}) = 0$, or when H_0 holds and $\phi(\mathbf{x}) = 1$. In most cases, H_1 corresponds to a significant event from an investigative point of view, so it is customary to call the former type of error a missed detection, and the latter a false alarm. The corresponding error probabilities are usually denoted missed detection, or false negative, error probability ($P_{fn} = Pr\{\mathbf{x} \in \Lambda_0 | H_1\}$), and false alarm, or false positive, error probability ($P_{fp} = Pr\{\mathbf{x} \in \Lambda_1 | H_0\}$). As an example, we may consider a forensic test aiming at deciding whether an image was captured by a given camera. When an image is said to be taken from the camera even when this was not the case, we have a false alarm event, while failing to recognise that an image was taken by the camera that actually shot it is referred to as a missed detection, or false negative, event. Even if such a terminology is not justified in all application scenarios, in the following we will adhere to this notation, its adaption to different contexts being straightforward.

In the above framework, a threat model consists of three pieces of information: i) the goal of the attacker; ii) the information that the attacker has regarding ϕ; iii) the actions that the attacker can undertake to reach his goal, e.g., if and to which extent we can modify the feature vector $\mathbf{x}$. In the following sections we introduce several possible threat models for two of the most common application scenarios, namely when the forensic analysis is used during an investigation and when the analysis is used to provide evidence in court.

2 INVESTIGATIVE SCENARIO

When the forensic tools are used as part of an investigative process, they usually must ensure a low missed detection probability, possibly at the expense of false positive error probability, which, however, should not be so high to flood the investigator with false alarms. The forensic tools are usually applied to large amounts of data, thus calling for operational simplicity. Eventually, the analyst may keep the details of the forensic algorithm he is using (or part of them) secret.

2.0.1 Goal of the adversary. In an investigative scenario, the goal of the attacker is to delete, or hide, the footprints contained in a document so to hinder their detection, since they can help the investigator to solve the case he is working on. In other words, the attacker is interested in generating a missed detection (false negative) event. Often the attacker is not interested in increasing the overall false negative error probability, but to induce a missed detection error on a specific document. On the other hand, he usually has no interest to induce a false positive error, unless in some specific cases where he aims at sidetracking the investigation.

In the following we distinguish between model-based and machine-learning (or data-driven) techniques. The latter, in fact, need to be trained on a properly defined training set which introduces an additional degree of freedom for the attacker with respect to model-based approaches.

2.1 Model-based techniques

2.1.1 Information available to the adversary. By following Kerckhoff's principle [13], we assume that the attacker knows the algorithm used by the investigator. At most we can assume that he does not know the exact implementation details. This is the case, for instance, of the data used to calibrate the model which may not be known to the attacker and hence can be assimilated to a kind of secret key. The extensive knowledge that the attacker has about the forensic tools means that he can build its own version of ϕ which is exactly equal to or a very good approximation of ϕ. Let us call such an approximation ϕ'. The attacker, can use ϕ' to understand if his attack was successful or even to exploit the output of ϕ' to carry out its attack. For instance, he could use ϕ' to implement a gradient descent, so to find the direction of the shorter path to the decision boundary, or to carry out a sensitivity attack [7]. The sensitivity attack works by changing one component of the signal at a time and observing the binary output of the decision function to learn the normal vector that (locally) represents the detection region boundary.

2.1.2 Actions available to the adversary. In an investigative scenario the attacker has a signal $\mathbf{s}$ for which $\phi(\mathbf{x}(\mathbf{s})) = 1$ ($\mathbf{x}(\mathbf{s})$ is a feature vector extracted from $\mathbf{s}$) and wants to modify it in such a way that when ϕ is applied to the modified signal $\mathbf{z}$ we have $\phi(\mathbf{x}(\mathbf{z})) = 0$. In doing so he must respect a constraint on the amount of distortion he can introduce as a result of the attack. In fact, should the attacker be allowed to modify $\mathbf{s}$ at will, he would be able to prevent footprint detection by replacing $\mathbf{s}$ with a completely different signal $\mathbf{z}$, but the attacked signal $\mathbf{z}$ would no more serve the original purpose of $\mathbf{s}$. Consider for instance the case of photographer who wants to delete from an image the footprint of the camera he used to capture it. He must do so without degrading too much the quality of the image. Mathematically speaking, the attacker looks for a signal $\mathbf{z}$ such that:

$$\begin{cases} \phi'(\mathbf{z}) = 0 \\ d(\mathbf{z}, \mathbf{s}) \leq D_{max} \end{cases} \tag{1}$$

where $d()$ is s properly defined distance function, D_{max} is the maximum distortion the attacker can introduce, and ϕ' is either equal to or a good approximation of ϕ.

2.2 Machine learning techniques

When the investigator relies on a machine learning approach to detect the trace he is looking for, we must take into account the role of the training data $\mathcal{T}$ in the definition of ϕ, which, for this reason, will be referred to as $\phi(\mathbf{x}; \mathcal{T})$. As we will see, the presence of $\mathcal{T}$ marks a significant difference with respect to the case of model-based footprint detection.

2.2.1 Information available to the adversary. While Kerckhoff's principle requires that we assume that the attacker knows ϕ, we can safely presuppose that $\mathcal{T}$ is kept secret. As a consequence, the attacker can not build an exact replica of ϕ, so he must either attack the signal **s** blindly, or build a detector based on its own training set $\mathcal{T}'$. This is not an impossible task, since the investigator can not choose $\mathcal{T}$ at random. However, it is known that the performance of machine learning methods used in forensics or in the contiguous field of steganalysis are sensitive to database mismatch [15, 19], hence making the replication of the results produced by the investigator difficult. Such a difficulty may diminish the effectiveness of attacks obtained by applying gradient descent or sensitive attacks built starting from $\phi(\mathbf{x}(\mathbf{s}); \mathcal{T}')$ instead of $\phi(\mathbf{x}(\mathbf{s}); \mathcal{T})$.

With regard to the actions available to the adversary, they are the same as for model-based forensic techniques, with $\phi(\mathbf{x}(\mathbf{s}); \mathcal{T}')$ instead of ϕ'.

3 USE IN COURT

When the footprints detected by the forensic analysis must be used in a lawsuit, the analysis must satisfy more rigid constraints dictated by the strict rules adopted in court.

3.1 Model-based techniques

As opposed to the investigative scenario now the signal **s** is not accessible to the attacker, since **s** has been seized during the investigative process and delivered to the court. Furthermore, all the details of the procedure used to detect the footprints are now public since it must be possible for all the parties in the lawsuit to re-obtain the same results.

3.1.1 Goal of the adversary. Since **s** can not be modified and ϕ is fixed, there is no way for the attacker to induce a false negative event ($\phi(\mathbf{x}(\mathbf{s})) = 0$). Still the adversary can undermine the credibility of the forensic test by inducing a sufficiently large number of false positive errors, thus showing that the false positive error probability is larger than expected.

3.1.2 Information available to the adversary. Due to the reproducibility constraint required for the use in court, the attacker has a complete knowledge of function ϕ. He can exploit such a knowledge to verify whether or not his attacks were successful and to mount gradient descend and/or sensitivity attacks.

3.1.3 Actions available to the adversary. Given that now the goal is to increase P_{fp}, the attacker aims at forging a number of examples $\{\mathbf{z}_1, \mathbf{z}_2 \dots \mathbf{z}_m\}$ such that $\phi(\mathbf{x}(\mathbf{z}_i)) = 1$ and for which the expected answer would be 0. The exact meaning of the latter requirement depends on the case at hand. Suppose, for instance, that ϕ is used to detect the footprint left within an image by a camera Y. The attacker could take a number of images which *are known* to have been taken by a camera X and modify them in such a way that the detector finds in them the footprint of camera Y. Note that even if the judge knows that the forged images have been crafted by the attacker, the mere fact that creating such forgeries is possible may undermine the validity of the forensic analysis, since the attacker may always claim that the investigator crafted the image **s** to inculpate (or exculpate) the defendant.

3.2 Machine learning techniques

The main peculiarity of the use of machine-learning methods in court regards the role and definition of the training set. If a standard training set is available and used, then machine learning methods are not different from model-based techniques, since the training set can be considered as part of the detector itself. This is rarely the case, though, so the attacker can exploit the additional degree of freedom provided by a loose definition of the training set to reach his goal.

3.2.1 Goal of the adversary. If the training data is standardised, then the only possibility for the attacker is to forge a number of false positive examples, to undermine the credibility of the detector. When this is not the case, however, he may also try to craft an ad-hoc, yet plausible, training set which gives a negative result on the to-be-analysed signal, thus raising doubts on the real presence of the footprint within it.

With regard to the information available to the adversary, as with model-based techniques, use in court of forensic evidence requires that all the details of the detector ϕ are public, including the training set.

3.2.2 Actions available to the adversary. Two classes of attacks can be carried out depending on whether a standard training set is used or not. In the former case, the attacker operates as in the case of model-based techniques, by artificially creating a number of forged signals to show that P_{fp} is too large for the detector to be credible. If no standard training set exists, the attacker may build an ad-hoc (yet *reasonable*) training set resulting in a larger false positive probability insinuating that the investigator did the same, or to show that with an alternative training set the presence within **s** of the to-be-looked footprint is not revealed. More precisely, in this case the attacker aims at creating a training set $\mathcal{T}_A$ such that

$$\phi(\mathbf{x}(\mathbf{s}); \mathcal{T}_A) = 0, \tag{2}$$

and where $\phi(\mathbf{x}(\mathbf{s}); \mathcal{T}_A)$ provides sufficiently good results on a reference test set recognised by the judge, e.g. because it is widely adopted by the scientific community. As with the model-based scenario, proving that the attacker crafted the training set with the explicit aim of causing a false negative event on **s**, does not help, since he could claim that the investigator did the same with the training set $\mathcal{T}_I$ he used to obtain $\phi(\mathbf{x}(\mathbf{s}); \mathcal{T}_I) = 1$.

A summary of the threat models that we have discussed in the paper are summarised in Table 1.

4 GAME-THEORETIC FORMULATION

In this section, we introduce a number of games associated to the threat models discussed so far. As we will see, whereas the task of the forensics analyst (FA) is always to define a test to accept or reject the hypothesis that the looked-for trace is present (H_1), the specific goal of the adversary (AD) and his behaviour depend on the specific scenario (investigative or use-in-court), thus leading to different game definitions. A 2-player game is defined by a set of strategies for first and second player, namely $\mathcal{S}_1 = \{s_{1,1} \dots s_{1,n_1}\}$ and $\mathcal{S}_2 = \{s_{2,1} \dots s_{2,n_2}\}$, and the payoff functions. Specifically, $u_l(s_{1,i}, s_{2,j})$, $l = 1, 2$ is the payoff for player l, when the players play the strategies $(s_{1,i}, s_{2,j})$. When $u_1(s_{s1,i}, s_{2,j}) = -u_2(s_{1,i}, s_{2,j})$, the game is said to be zero-sum. A profile $(s^*_{1,i}, s^*_{2,j})$ is a Nash equilibirum (NE) if $u_1(s^*_{1,i}, s^*_{2,j}) \geq u_1(s_{1,i}, s^*_{2,j})$, $\forall s_{1,i} \in \mathcal{S}_1$ and $u_2(s^*_{1,i}, s^*_{2,j}) \geq u_2(s^*_{1,i}, s_{2,j})$, $\forall s_{2,j} \in \mathcal{S}_2$. A Nash equilibrium represents a solution for a simultaneous game. For sequential games,

Threat model	**Investigation**		**Court use**		
	Model-based	*Data-driven*	*Model-based*	*Data-driven*	
Goal	False negative event		Increase P_{fp}	Increase P_{fp}	False negative event
Available information	Complete information	Type of classifier*	Complete information		
Attack	Modify the to-be-analyzed signal		Forge false positive examples	Forge false positive examples Build ad-hoc training set†	Build ad-hoc training set†
Informed about success?	YES (good confidence)	Homemade guess‡	YES	YES	

*The attacker knows the type of classifier used (e.g. a neural network with a certain architecture), but he does not know the training set used by the investigator.
†Possible only if the training set is not standardized.
‡The guess is based on a copy of the detector trained with a homemade training set.

Table 1: Summary of the threat models characterising the scenarios addressed in this paper.

where later players have some knowledge about previous actions, an equivalent equilibrium notion is given by the subgame perfect equilibrium (SPE). A profile is a SPE if it represents a NE of every subgame of the original game. SPE can be found through backward induction [17].

4.1 Investigative scenario

In this scenario, the adversary is interested in modifying the signal $\mathbf{s}$ to cause a missed detection event (see discussion in Section 2). Reasonably, both players are unaware of the action chosen by the other and play simultaneously. By assuming that the analyst adopts a Neyman-Pearson (NP) approach, we can model the interplay in the investigative scenario as in the following game:

Definition 4.1. The $G_I(\mathcal{S}_{FC}, \mathcal{S}_{AD}, u)$ is a simultaneous, zero-sum game with:

$$\begin{aligned} \mathcal{S}_{FA} &= \{\phi : Pr\{\phi(\mathbf{x}) = 1|H_0\} \leq P^*_{fp}\}, \\ \mathcal{S}_{AD} &= \{g : d(g(\mathbf{s}), \mathbf{s}) \leq D_{max}\}, \\ u(\phi, g) &= Pr\{\phi(\mathbf{x}(g(\mathbf{s}))) = 0|H_1\} = -P_{fn}, \end{aligned} \quad (3)$$

for a prescribed maximum false positive value P^*_{fp}. We have that $u = u_{FA} = -u_{AD}$.

Then, the analyst can choose any statistical test which guarantees the false alarm constraint, whereas the attacker can choose any function satisfying the distortion constraint. Solving the game means looking for the existence of Nash equilibria. When the features extracted from $\mathbf{s}$ correspond to the empirical probability mass function (pmf) of the signal, that is, $\mathbf{x}(\mathbf{s}) = \hat{P}_{\mathbf{s}}$ (i.e., the FA performs a first order analysis), an asymptotic version of the above game has been solved in [3]. In this case, the winning regions for the two players have also been determined. The theoretical results of the game analysis derived in [3] have been applied in [1] to the case of contrast-enhancement detection.

For the data-driven case, the probability under the two hypotheses is estimated from the training set $\mathcal{T}$ (i.e., the evidence provided by the training is used as an estimation for the model). Reasonably, the attacker will rely on a set $\mathcal{T}'$ different from $\mathcal{T}$, thus basing his action on a different estimation. An asymptotic version of such game has been solved in [4] for the case of first order analysis.

4.2 Use in court

We focus on the case in which the attacker aims at forging false positive examples. Accordingly, under hypothesis H_0 (absence of the trace), the signal under inspection may have been forged by the attacker to induce the analyst to decide that the trace is present. Formally, we let α be the probability that the signal is forged by the attacker (this determines the percentage of forgeries on the total number of signals under H_0 examined at the court).

For the court scenario, it is natural to define the game as a sequential game where the attacker plays second and perfectly knows the action chosen by the analyst. We observe that, since in this case the attacker can not influence the behavior under H_1 (i.e., act on the false negative probability), it is reasonable for the defender to put a constraint on P_{fn} and consider a NP test. Then, for the court scenario we give the following definition:

Definition 4.2. The $G_C(\mathcal{S}_{FA}(1), \mathcal{S}_{AD}(2), u)$ game is a sequential, zero-sum game in which the analyst plays first and the attacker plays second, where:

$$\begin{aligned} \mathcal{S}_{FA}(1) &= \{\phi : Pr\{\phi(\mathbf{x}) = 0|H_1\} \leq P^*_{fn}\}, \\ \mathcal{S}_{AD}(2) &= \{g : d(g(\mathbf{s}), \mathbf{s}) \leq D_{max}\}, \\ u(\phi, g) &= -P_{fp}, \end{aligned} \quad (4)$$

where P^*_{fn} is a prescribed false negative error probability and

$$P_{fp} = (1 - \alpha)Pr\{\phi(\mathbf{x}(\mathbf{s}))=1|H_0\} + \alpha Pr\{\phi(\mathbf{x}(g(\mathbf{s})))=1|H_0\}.$$

It is easy to derive the subgame perfect equilibrium of the G_C game corresponding to the profile (ϕ^*, g^*) satisfying

$$\max_{\phi \in \mathcal{S}_{FA}} \min_{g \in \mathcal{S}_A} u(\phi, g).$$

This is indeed the solution we obtain by maximising the utilities of the analyst and the attacker according to the playing order (backward induction). The above game definition also holds for the case of machine learning techniques, by replacing $\phi(\mathbf{x})$ with $\phi(\mathbf{x}; \mathcal{T})$[1].

5 CONCLUSIONS

In this paper, we have discussed a number of threat models for adversarial multimedia forensics suited to describe the behaviour of the attacker in two of the most common forensic scenarios, namely the investigative and court scenarios. We have shown that the two scenarios are characterised by different goals, available information and possible actions for the attacker. We have also defined a number of games associated to the threat model introduced. We believe this work will help guiding further research in the field and contribute to a systematic development of multimedia forensics as a solid and well-funded discipline.

[1] We are considering the case in which the training set is standardised.

REFERENCES

[1] M. Barni, M. Fontani, and B. Tondi. 2012. A universal technique to hide traces of histogram-based image manipulations. In *Proc. of the ACM Multimedia and Security Workshop*. Coventry, UK.

[2] M. Barni and F. Pérez-González. 2013. Coping with the Enemy: advances in Adversary-Aware Signal Processing. In *ICASSP 2013, IEEE Int. Conf. Acoustics, Speech and Signal Processing*.

[3] M. Barni and B. Tondi. 2013. The source identification game: an information-theoretic perspective. *IEEE Trans. Information Forensics and Security* 8, 3 (2013).

[4] M. Barni and B. Tondi. 2014. Binary Hypothesis Testing Game with Training Data. *IEEE Trans. on Information Theory* (2014).

[5] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar. 2010. The security of machine learning. *Machine Learning* 81, 2 (2010), 121–148.

[6] A. Costanzo, I. Amerini, R. Caldelli, and M. Barni. 2014. Forensic analysis of SIFT keypoint removal and injection. *IEEE Trans. on Information Forensics and Security* 9, 9 (2014).

[7] Ingemar J Cox and Jean-Paul MG Linnartz. 1997. Public watermarks and resistance to tampering. In *International Conference on Image Processing (ICIPfi97)*. 26–29.

[8] N. Dalvi, P. Domingos, P. Mausam, S. Sanghai, and D. Verma. 2004. Adversarial classification. In *Proc. of the tenth ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*. 99–108.

[9] D. T. Dang-Nguyen, I. D. Gebru, V. Conotter, G. Boato, and F. De Natale. 2013. Counter-forensics of median filtering. In *IEEE 15th Int. Workshop on Multimedia Signal Processing*.

[10] T. Gloe, M. Kirchner, A. Winkler, and R. Bohme. 2007. Can we trust digital image forensics ?. In *ACM Multimedia 2007, Augsburg, Germany*.

[11] M. Goljan, J. Fridrich, and M. Chen. 2011. Defending against fingerprint-copy attack in sensor-based camera identification. *IEEE Trans. on Information Forensics and Security* 6, 1 (2011).

[12] A. K. Jain, A. Ross, and U. Uludag. 2005. Biometric template security: Challenges and solutions. In *Proc. of EUSIPCO'05, European Signal Processing Conference*.

[13] A. Kerckhoffs. 1883. La cryptografie militaire. *Journal des Sciences Militaire* 9 (1883), 5–38.

[14] M. Kirchner and R. Bohme. 2008. Hiding traces of resampling in digital images. *IEEE Trans. on Information Forensics and Security* 3, 4 (2008).

[15] J. Kodovsky, V. Sedighi, and J. Fridrich. Study of cover source mismatch in steganalysis and ways to mitigate its impact. In *IS&T/SPIE Electronic Imaging*.

[16] S. Y. Lai and R. Böhme. 2011. Countering counter-forensics: the case of JPEG compression. In *Information Hiding*. Springer.

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

[18] R. Böhme and M. Kirchner. 2012. Counter-Forensics: Attacking Image Forensics. In *Digital Image Forensics*, H. T. Sencar and N. Memon (Eds.). Springer Berlin / Heidelberg.

[19] D. Ramos, J. Gonzalez-Rodriguez, J. Gonzalez-Dominguez, and Jose Juan J. J. Lucena-Molina. 2008. Addressing database mismatch in forensic speaker recognition with Ahumada III: a public real-casework database in Spanish. In *Interspeech*. International Speech Communication Association.

[20] C. Roberts. 2007. Biometric attack vectors and defences. *Computers & Security* 26, 1 (2007).

[21] M. C. Stamm, W. S. Lin, and K. J. R. Liu. 2012. Forensics vs Anti-Forensics: a Decision and Game Theoretic Framework. In *Proc. of ICASSP 2012, IEEE Int. Conf. Acoustics, Speech and Signal Processing*. Kyoto, Japan.

[22] M. C. Stamm and K. J. R. Liu. 2011. Anti-forensics of digital image compression. *IEEE Trans. on Information Forensics and Security* 6, 3 (2011).

[23] H. Zeng, X. Kang, and J. Huang. 2013. Mixed-strategy Nash equilibrium in the camera source identification game. In *Proc. ICIP 2013, 20th IEEE Int. Conf. on Image Processing*.

[24] H. Zeng, T. Qin, X. Kang, and L. Liu. 2014. Countering anti-forensics of median filtering. In *Proc. ICASSP 2014, IEEE Int. Conf. Acoustics, Speech and Signal Processing*.