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? Or a battlefield?

Search engines

Reputation scores

User-generated contents

Biometric identification

Fake images

Identity theft

Network intrusion

SPAM

Denial of Service
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
• Same solutions are re-invented again and again

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

• We do not understand the real essence of problems
• 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 noisiness
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
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

\[ G(S_1, S_2, u_1, u_2) \]

\[ S_1 = \{ s_{1,1}, s_{1,2} \ldots s_{1,n_1} \} \quad \text{Set of strategies available to first player} \]

\[ S_2 = \{ s_{2,1}, s_{2,2} \ldots s_{n_2} \} \quad \text{Set of strategies available to second player} \]

\[ u_1(s_{1,i}, s_{2,j}) \quad \text{Payoff of first player for a given profile} \]

\[ u_2(s_{1,i}, s_{2,j}) \quad \text{Payoff of second player for a given profile} \]

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

(strictly) 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

A

D

S_A

S_D
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

<table>
<thead>
<tr>
<th>A</th>
<th>S_A</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>S_D</td>
</tr>
</tbody>
</table>

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)

Does it come from X?
A motivating example (2)

Normal Network Traffic

Network under attack

Is an attack on-going?

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

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 $S_A$

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

$\Lambda_0$ 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

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

\[ P_{fn} = \sum_{x: f(x) \in \Lambda_0} P_Y(x) \]
Fourth step: study the equilibrium

Lemma: dominant strategy for D

$$\Lambda^*_0 = \left\{ x^n : D(\hat{P}_x^n \parallel P_X) < \lambda - \left| \chi \frac{\log(n+1)}{n} \right| \right\} \text{regardless of } P_Y$$

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

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

\[ f^*(y^n) = \arg\min_{z^n : d(z^n, y^n) \leq nD} D(\hat{P}_{z^n} \parallel P_X) \]
Fifth step: who wins?

Theorem 2: distinguishability region

Given $P_X \lambda$ and D, we can define a region $\Gamma_{fn}^\infty$ 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 $\lambda \to 0$ we obtain the ultimate distinguishability region for a certain distortion level D.
Fifth step: who wins?

Security margin*

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

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

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

Example

Original Image
Example

Processed image (gamma-correction)

Resulting histogram
Prior to Counter-Forensics

After Counter-Forensics

DB histogram
Another example

DB histogram

$D_{\text{max}} = 4$
Experimental results: gamma correction

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

<table>
<thead>
<tr>
<th>$D_{\text{max}}$</th>
<th>PSNR (db)</th>
<th>SSIM</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>44.8</td>
<td>0.994</td>
<td>0.600</td>
</tr>
<tr>
<td>4</td>
<td>39.3</td>
<td>0.981</td>
<td>0.559</td>
</tr>
<tr>
<td>6</td>
<td>36.2</td>
<td>0.964</td>
<td>0.539</td>
</tr>
</tbody>
</table>

No CF AUC = 95.4%
Max D = 6 AUC = 53.9%
Max D = 4 AUC = 55.9%
Max D = 2 AUC = 60.0%
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


Thank you for your attention