HKBU – May 23, 2017
Distinguished Lecture Series

Adversarial Signal Processing and the Hypothesis Testing Game

Mauro Barni
University of Sienna, Italy

Outline of the talk

- Motivation
- Adversarial signal processing and game theory
- Hypothesis testing game
  - Definition
  - Equilibrium point
  - Security Margin
- Application to Multimedia Forensics
- Conclusion

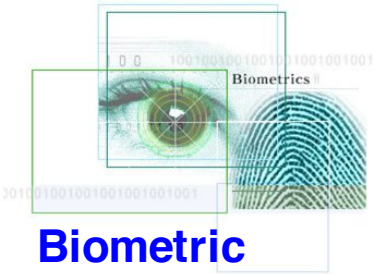
The digital ecosystem we live in

A digital paradise ?

Search enginees

Reputation scores

User-generated contents



Biometric identification

Or a battlefield ?

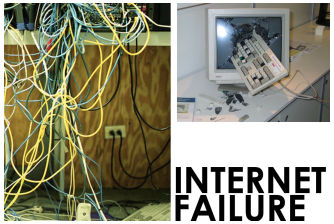
Denial of Service

Fake images

SPAM

Identity theft

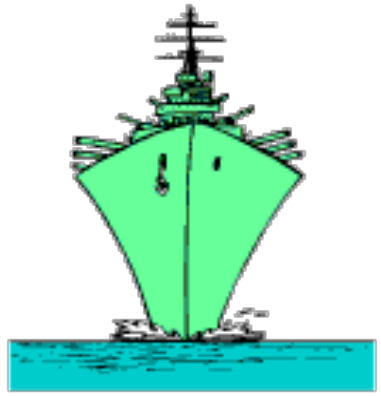
Network intrusion



To the rescue

- Researchers with diverse background have started looking for countermeasures
  - Spam filtering
  - Network intrusion detection
  - Secure reputation systems
  - Watermarking - fingerprinting
  - Multimedia forensics
  - Secure classification/learning
  - Anti-spoofing biometrics

  ... and many many others



To a closer look ...

- All these fields face with similar problems ...
- ... but interact each other to a very limited extent

We keep patching techniques thought to work in a digital paradise while we should develop tools explicitly designed for a battlefield

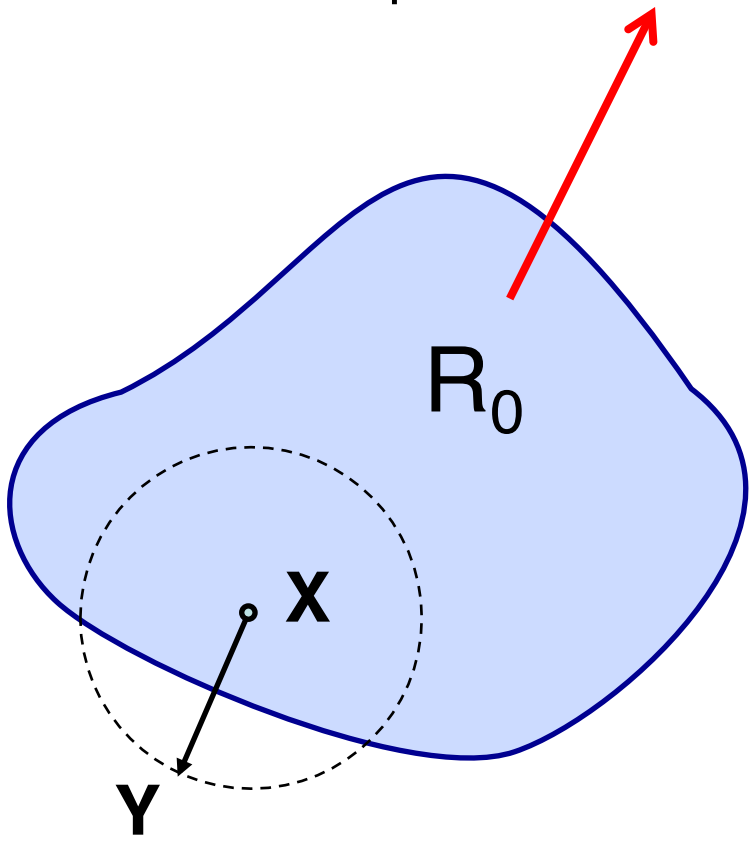
- Solutions are less effective than possible
- Basic concepts are misunderstood
  - Security vs Robustness

Binary decision: most recurrent problem

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

Attacks are also similar

- Images taken by camera X
- SPAM e-mails
- Biometric template



- Anomalous network traffic
- Malevolent users in reputation systems

  - Exit (or enter) R_0 under a distortion constraint
  - Exit (or enter) R_0 with the minimum distortion
  - If R_0 is known, then look for optimal solution
  - If R_0 is not known: oracle attacks are possible
    - **Gradient descent**

The cat and mouse loop

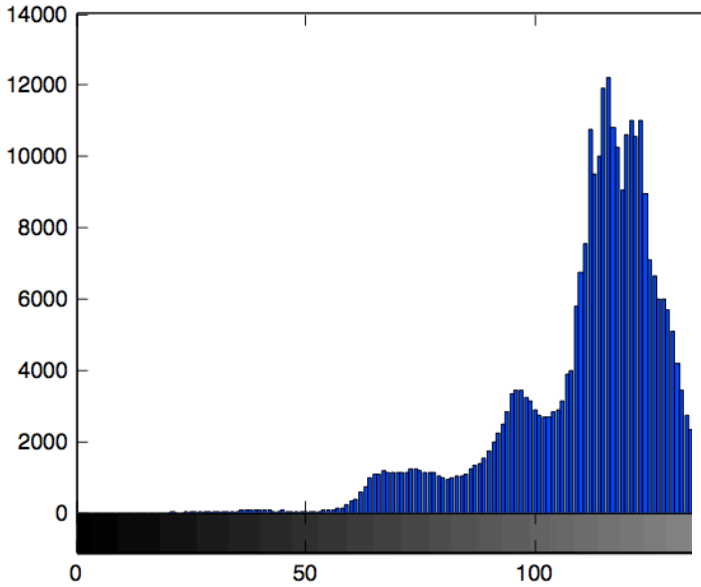
An example from Multimedia Forensics

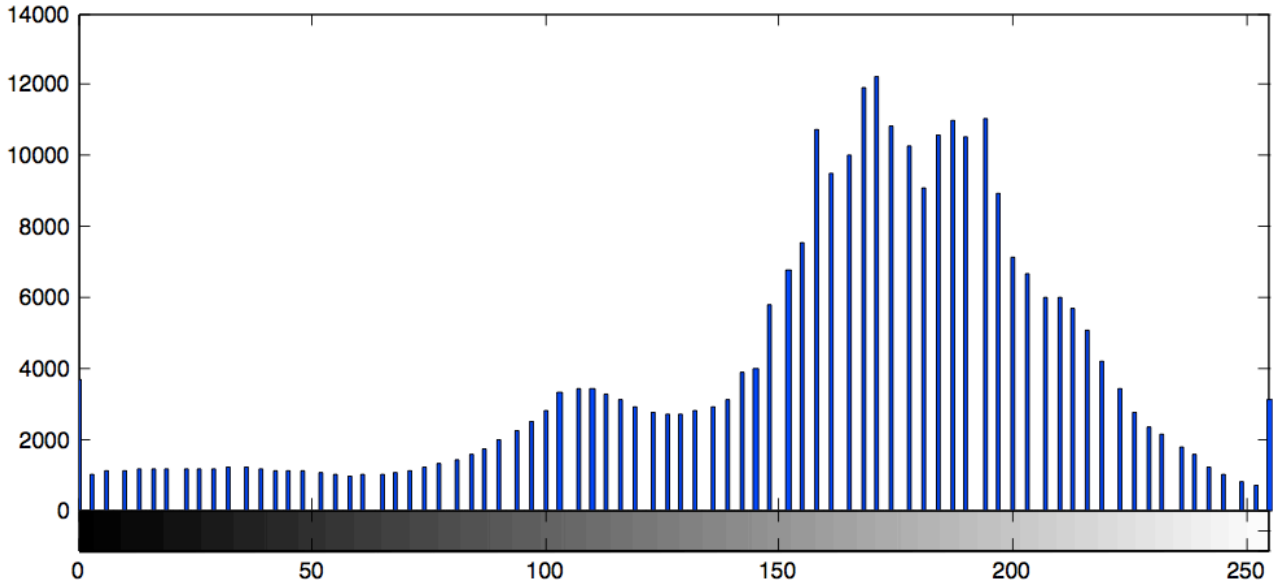
Was a certain image contrast-enhanced ?

Avoid entering a cat and mouse loop

Was a certain image contrast-enhanced ?

Look at histogram gaps !!!



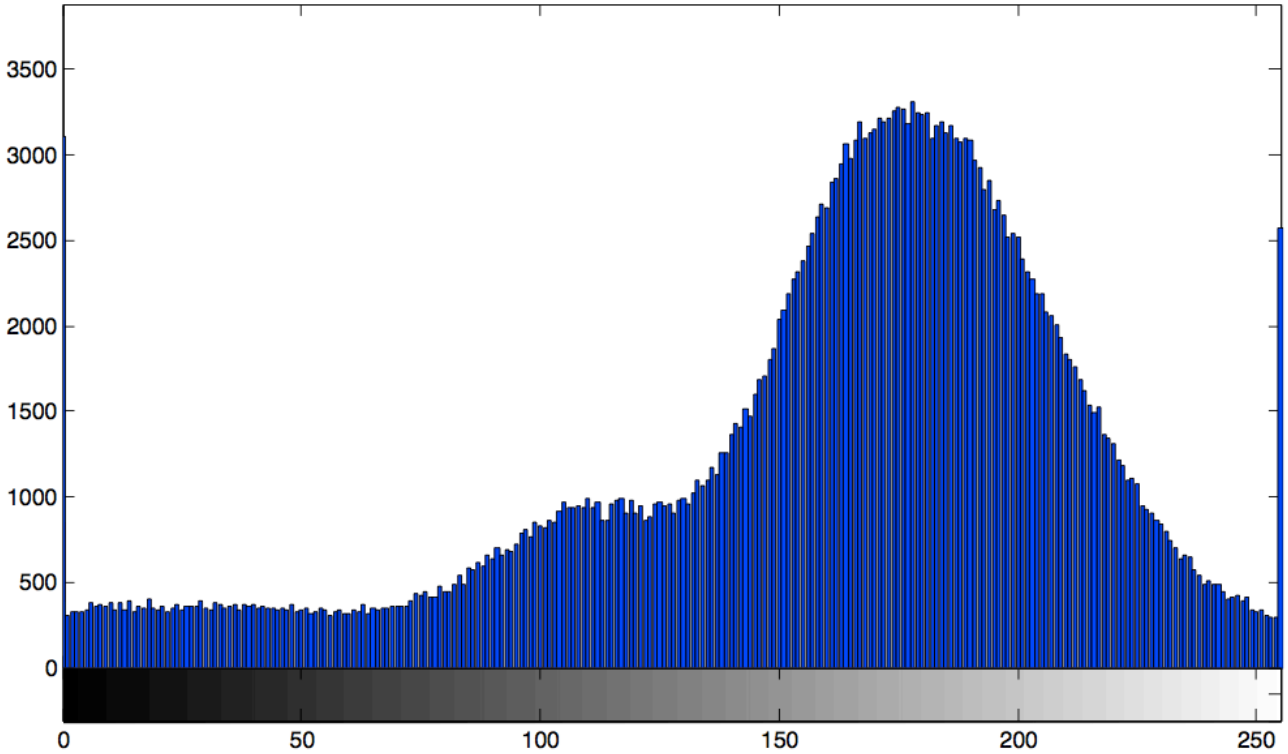


Avoid entering a cat and mouse loop

Was a certain image gamma-enhanced ?

Look at histogram gaps !!!

Add noise so to fill the gaps



Avoid entering a cat and mouse loop

Was a certain image gamma-enhanced ?

Look at histogram gaps !!!

Look at image noi

Avoid entering a cat and mouse loop

Was a certain image gamma-enhanced ?

Look at his

e gaps

Look at loc

image

Avoid entering a cat and mouse loop

Was a certain image gamma-enhanced ?

Look at histogram gaps !!!

Add noise so to fill the gaps

Look at local gradient

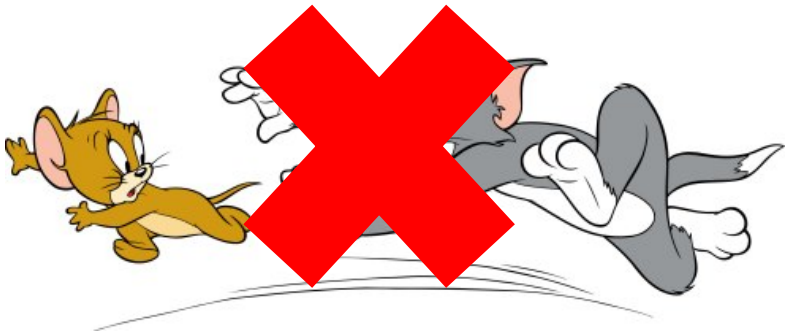
Smooth the image

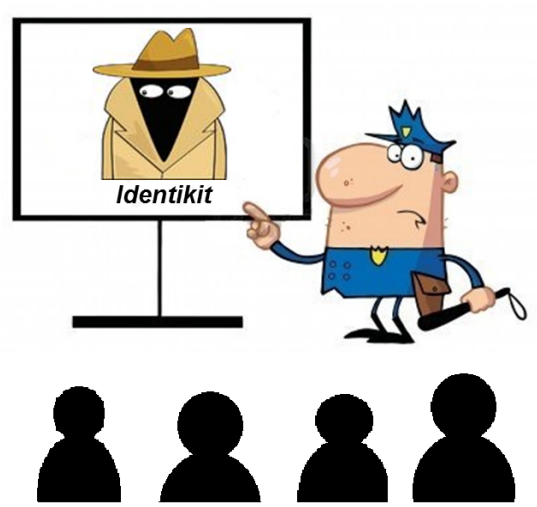
Develop a tool to detect filtered images

Adversarial Signal Processing

It is advisable to

- Avoid entering this never ending loop
- Catch the real essence of the problems
- Understand who's going to win this race of arms, at least under certain (reasonable) assumptions





Where do we start from ?

Adv-SP and Game-Theory: a perfect fit

Vast amount of results to rely on

Clear definition of **rational** players

Clear definition of goals

Optimality criteria (equilibrium notion)

Modelling social interactions

Definition of possible moves

Several game structures are possible

Game Theory in a nutshell

Two-player game

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

$S_1 = \{s_{1,1}, s_{1,2} \ldots s_{1,n1}\}$ Set of strategies available to first player

$S_2 = \{s_{2,1}, s_{2,2} \ldots s_{n2}\}$ Set of strategies available to second player

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

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

Competitive (zero-sum) game

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

Sequential vs strategic vs multiple moves games

Equilibrium

Optimal choices

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

(stricly) Dominant strategy

The best strategy regardless of the other player's move

$$u_1(s_1^*, s_2) > u_1(s_1, s_2) \quad \forall s_1 \in S_1 \quad \forall s_2 \in S_2$$

... then equilibrium is

(s_1^*, s_2^*) with s_2^* such that

$$u_2(s_1^*, s_2^*) \geq u_2(s_1^*, s_2) \quad \forall s_2 \in S_2$$

Equilibrium

Nash equilibrium

No player gets an advantage by changing his strategy assuming the other does not change his own

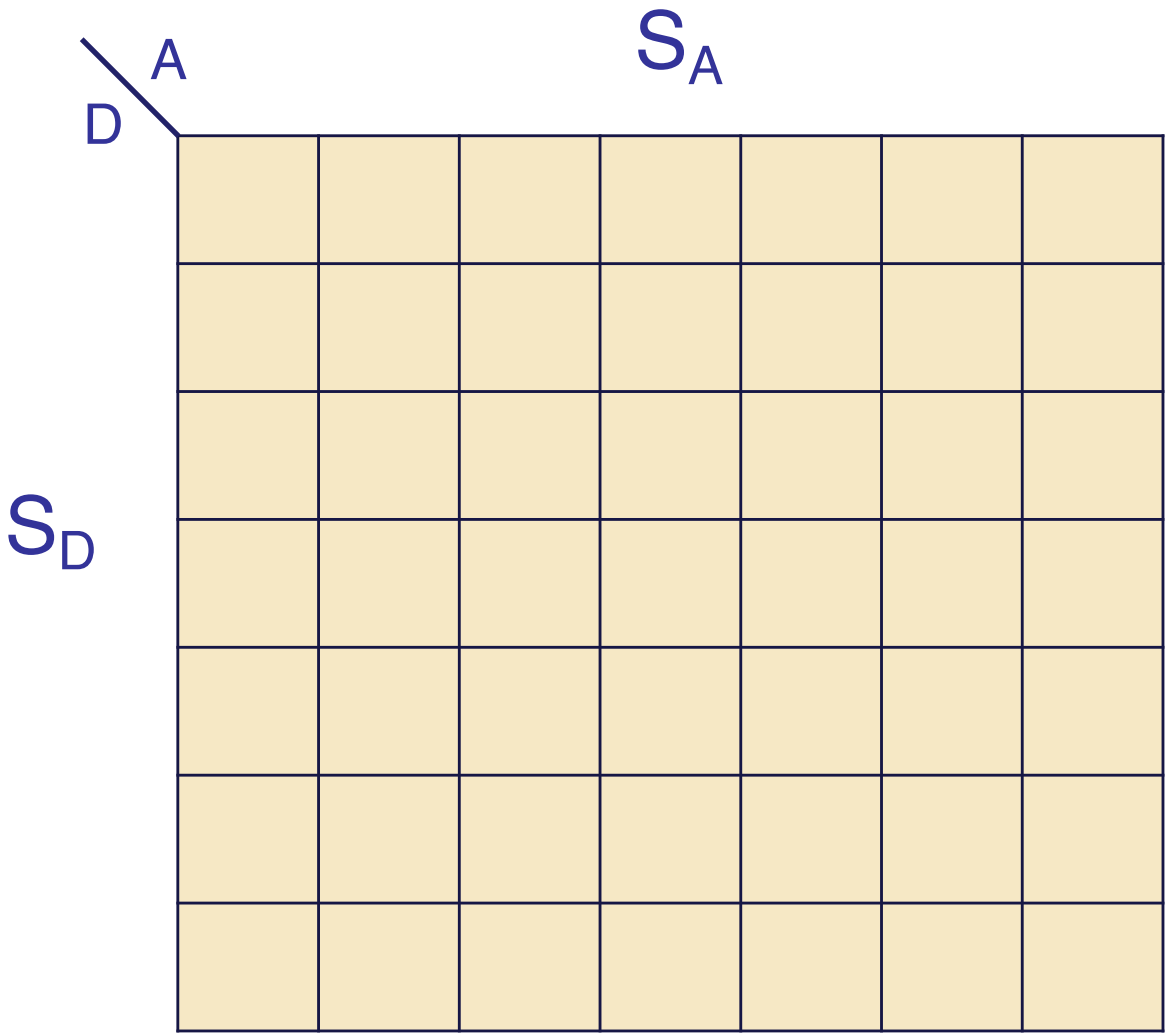
$$u_1(s_1^*, s_2^*) \geq u_1(s_1, s_2^*) \quad \forall s_1 \in S_1$$

$$u_2(s_1^*, s_2^*) \geq u_2(s_1^*, s_2) \quad \forall s_2 \in S_2$$

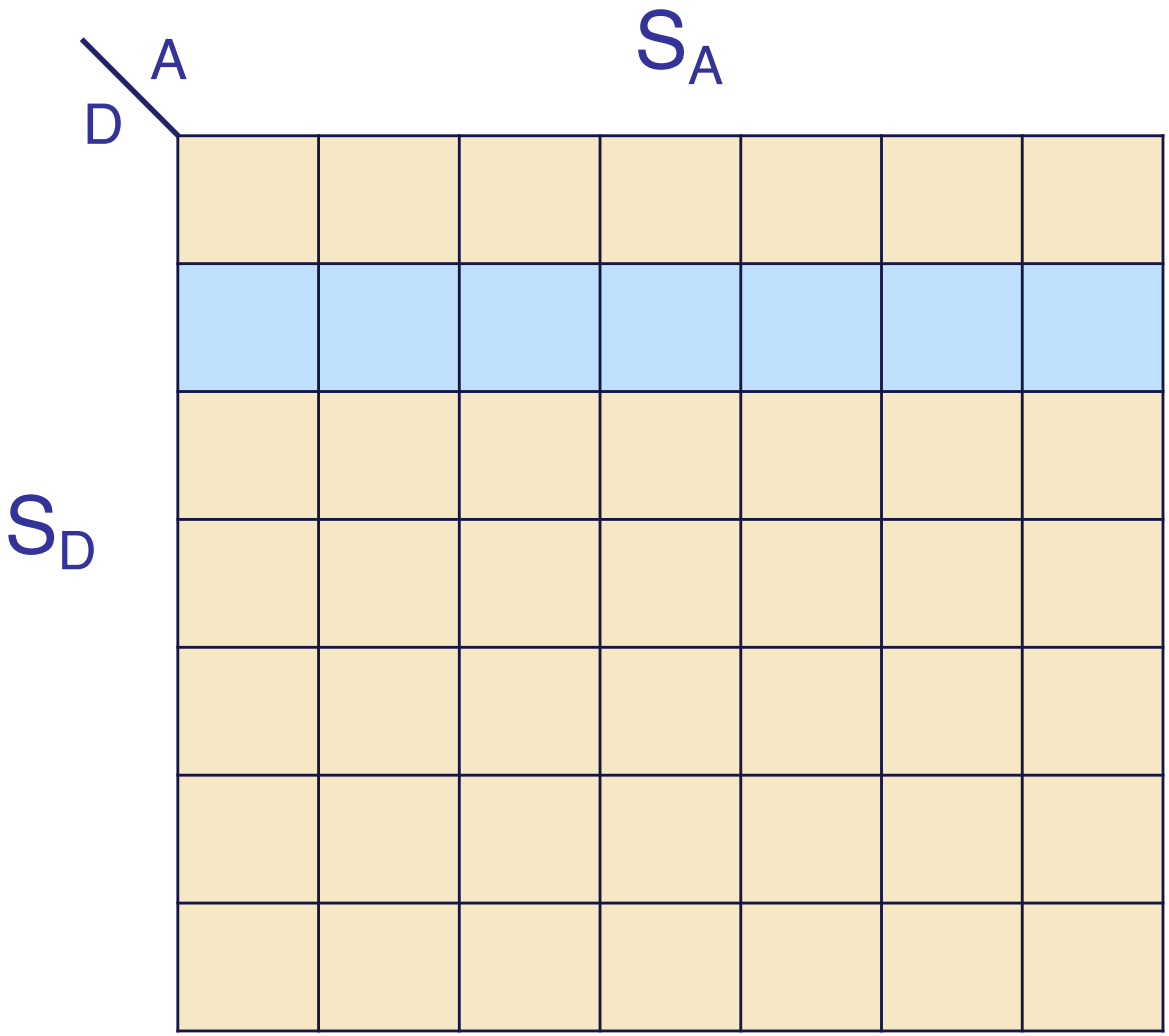
... and many others

- worst case assumption
- rationalizable equilibrium
- ...

The cat & mouse loop and Adv-SP

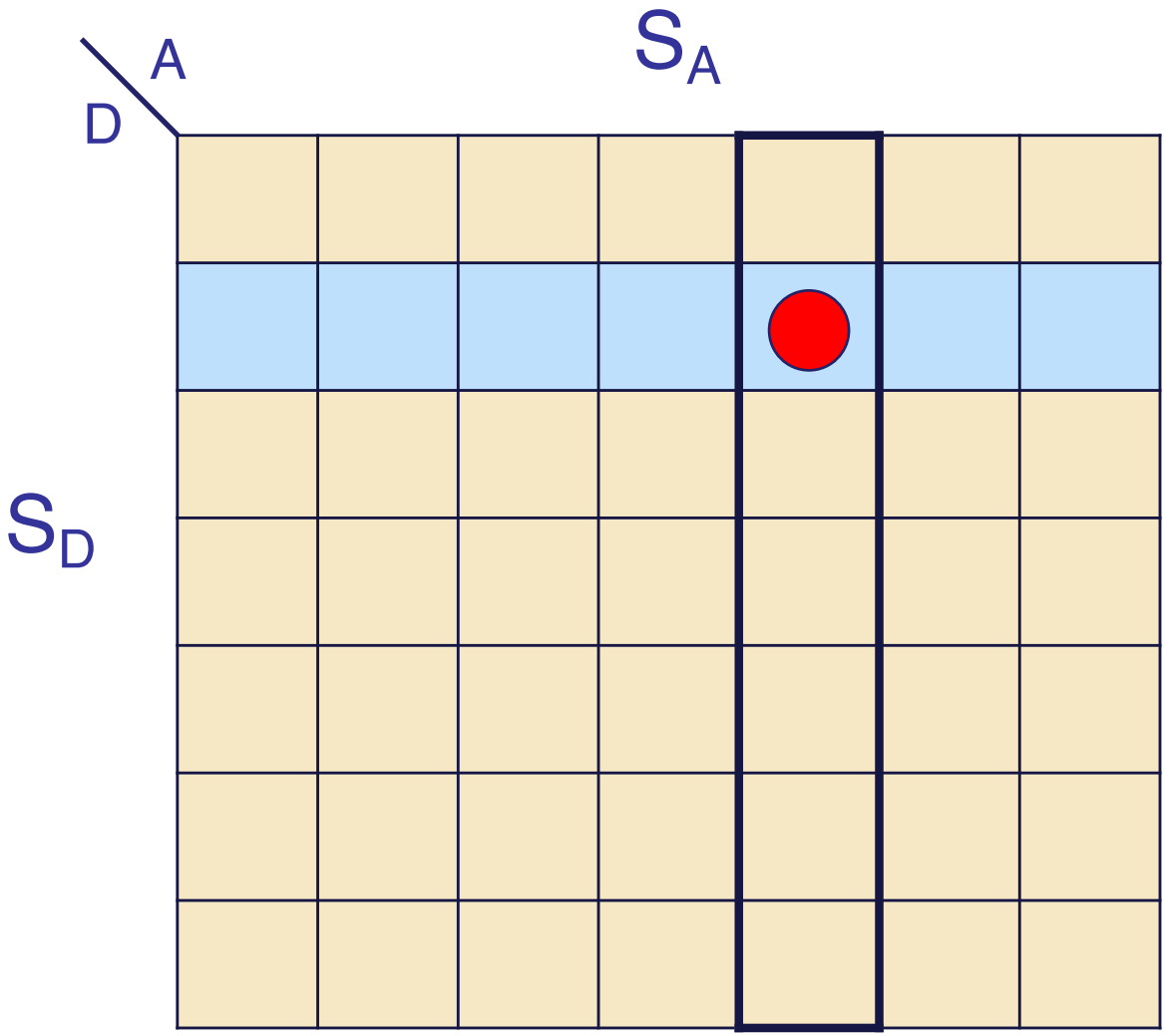


The cat & mouse loop and Adv-SP



Paper 1: The defender chooses a strategy according to a certain optimality criteria

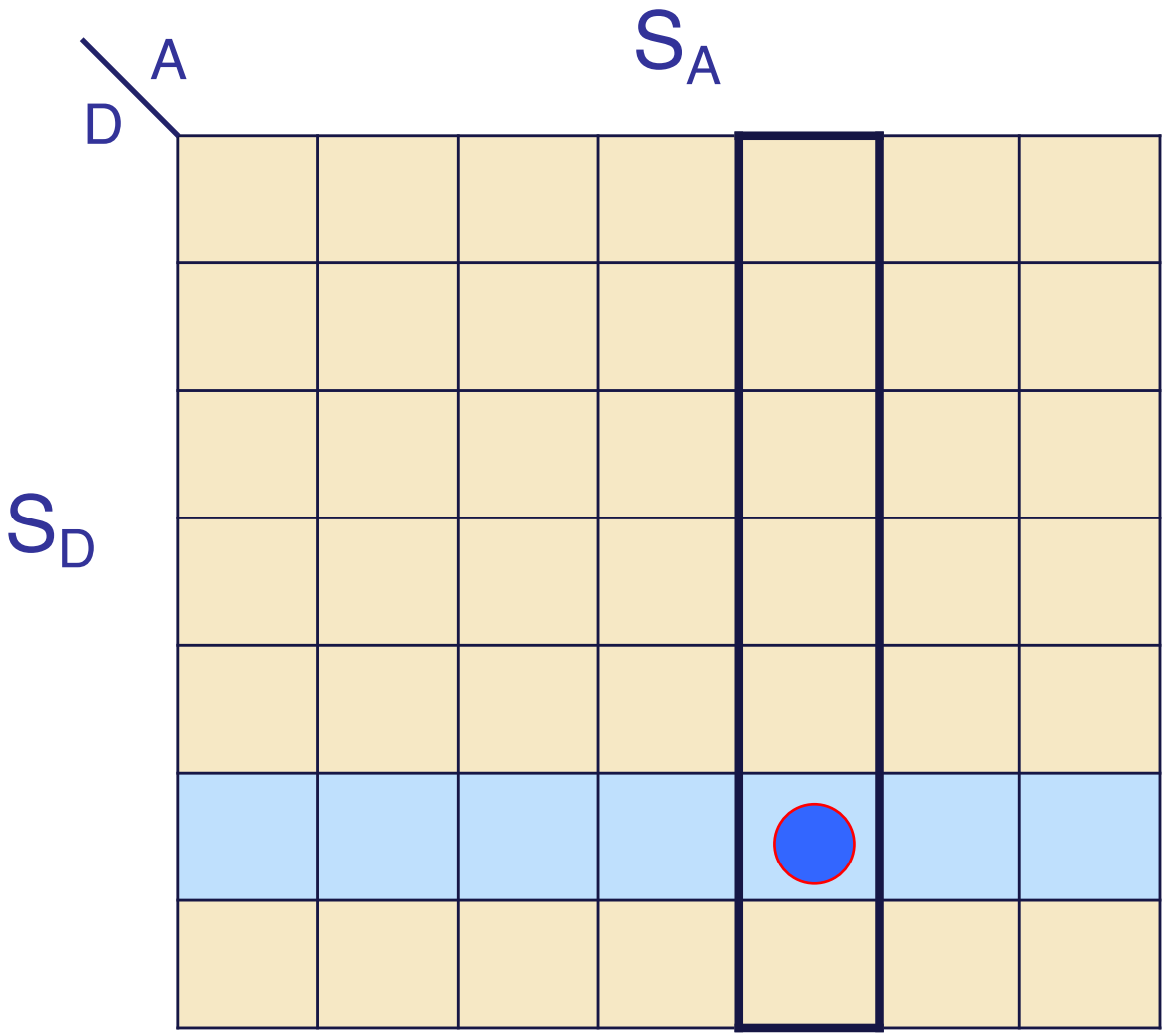
The cat & mouse loop and Adv-SP



Paper 1: D chooses a strategy

Paper 2: A derives the optimum attack

The cat & mouse loop and Adv-SP

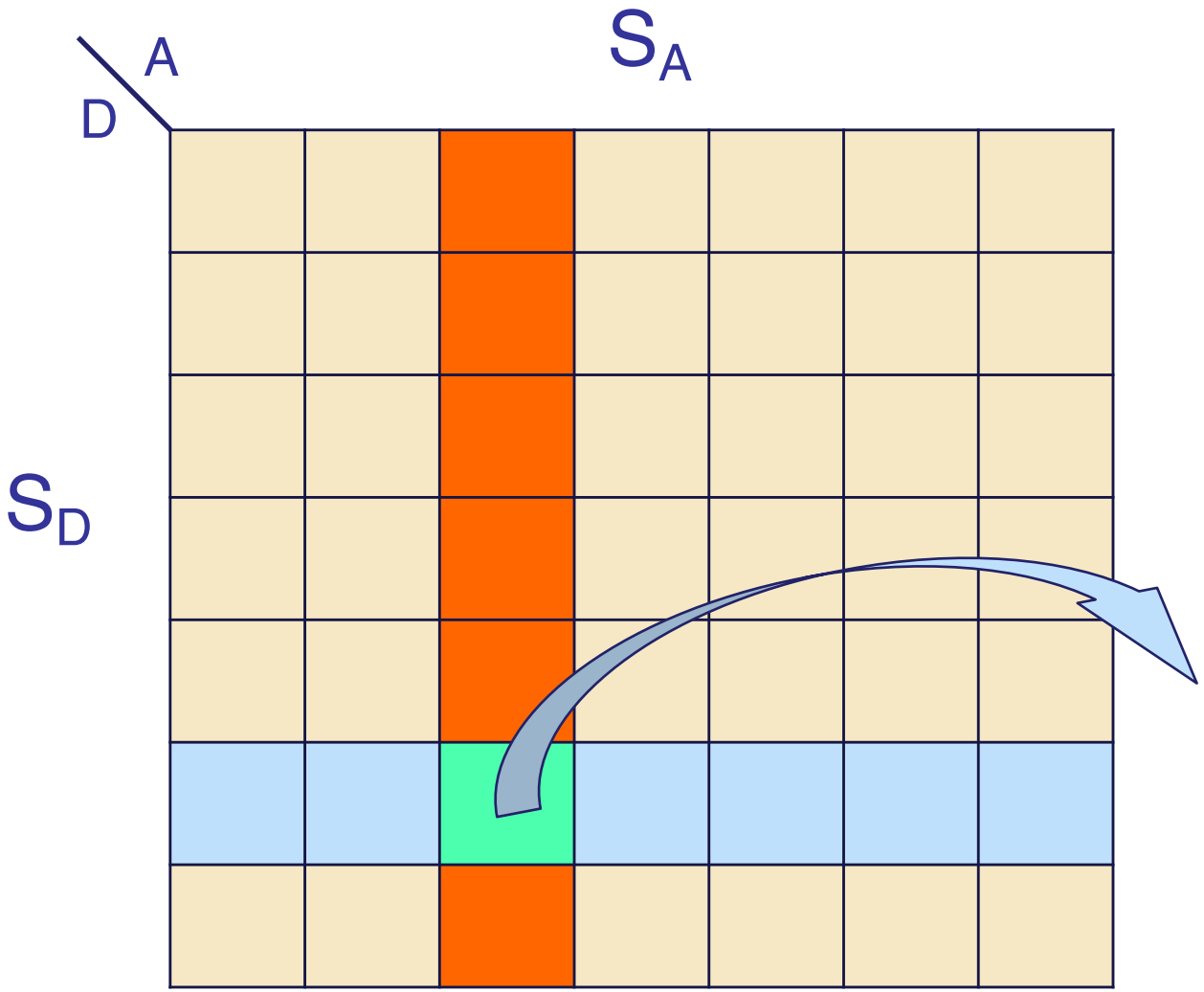


Paper 1: D chooses a strategy

Paper 2: A derives the optimum attack

Paper 3: D derives the optimum countermeasure (forgetting the initial optimality criteria)

The cat & mouse loop and Adv-SP



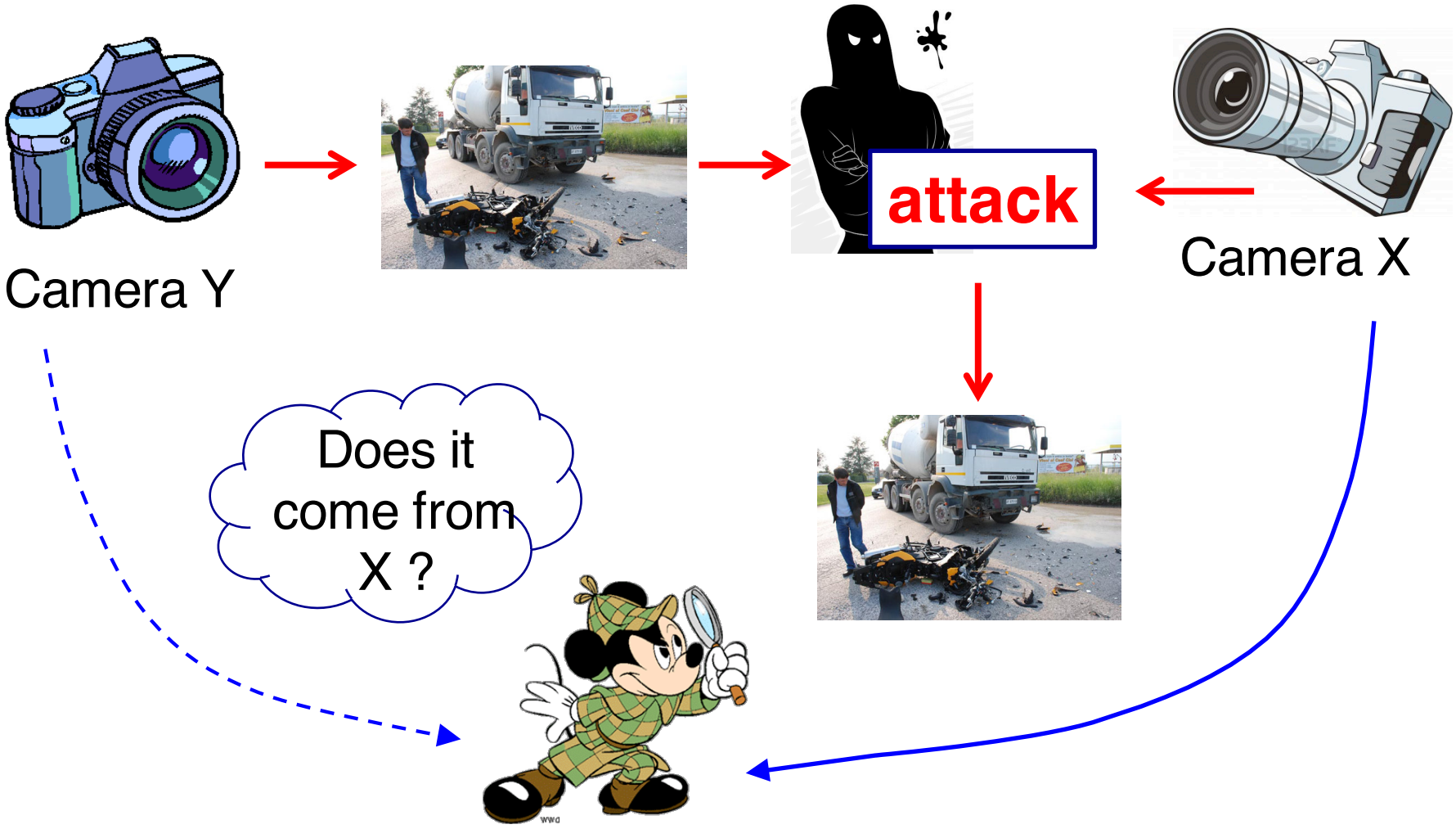
The equilibrium of the game represents the *optimum* choice for both players

It determines the security of the system

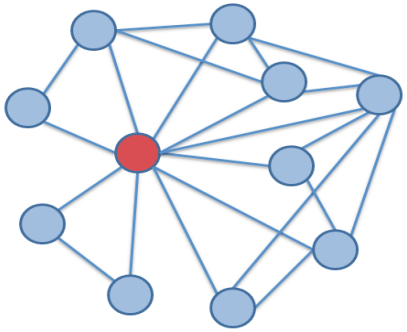
AdvSP at work

The Source Identification Game

A motivating example (1)

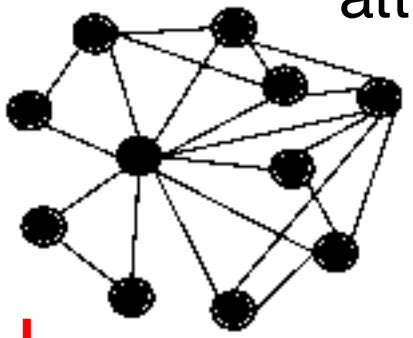


A motivating example (2)



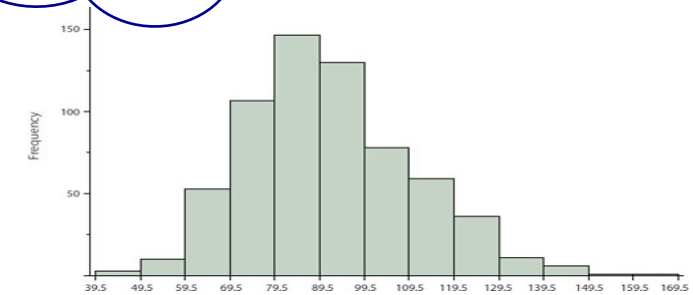
Normal
Network Traffic

Network under attack



Re-shape traffic so that the attack can not be detected

Is an attack on-going?



The SI Game with known sources*

First step: structure of the game

- Two DM sources X and Y with known pmf's P_X and P_Y
- Task of Defender (D): decide whether a sequence has been drawn from X
- Task of Attacker (A): modify a sequence drawn from Y so that it looks as if it were drawn from X subject to a distortion constraint

* M. Barni, B, Tondi, "The Source Identification Game: an Information-Theoretic Perspective", *IEEE Trans. Inform. Forensics and Security*, March 2013

Second step: explore S_D and S_A

For the defender

- All possible acceptance regions ...
- ... subject to a constraint of false I-type error probability
- ... including possible limitations on the kind of analysis the defender can carry out
- Asymptotic analysis

For the attacker

- All modifications subject to a constraint on the maximum distortion introduced by the attack

Second step: explore S_D and S_A

$$S_D = \left\{ \Lambda_0 : P_{fp} \leq 2^{-\lambda n} \right\}$$

Λ_0 is defined by relying on first order statistics only

$$S_A = \left\{ f(y^n) : d(y^n, f(y^n)) \leq nD \right\}$$

D = maximum average per letter distortion

Third step: define the payoff

Neyman-Pearson set up

- Zero sum game
- Payoff linked to II-type error probability

$$u_A(\Lambda_0, f) = -u_D(\Lambda_0, f) = P_{fn}$$

$$P_{fn} = \sum_{\mathbf{x}: f(\mathbf{x}) \in \Lambda_0} P_Y(\mathbf{x})$$

Fourth step: study the equilibrium

Lemma: dominant strategy for D

$$\Lambda_0^* = \left\{ x^n : D(\hat{P}_{x^n} \,\|\, P_X) < \lambda - |\chi| \frac{\log(n+1)}{n} \right\} \quad regardless\ of\ P_Y$$

is a dominant strategy for the defender.

Remark

The optimum strategy of D depends neither on P_Y nor on A's strategy (semi-universal and dominant strategy).

Fourth step: study the equilibrium

Having fixed the strategy of the Defender, the optimum strategy of the attacker is easy to derive

Theorem: dominance-based equilibrium

$$\Lambda_0^* = \left\{ x^n : D(\hat{P}_{x^n} \,\|\, P_X) < \lambda - |\chi| \frac{\log(n+1)}{n} \right\}$$

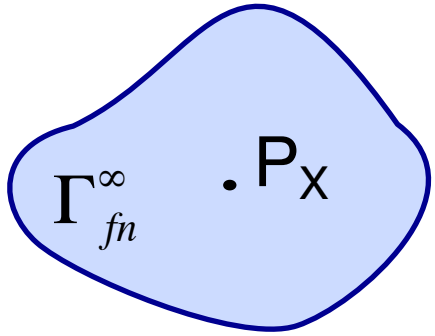
$$f^*(y^n) = \underset{z^n : d(z^n, y^n) \le nD}{\operatorname{argmin}} \; D(\hat{P}_{z^n} \,\|\, P_X)$$

Fifth step: who wins ?

Theorem 2: distinguishability region

Given P_X λ and D, we can define a region Γ_{fn}^{∞} such that

$$\begin{cases} \text{if } P_Y \in \Gamma_{fn}^{\infty} \text{ then } P_{fn} \to 1 \\ \text{if } P_Y \notin \Gamma_{fn}^{\infty} \text{ then } P_{fn} \to 0 \end{cases}$$



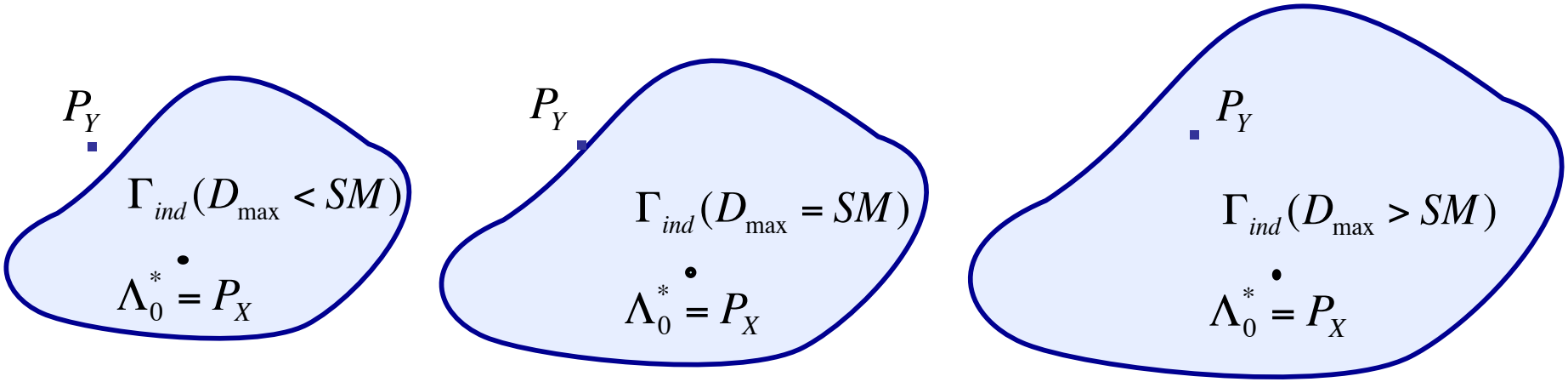
By letting λ -> 0 we obtain the ultimate distinguishability region for a certain distortion level D.

Fifth step: who wins ?

Security margin*

Let D_{max} = maximum value of D for which P_X and P_Y are distinguishable, we can say that P_X and P_Y are distinguishable up to an attack of power D_{max}

SM = D_{max} is said the *security margin between P_X and P_Y*



* M. Barni, B. Tondi, "Source Distinguishability under Distortion-Limited Attack: an Optimal Transport Perspective", *IEEE Trans. Information Forensics and Security*, vol. 11, no. 10, Oct. 2016,

SM and optimal transport theory

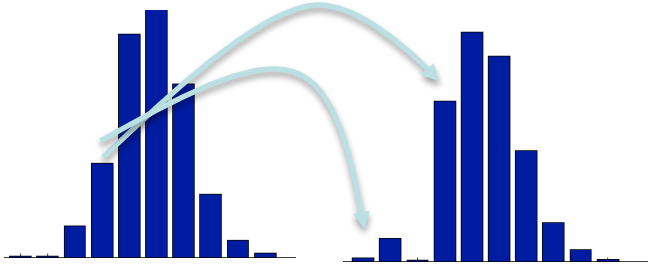
We can compute SM by resorting to optimal transport theory

Let us interpret P_Y and P_X as two different ways of piling up a certain amount of soil

Let c(i,j) be the cost of moving a unitary amount of earth from the i-th to the j-th bin

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

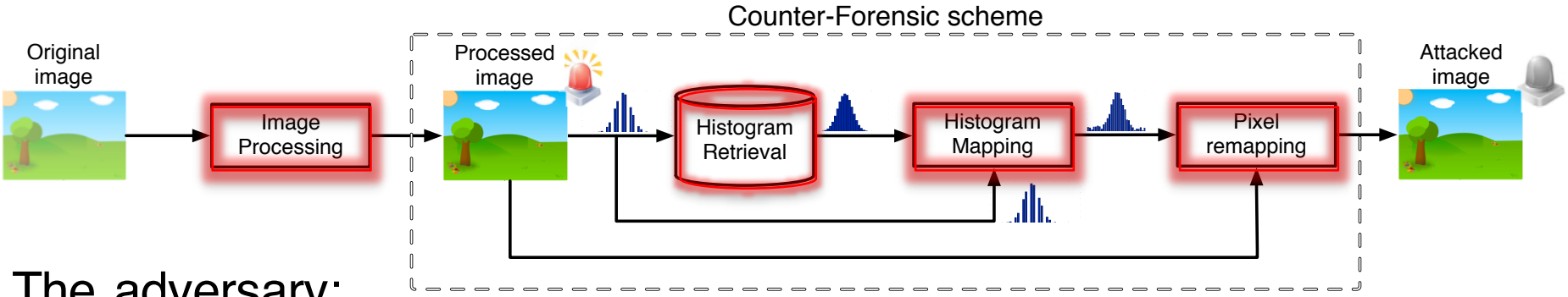
We have: **$SM(P_Y,P_X) = EMD(P_Y,P_X)$**
which can be computed numerically



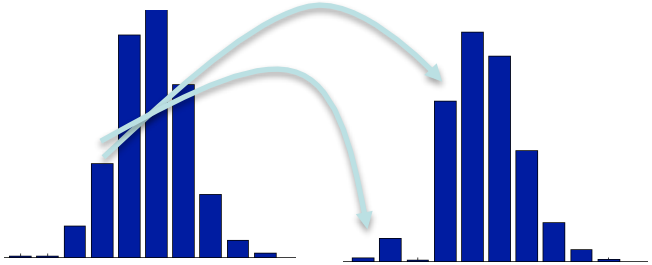
From theory to practice

- Histogram-based detection of contrast enhancement or gamma correction
- Thanks to theory
  - We avoid cat and mouse game
  - Universal attack: the attack is optimum against any detector based on first order statistics

From theory to practice: an example in histogram-based image forensics*



- The adversary:
  - Processes the image
  - Searches the DB for the nearest untouched histogram
  - Computes a transformation map from one histogram to the another
  - Applies the transformation, minimizing perceptual distortion



- M. Barni, M. Fontani, B. Tondi, "A Universal Technique to Hide Traces of Histogram-Based Image Manipulations", Proceedings of MMSEC 2012, Coventry (UK), Sept. 2012.

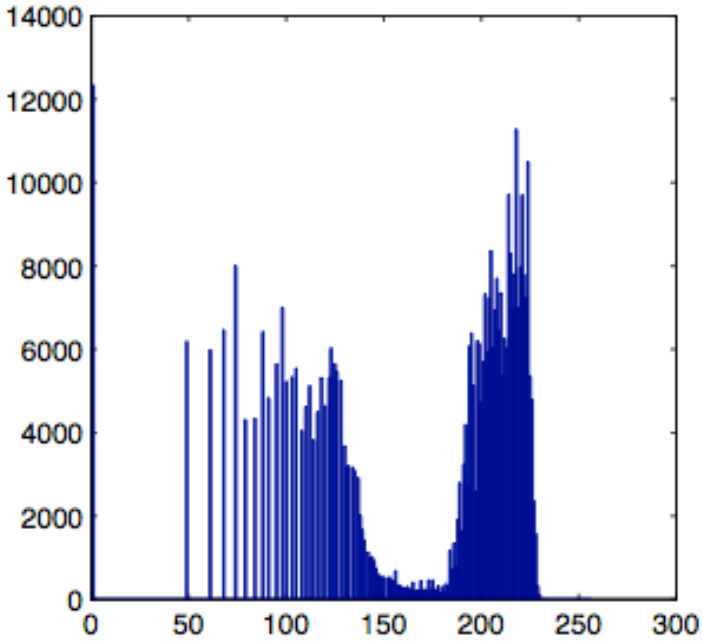
Example

Original Image

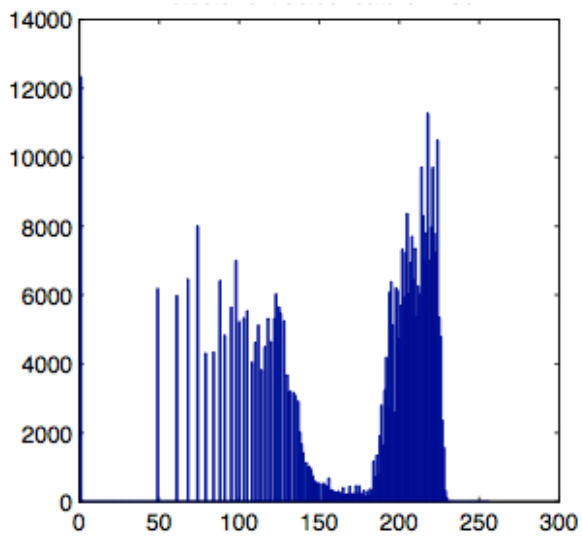
Example

Processed image (gamma-correction)

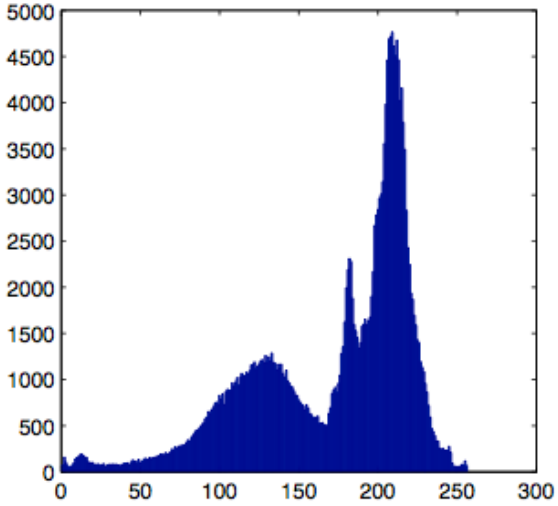
Resulting histogram



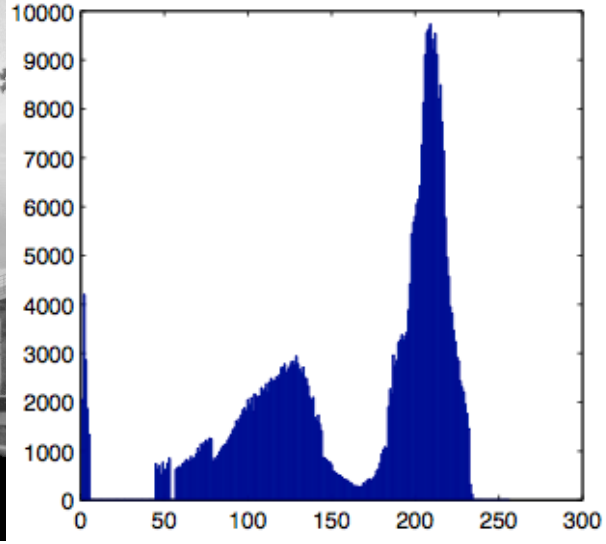
Prior to Counter-Forensics



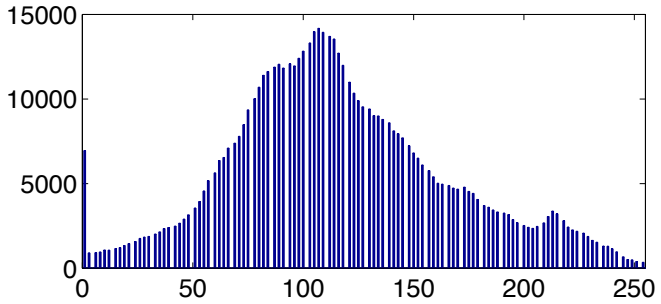
DB histogram



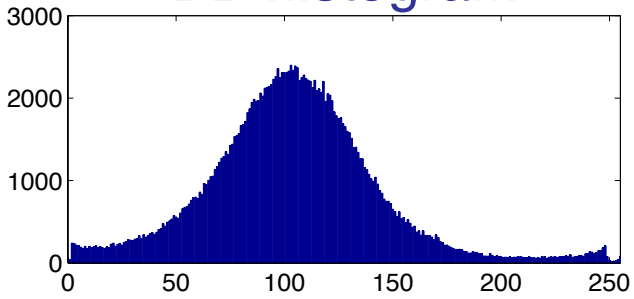
After Counter-Forensics

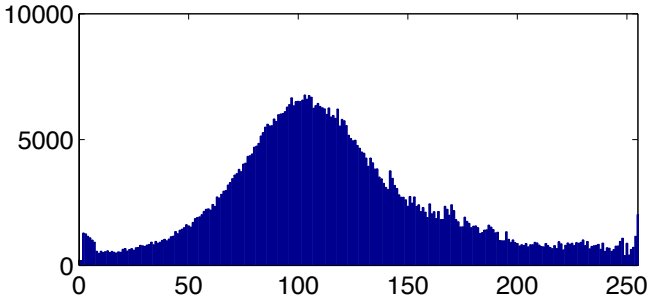


Another example



DB histogram

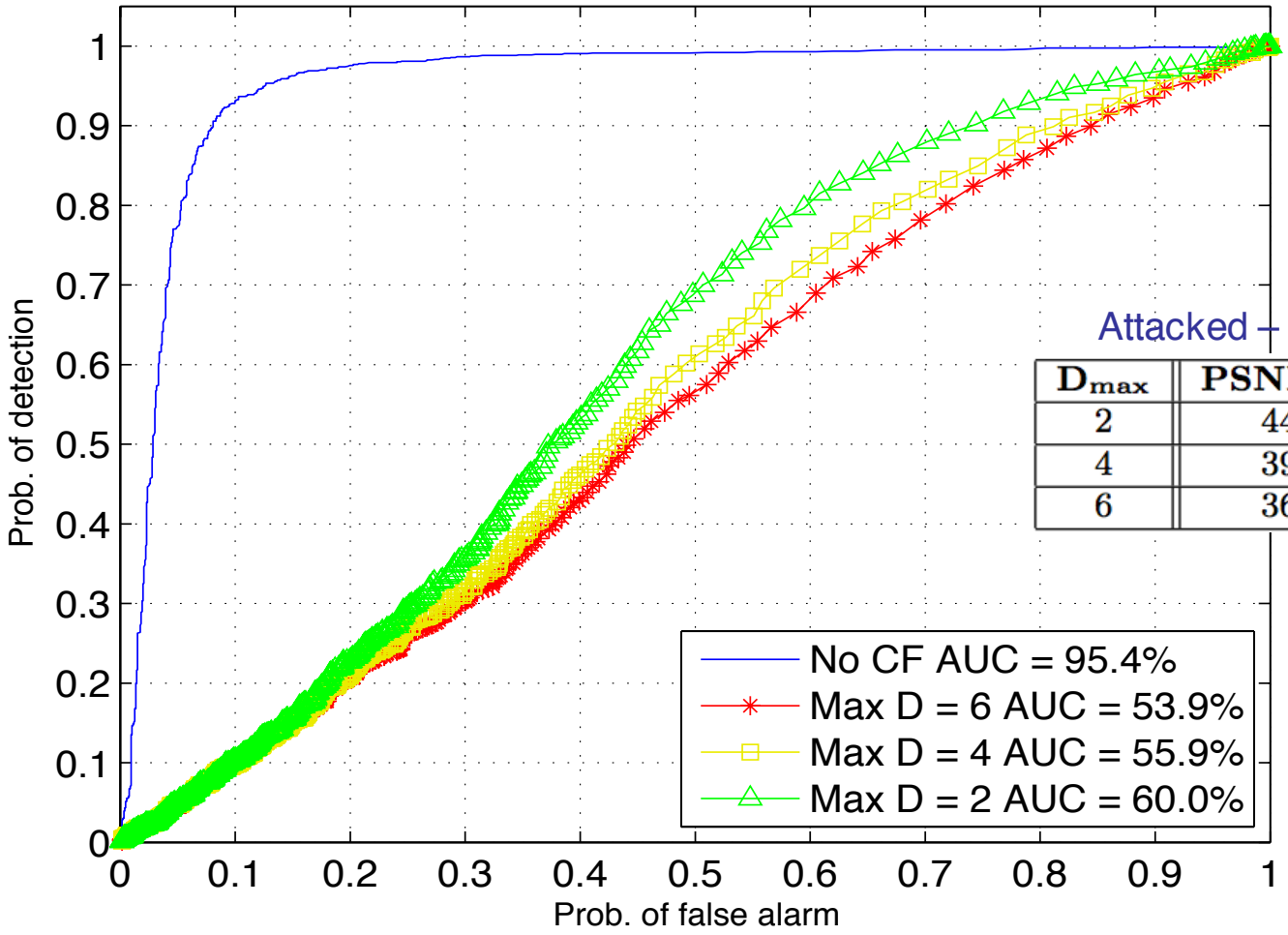




$D_{max} = 4$

Experimental results: gamma correction

ROC curves for Contrast Enhancement (γ–correction) detection before (solid line) and after (marked lines) CF attack.



Attacked – Processed distance

D_{max}	PSNR(db)	SSIM	AUC
2	44.8	0.994	0.600
4	39.3	0.981	0.559
6	36.2	0.964	0.539

A practical meaning of the SM

Database of images belonging to a certain class C_0

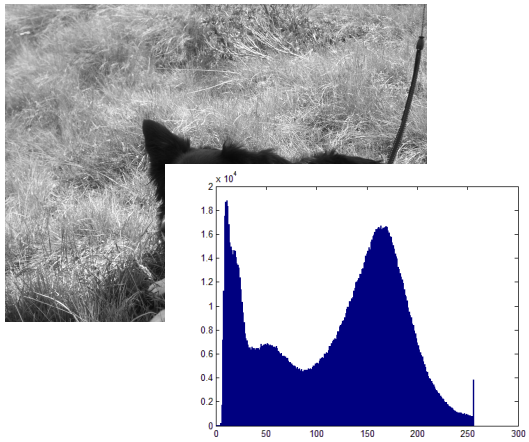
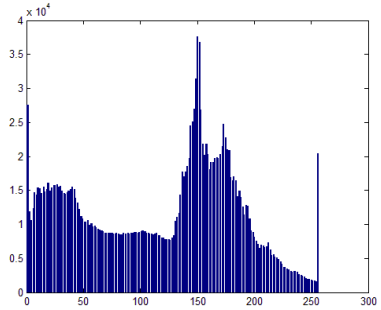


Image Y that does not belong to C_0



The minimum SM between the histogram of Y and those of the images in the database gives the minimum effort required to the attacker to make Y indistinguishable from the images in C_0

Conclusions

Extensions

- Source identification with training date
- Source identification with multiple observations
- Source identification with corrupted training
- Fully active adversary

Future research: there's a lot to work on

- Non-asymptotic analysis
- Go beyond binary HT
- Machine learning
- Coalition games
- Computational aspects

References

1. M. Barni, F. Perez-Gonzalez, “Coping with the enemy: advances in adversary-aware signal processing”, *Proc. of ICASSP 2013, IEEE International Conference on Acoustic Speech and Signal Processing*, 26-31 May 2013, Vancouver - Canada.
2. 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
3. M. Barni, B. Tondi, “Binary Hypothesis Testing Game with Training data”, *IEEE Transactions on Information Theory,* vol. 60. no. 8, August 2014, pp. 4848-4866
4. M. Barni, B. Tondi, “Source Distinguishability under Distortion-Limited Attack: an Optimal Transport Perspective”, *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 10, pp. 2145-2159, Oct. 2016
5. M. Barni, B. Tondi, “Adversarial Source Identification Game with Corrupted Training”, submitted to *IEEE Trans. on Information Theory* (available on arXiv)

Thank you
for your attention