Département Informatique
Année Universitaire 2022-2023

Chapitre 4

SECURITY ATTACKS

Outline

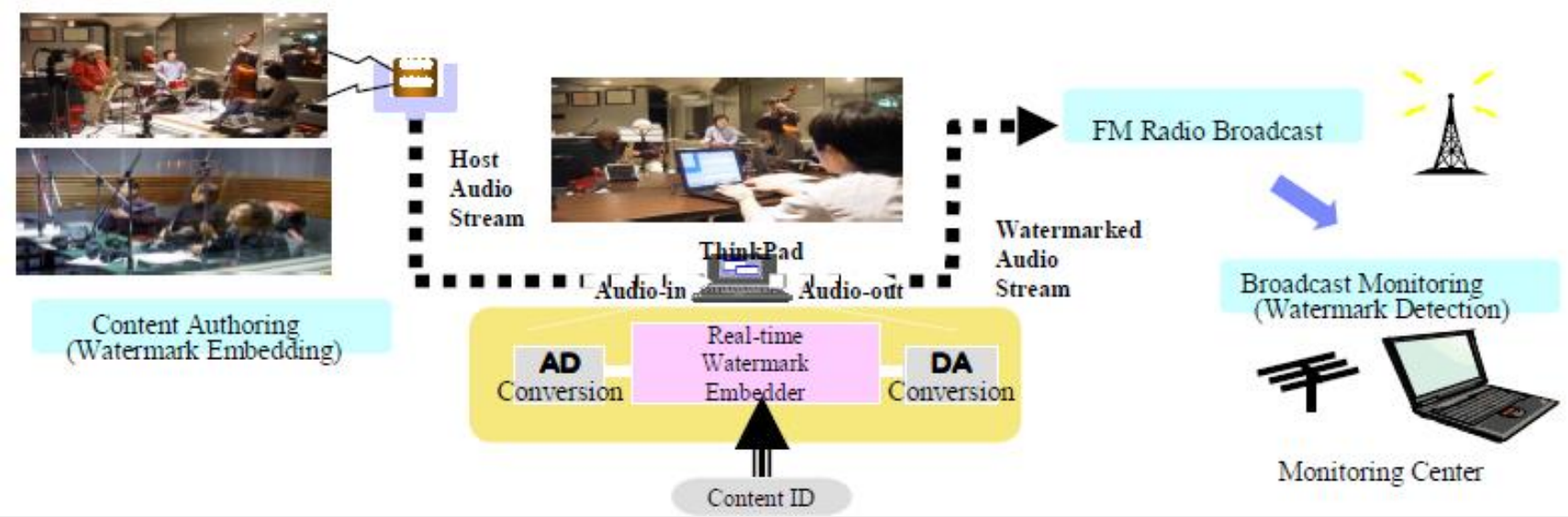
- Major Issues in Watermarking
- Security Requirements
- Categories of Attack
- Some Significant Known Attacks

Major Issues in Watermarking

- Invisibility:
  - Least-Significant Bits
    - Spatial Domain
    - Compression-Compliant Block-Frequency Domain
    - Global Frequency Domain
  - Human Perceptual Models
    - Domain-Specific Models
    - Generic Models
- Robustness:
  - Lossy Compression
  - Format Transformation
  - Scaling, Translation, Cropping
  - Rotation, Scan-and-Print
- Embedding Information Payload:
  - Information Theory
  - Writing on Dirty Paper
  - Zero-Error Embedding Capacity
- Security:
  - Attacks

Security Requirements Depend on Applications – Example (1)

- **Scenario 1:** Alice is an advertiser who embeds a watermark in each of her radio commercials before distribute them to 600 radio stations.
  - Alice monitors radio station broadcasts with a watermarking detector.
  - She matches her logs with the 600 invoices.
  - **[Attack]:**
    - Bob secretly embed Alice's watermark into his own advertisement and airs it in place of Alice's commercial.



Unauthorized Embedding / Forgery Attack

Security Requirements Depend on Applications – Example (2)

- **Scenario 2:** Alice owns a watermarking service that, for a nominal fee, adds an owner identification watermark to images that will be accessed through the Internet.
  - Alice provides an expensive reporting service to inform her customers of all instances of their watermarked images found on the Web.
  - **[Attack]:** Bob builds his own web crawler that detects watermarks embedded by Alice and offers a cheaper reporting service.

Unauthorized Detection / Passive Attack

Security Requirements Depend on Applications – Example (3)

- **Scenario 3:** Alice owns a movie studio, and she embeds a copycontrol watermark in her movies before they are distributed.
  - She trusts that digital recorders capable of copying these movies contain watermark detectors and will refuse to copy her movie.
  - **[Attack]** Bob is a video pirate who has a device designed to remove the copy protection watermark.

Unauthorized Removal

Operational Table of the Three Scenarios

	Embed	Detect	Remove
Scenario1: Broadcast Monitoring			
Advertiser	Y	Y	-
Broadcaster	N	N	-
Public	N	N	-
Scenario2: Web Reporting			
Marking Service	Y	Y	-
Reporting Service	-	Y	-
Public	N	N	N
Scenario3: Copy Control			
Content Provider	Y	Y	-
Public	-	Y	N

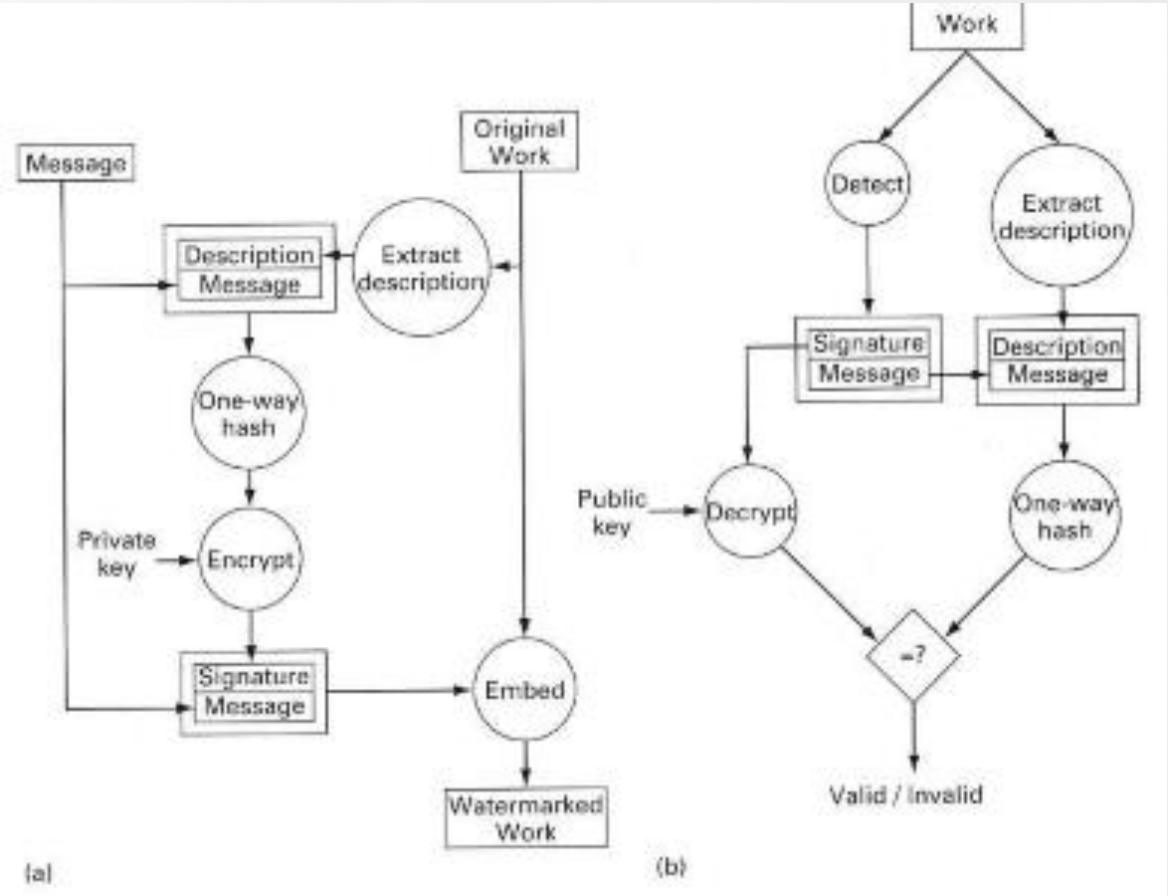
Y: must be allowed, N: must not be allowed, - : don't care

Categories of Attack (1)

- **Unauthorized Embedding:**
  - Being able to composing and embedding an original message..
    - Another example, in Scenario 2, Alice charges for embedding and gives away the monitoring tool..
    - Possible Solution: using standard cryptographic techniques.
  - Being able to obtain a pre-composed legitimate message and embeds this message in a Work.
    - E.g., in Scenario 1, Bob extracts the reference pattern and then uses it to his work – called copy attack.
    - Possible Solution: using content-related watermarks.

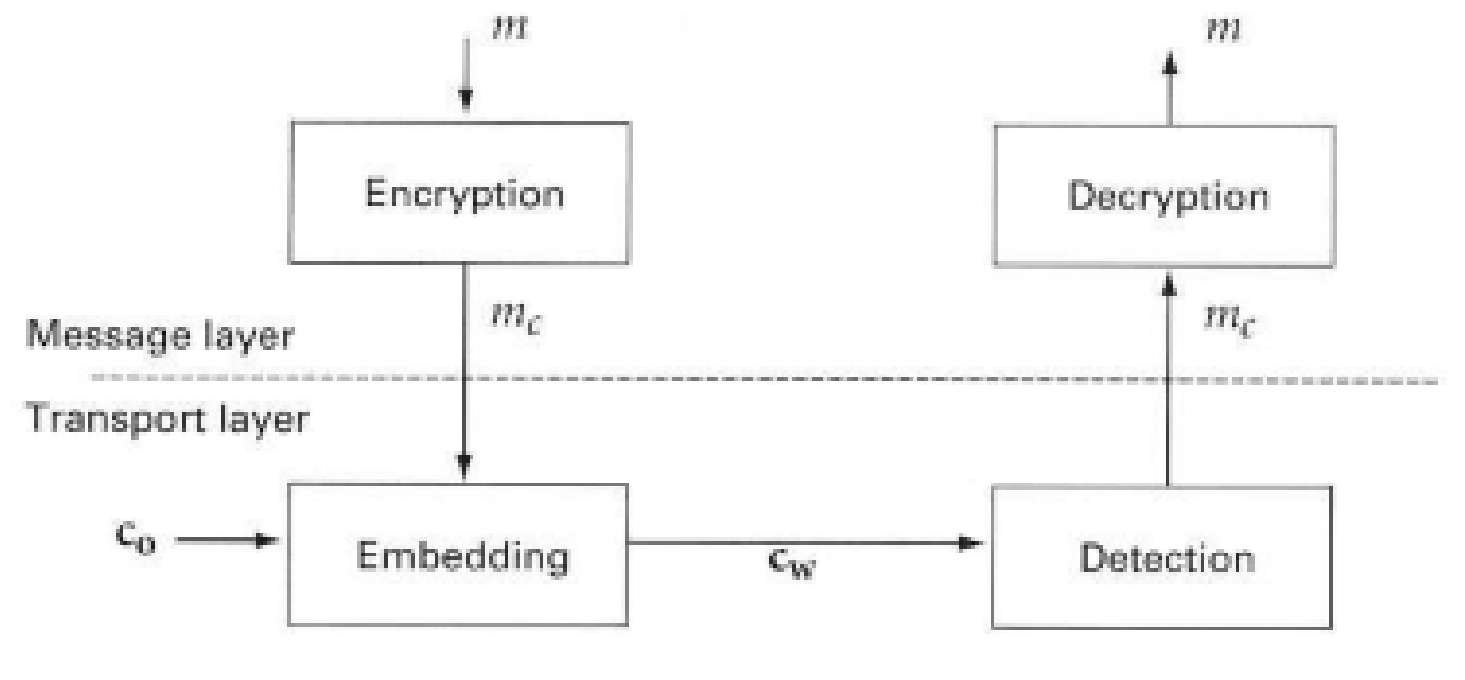
Methods to Prevent Unauthorized Embedding

- Make the embedding codes:
  - Content dependent
  - Signer dependent



Categories of Attack (2)

- **Unauthorized Detection:**
  - A hospital might embed the names of patients into their X-rays.
  - Knowing whether or not a watermark is present Steganography.
  - Intervention on the transmission process.



Categories of Attack (3)

- **Unauthorized Removal:**
  - Attackers try to modify the watermarked Work such that it resembles the original and yet does not trigger the detector.
  - Two types of attacks:
    - Elimination attacks ->The watermark is truly gone.
    - Masking attacks -> The watermark is still present but is weakened.

Methods to Prevent Unauthorized Removal

- Spread Spectrum Techniques are suggested.
- One-line of researching is based on the belief that watermarking can be made secure by creating something analogous to asymmetric-key encryption -> The detection key is not sufficient to remove a watermark -> May not survive sensitivity analysis.
- There are some fundamental differences between watermarking and cryptography that make the standard asymmetric-key encryption systems unsuitable.
  - In watermarking, the mapping between Works and messages must be many-to-one, so that a given message may be
  - embedded in any given Work.
  - In asymmetric-key cryptography, the mapping between cleartext and ciphertext is always one-to-one.
  - In watermarking, small changes in the Works should map to similar messages.
  - In assymmetric-key cryptography, a small change in cleartext results in large change in the ciphertext.

Categories of Attack (4)

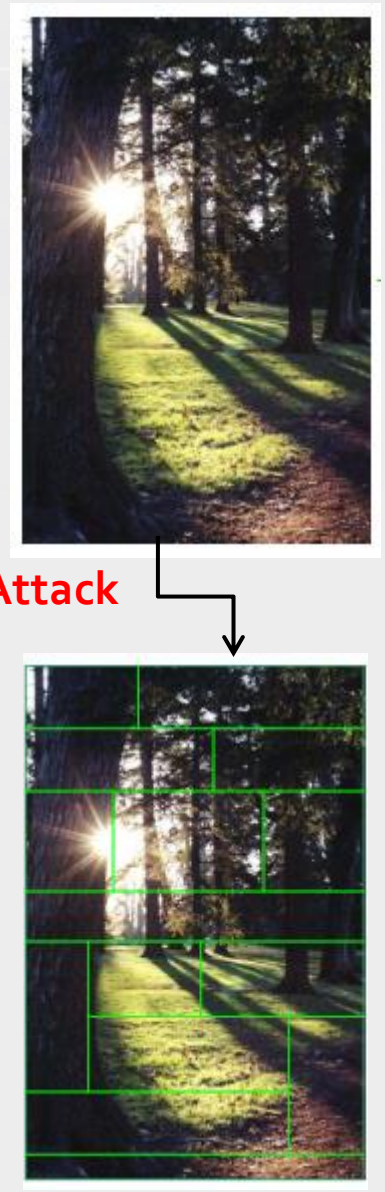
- **System-level Attacks:**
  - Attackers exploit the weakness in how the watermarks are used.
  - For instance, in a copy-control application, an attaker might open the recorder and just remove the chip.
  - Forge identification.

Some Significant Known Attacks

- Scrambling attacks
- Pathological distortions:
  - Synchronization attacks
  - Linear filtering and Noise Removal Attacks
- Copy attacks
- Ambiguity attacks
  - Ambiguity attacks with informed detection
  - Ambiguity attacks with blind detection
- Sensitivity analysis attacks
- Gradient descent attacks

Scrambling Attack

- System-level attack
  - An additional device is applied to scramble watermarked multimedia work to make the watermark undetectable by the detector.
  - Using a descramble device to invert the work.
- Example:
  - Mosaic Attack: partition the watermarked image into several individual smaller images that are organized with table when displayed.
- Effectiveness: avoid on-line image crawling

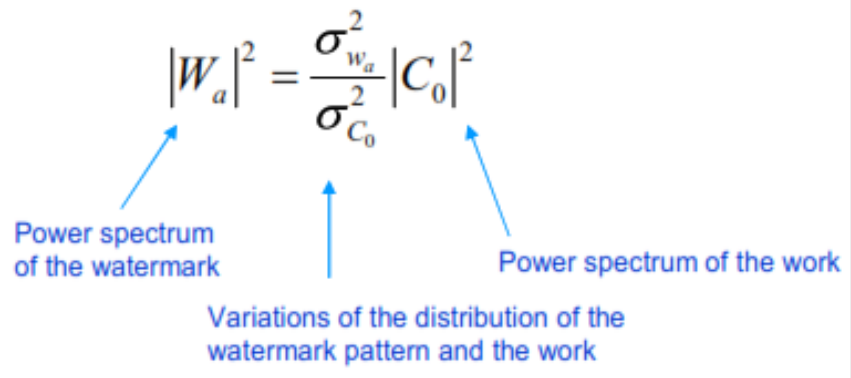


Pathological Distortions (I)

- **Synchronization Attacks:**
  - Most watermarking techniques are sensitive to synchronization
  - Audio and Video: delay and time scaling
    - Pitch-preserving scaling
    - Sample removing
  - Image and Video: rotation, scaling and translation
    - Shearing
    - Horizontal reflection
    - Column or line removal
    - Nonlinear warping
- Some of these attacks are applied by the StirMark – a watermark benchmarking system.

Pathological Distortion (II)

- **Linear Filtering and Noise Removal Attacks:**
  - May be effective while many watermarking system embed significant energy in the high frequencies.
  - Wiener filtering is an optimal linear-filtering/noise-removal attack. It is effective when:
    - The added pattern is independent of the work.
    - Both the work and the watermark are drawn from zero-mean Gaussian distribution.
    - Linear correlation is used as the detection statistic.
  - The security of a watermark against Wiener filtering can be maximized by selecting the power spectrum of the added pattern to be a scaled version of the power spectrum of the original work, as:



Copy Attack

- An adversary copies a watermark from one work to another. It is a form of unauthorized embedding.
- Example: (Kutter et al., 2000) given a legitimately watermarked work, *c1w*, and an unwatermarked target work, *c2*, this method begins by
  - Applying a watermark removal attack to *c1w* to obtain an approximation of the original, *c1'*, by using a nonlinear noise-reduction filter.
  - Estimate the added watermark pattern by subtracting the estimated original from the watermarked work:

$$w_a' = c_{1w} - c_1'$$

  - The estimated watermark pattern is added to the unwatermarked work:

$$c_{2w} = c_2 + w_a'$$

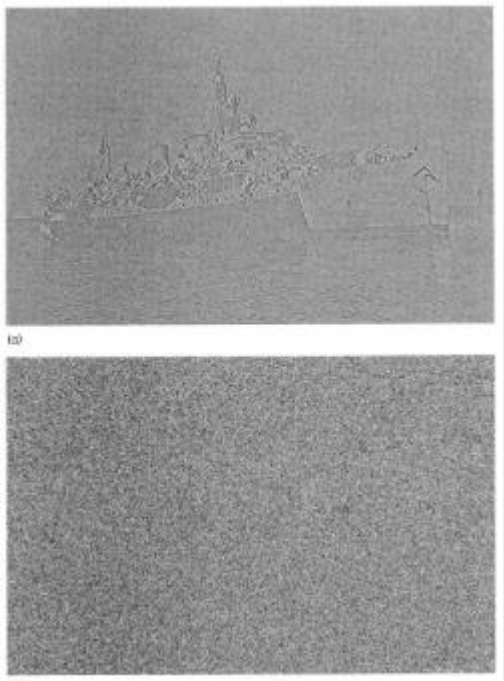
Ambiguity Attacks

- **Ambiguity attacks (or called the Cover attack, Craver et al., 1998):** create the appearance that a watermark has been embedded in a work when in fact no such embedding has taken place.
- **Objectives:** claiming false ownership.
- **Two situations:**
  - ambiguity attacks with informed detection
  - ambiguity attacks with blind detection

Ambiguity Attacks with Blind Detection

- Examples of Ambiguity Attack: (a) True original Image, (b) Distributed Watermarked Image.

Ambiguity Attacks with Blind Detection



Ambiguity Attack (a): Adding some random noise into the Fourier phase; (b) Add noise to the image and then scale Fourier coefficients with random magnitude changes

Faked original image constructed by subtracting 99.5% of the fake reference pattern

Defending Ambiguity Attacks

- The true owner of the Work uses a watermarking technique that can ensure that his original could not have been forged.
- Invertibility: a watermarking scheme is invertible if the inverse of the embedding is computationally feasible.
- Ambiguity attacks cannot be performed with non-invertible embedding techniques. For instance, the reference pattern should be dependent on the content of the original work.