UNIVERSITÀ DI SIENA

Annual GTTI meeting, Udine, 22 June 2017

Adversarial Detection: Theoretical Foundations and Applications to Multimedia Forensics

Benedetta Tondi

Post-Doc researcher

Department of Information Engineering and Mathematics, University of Siena (Italy)

Summary

- Introduction to Adversarial Signal Processing
- **Adversarial Binary Detection**
- Theoretical analysis:
  - General framework for the Binary Detection problem in the presence of adversary (simple case)
- [left out] Practical analysis:
  - Applications to Multimedia Forensics

Adversarial Signal Processing (AvdSP)

Motivations:

- Every digital system is exposed to *malicious* threats
- Security-oriented disciplines have to cope with the presence of adversaries
  - Watermarking - fingerprinting
  - Multimedia forensics
  - Spam filtering
  - intrusion detection
  - ….and many others

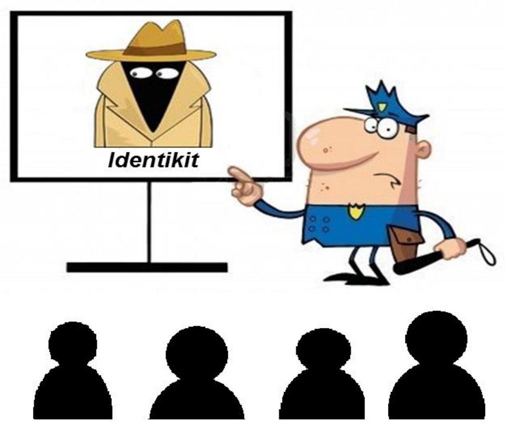
- Researchers have started looking for countermeasures, with *limited interaction*.

Adversarial Signal Processing (AvdSP)

- These fields face with similar problems
  - e.g. oracle attacks (in watermarking, in biometrics, in machine learning)
- ….and countermeasures are similar

Idea: a **unified view**

- ✓ catch the real essence of the problems
- ✓ work out effective and general solutions
- ✓ avoid the cat&mouse….



Tools: ***Game Theory*** -> a good fit !

Game Theory in a nutshell

Two players, strategic game

$$G(S_1, S_2, u_1, u_2)$$

$S_1 = \{s_{1,1}, s_{1,2}, ..., s_{1,m_1}\}$ Set of strategies of Player 1

$S_2 = \{s_{2,1}, s_{2,2}, ..., s_{2,m_1}\}$ Set of strategies of Player 2

$u_1(s_{1,i}, s_{2,j})$ Payoff of Player 1 for a given profile $(s_{1,i}, s_{2,j})$

$u_2(s_{1,i}, s_{2,j})$ Payoff of Player 2 for a given profile $(s_{1,i}, s_{2,j})$

Competitive (zero-sum) game

$$u_1(\cdot,\cdot) = -u_2(\cdot,\cdot) = u$$

In game theory we are interested in the optimal choices of rationale players.

Game Theory in a nutshell

Nash equilibrium

None of the players gets an advantage by changing his strategy (assuming the other does not change his own)

- Very Popular
- Often unsatisfactory (for the players)

Rationalizable equilibrium

The profile which survives to iterative elimination of strictly dominated strategies (for dominance-solvable games)

Dominated strategy

$$u_1(s_{1,k}, s_{2,j}) > u_1(s_{1,i}, s_{2,j}) \quad \forall s_{2,j} \in \mathcal{S}_2$$

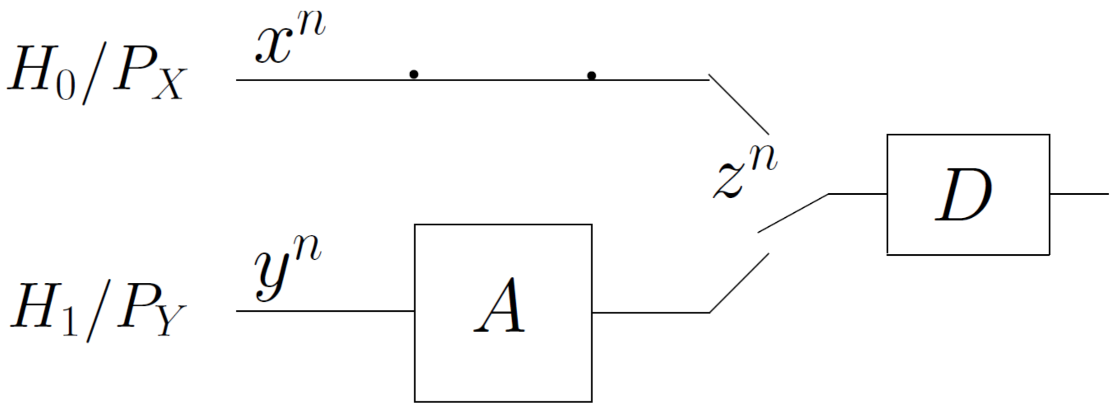
$s_{1,i}$ is strictly dominated by $s_{1,k}$

Binary Detection: a recurrent problem in SP

- Was a given image taken by a given camera ?
- Was this image resized/compressed twice ?
- Is this image a stego or a cover ?
- Does this face/fingerprint/iris belong to Mr X ?
- Is this e-mail spam or not ?
- Is traffic level indicating the presence of an anomaly/intrusion ?
- Is X a malevolent or fair user ?
  - Recommender systems, reputation handling
  - Cognitive radio

Common element: the presence of an adversary aiming at making the test fail

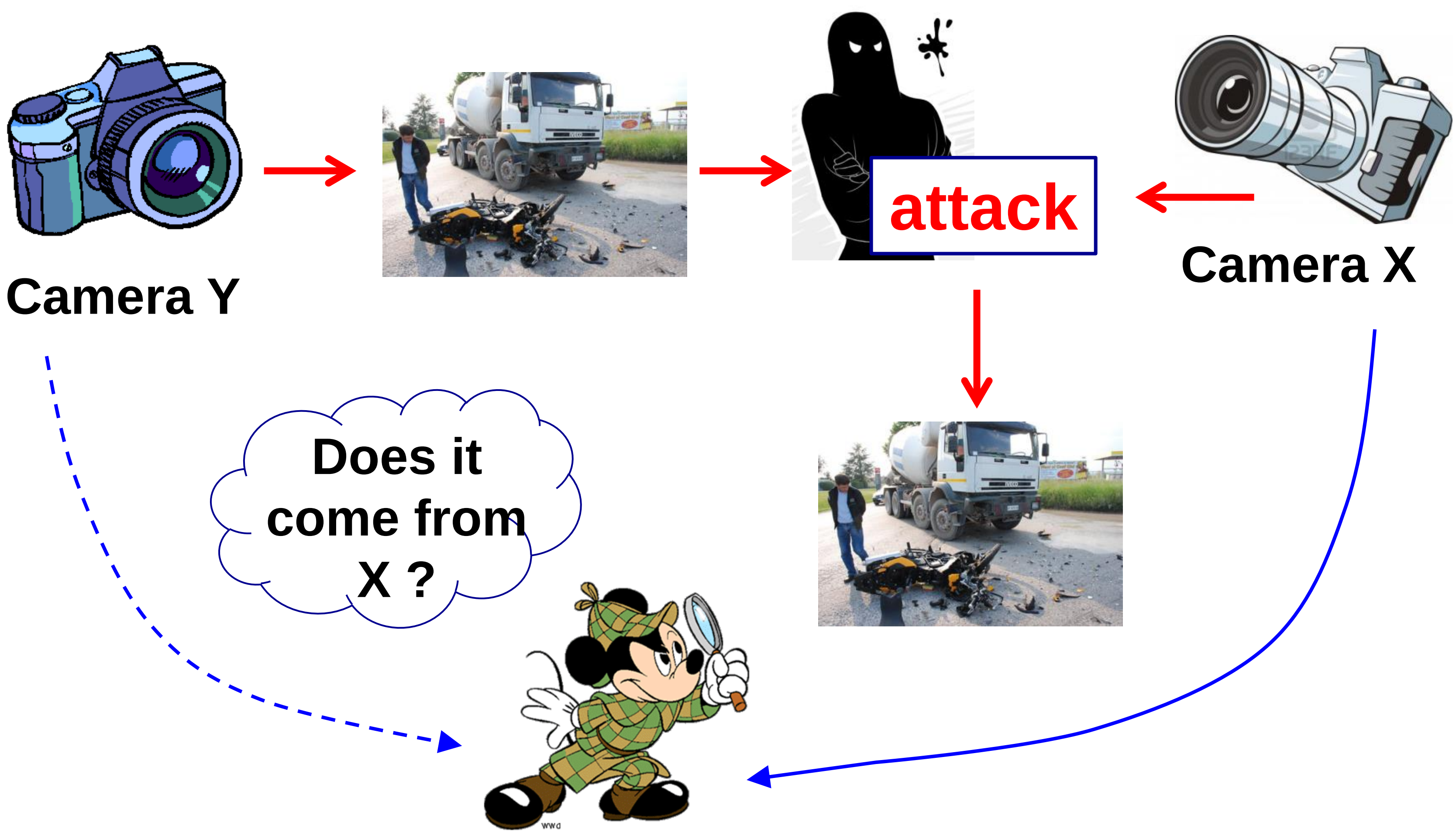
Detection problem: basic setup



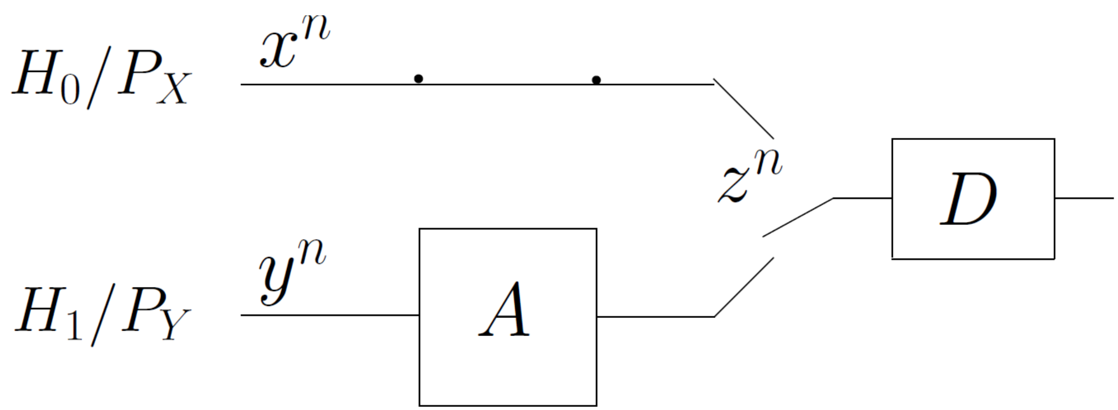
P_X and P_Y : pmf's of discrete memoryless sources X and Y

- **Goal of the Defender (D)**: decide if a sequence has been generated by P_X (under H_0) or P_Y (under H_1)
- **Goal of the Attacker (A)**: modify a sequence generated by P_Y so that it looks as if it were generated by P_X subject to a distortion constraint

A motivating example from Image Forensics



Detection problem: basic setup



P_X and P_Y : pmf's of discrete memoryless sources X and Y

- **Goal of the Defender (D)**: decide if a sequence has been generated by P_X (under H_0) or P_Y (under H_1)
- **Goal of the Attacker (A)**: modify a sequence generated by P_Y so that it looks as if it were generated by P_X subject to a distortion constraint

Starting from this setup….

- We studied the problem of the Adversarial Binary Detection in different scenarios depending on:
  - Threat setup: attack under H_0 only or under both H_0 and H_1
  - Decision setup: based on single or multiple observations
  - Knowledge available to Defender and Attacker (full or based on training data)
  - Possibility for the attacker of corrupting the training data

What we will cover….

- **Binary Detection Game with known sources**
  - Attack under H_1 only, known statistics, single observation-based decision

Binary Detection Game with known sources (DT_{ks})

The DT_{ks} game

Set of strategies for D

$$\mathcal{S}_D = \{\Lambda^n : P_{\text{FP}} \leq 2^{-\lambda n}\}$$

Λ^n defined by relying on P_{z^n} (first-order analysis)

λ decay rate (asymptotic analysis)

Set of strategies for A

$$\mathcal{S}_A = \{g(\cdot) : d(y^n, g(y^n)) \leq nL\}$$

L, maximum average per letter distortion

Payoff (zero-sum game)

$$u(\Lambda^n, g) = -P_{\text{FN}} = -\sum_{y^n : g(y^n) \in \Lambda^n} P_Y(y^n)$$

The $\mathrm{DT_{ks}}$ game: equilibrium point

Lemma **(optimum defence strategy)**

$$\Lambda^{n,*} = \left\{ P_{z^n} : \mathcal{D}(P_{z^n} || P_X) < \lambda - |\mathcal{X}| \frac{\log(n+1)}{n} \right\}$$

is a *dominant strategy* for the Defender.

Remarks:

- regardless of the attacking strategy (the optimum strategy is *dominant!*)
- regardless of P_Y (the optimum strategy is *universal* w.r.t. Y)

The DT_{ks} game: equilibrium point

Optimum attack strategy

Given that D will play the dominant strategy, A must solve a minimization problem

$$g^*(y^n) = \arg \min_{z^n : d(z^n, y^n) \leq nL} \mathcal{D}(P_{z^n} || P_X)$$

Theorem (equilibrium point): the profile $(\Lambda^{n,*}, g^*)$ is the only **rationalizable equilibrium** of the game

The DT_{ks} game: who wins?

Theorem (asymptotic payoff at the equilibrium)

Given P_X, λ and L, it is possible to define a region Γ for which we have:

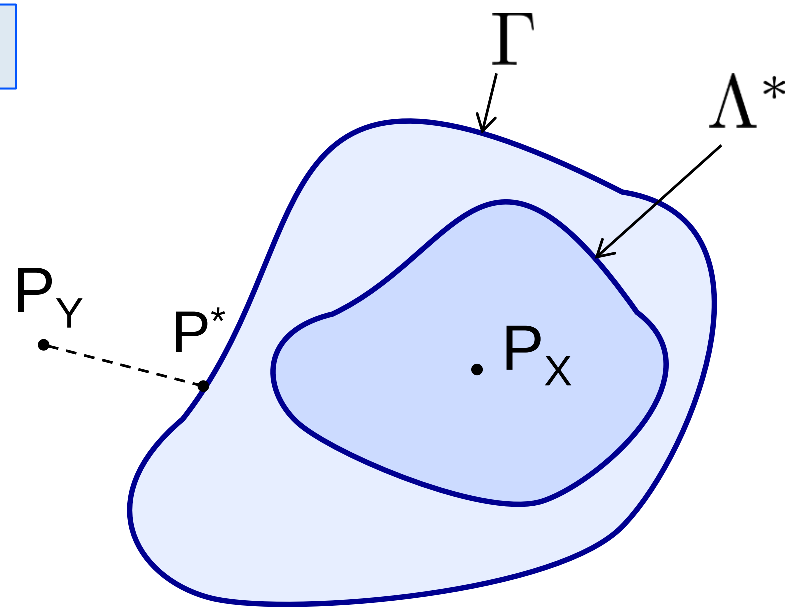
$$\begin{cases} P_Y \in \Gamma, & \text{then } P_{FN} \to 1 \\ P_Y \notin \Gamma, & \text{then } P_{FN} \to 0 \end{cases}$$

A wins

D wins

In the latter case we have:

$$\varepsilon = \min_{P \in \Gamma} \mathcal{D}(P || P_Y)$$



Γ -> ***indistinguishability region of the test***

(set of the pmf's P that cannot be distinguished from P_X)

The Security Margin (in the DT_{ks} setup)

Given Px and Py…..

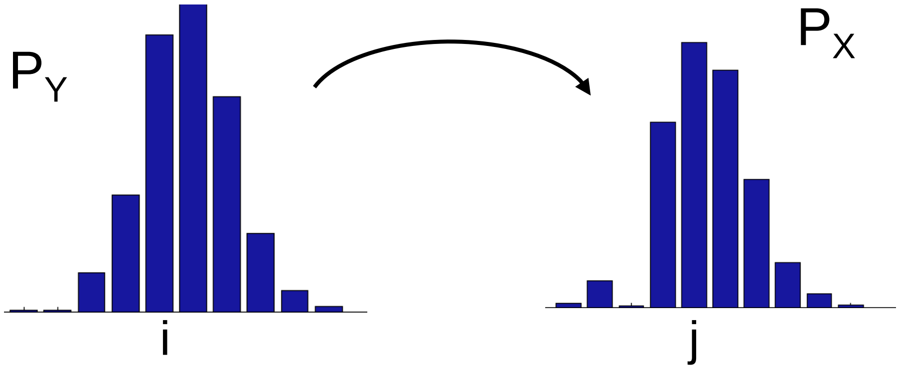
Security Margin between P_X and P_Y = maximum L for which P_X and P_Y can be *reliably* distinguished, $\mathcal{SM}(P_Y, P_X)$

SM and Optimal Transport

If we interpret P_Y and P_X as two different ways of piling up a certain amount of soil…..

The **Earth Mover Distance (EMD)** is the *minimum cost* necessary to transform P_Y into P_X

$$\mathcal{SM}(P_Y, P_X) = EMD(P_Y, P_X)$$



Further work

- Extension to
  - higher-order statistics (adversary-aware data driven classification)
  - continuous sources (on-going)
  - sources with memory
- Multiple-hypothesis testing or classification
- Applications to other fields (not only MM-Forensics)

References

CONFERENCE PUBLICATIONS

M. Barni, M. Fontani, B. Tondi. “A Universal Technique to Hide Traces of Histogram- Based Image Manipulations". In proc. of the 14th ACM workshop on Multimedia and Security, MMSEC 2012.

M. Barni, B. Tondi. “Optimum Forensic and Counter-forensic Strategies for Source Identification with Training Data". In Proc. of IEEE International Workshop on Information Forensics and Security, WIFS 2012.

M. Barni, B. Tondi. “Multiple-Observation Hypothesis Testing under Adversarial Conditions", Proc. of WIFS'13, IEEE International Workshop on Information Forensics and Security, 18-21 November 2013, Guangzhou, China

M. Barni, B. Tondi. “The Security Margin: a Measure of Source Distinguishability under Adversarial Conditions", Proc. of GlobalSip'13, IEEE Global Conference on Signal and Information Processing, 3-5 December 2013, Austin, Texas

M. Barni, B. Tondi. “Source Distinguishability under corrupted training". Proc. of WIFS'14, IEEE International Workshop on Information Forensics and Security, 3-5 December 2014, Atlanta, Georgia.

M. Barni, B. Tondi. “Universal Counterforensics of Multiple Compressed JPEG Images". IWDW 2014, The 13th International Workshop on Digital-forensics and Watermarking, October 01-04, 2014, Taipei, Taiwan

B. Tondi, M. Barni, N. Merhav. “Detection Games with a Fully Active Attacker". WIFS'15, IEEE International Workshop on Information Forensics and Security (WIFS), 16-19 Nov. 2015, Rome, Italy

References

JOURNAL PUBLICATIONS

M. Barni, B. Tondi, “The Source Identification Game: an Information Theoretic Perspective", IEEE Transactions on Information Forensics and Security, Vol. 8, no. 3, pp 450-463, March 2013.

M. Barni, M. Fontani, B. Tondi, “A Universal Attack Against Histogram-Based Image Forensics", International Journal of Digital Crime and Forensics (IJDCF), IGI Global, USA, Vol. 5, no. 3, 2013.

M. Barni, B. Tondi, “Binary Hypothesis Testing Game with Training Data", IEEE Transactions on Information Theory, Vol.60, no.8,pp 4848-4866, August 2014.

M. Barni, B. Tondi. “Source Distinguishability under Distortion-Limited Attack: an Optimal Transport Perspective”, IEEE Trans. on Information Forensics and Security, Vol. 11, No.10, May 2016

M. Barni, B, Tondi, “Adversarial Source Identification Game with Corrupted Training”, submitted *to IEEE Trans. on Information Theory*, on January 2017

AWARDS:

Best Student Paper Award at the IEEE International Workshop on Information Forensics and Security (WIFS), December 3-5, 2014, Atlanta, Georgia, USA

Best Paper Award at the IEEE International Workshop on Information Forensics and Security (WIFS), November 16-19, 2015, Rome, Italy

Thank you
for your attention