Outcomes

- Be familiar with the different taxonomies used to classify steganalysis techniques.
- Have an overview of different techniques and approaches used.

Classification by question asked

Steganalysis:
- Does this file contain a hidden message?

Extended steganalysis:
- How long is the embedded message?
- What is the contents of the hidden message?
- Which stego-algorithm has been used?
- What is the secret key used for the embedding?

Other attacks:
- Disable the message.

The questions we don’t ask

- Contents
  - Encryption protects the contents
  - Trivial to combine steganography and encryption
  - Break the encryption to recover contents
- Disable the message
  - The problem of Robust Watermarking
**Questions we do ask**

- Steganalysis – the existence of the message
  - Wendy the Warden’s only question
- Key-Recovery Attacks
  - Breaks all steganograms made with one key
- Diagnostics – potentially useful extra information
  - Message length
  - Algorithm used
  - This information is not easily protected by encryption

**Neil Johnson’s classification**

- Stego-only attack (secret algorithm)
- Known cover attack
- Known message attack
- Chosen stego attack (known algorithm and stego object)
- Chosen message attack
- Known stego attack (known algorithm, original, and stego-image)

Source: Katzenbeisser and Petitcolas (Ch. 4 by Neil Johnson)

**Cryptoanalysis**

- Neil Johnsen followed classification of steganalysis.
- Cryptanalysis classes
  - Ciphertext only
  - Known plaintext
  - Chosen plaintext
  - Chosen ciphertext
  - Aims to recover, either
    - Key, or
    - Message
  - The algorithm is always known

**Clearing up the mess**

- Steganalysis has more variables than cryptanalysis
  - Stego-algorithm may be secret
  - Cover image
  - Therefore the classification becomes more complex
  - Neil Johnson’s classification is ambiguous
    - Does the known cover/message correspond to the object being analysed?
  - Chosen ciphertext in cryptanalysis means
    - Attacker can choose and decrypt ciphertexts with the secret key
      - barring only the ciphertext to be analysed
  - What is meant in steganalysis
Chosen cover-text oracle model

1. Alice chooses a key pair \((k_P, k_S)\).
2. Eve can choose any ciphertext \(c\) and get decryption \(d_{k_S}(c)\) from an oracle.
3. Eve chooses a message \(m\).
4. Alice draws \(b \in \{0, 1\}\) at random, and creates \(c^*\), where
   - if \(b = 1\), \(c^*\) is a random (innocent) covertext
   - if \(b = 0\), \(c^*\) is a stegogramme containing \(m\).
5. Eve can choose ciphertexts \(c \neq c^*\) and get decryption \(d_{k_S}(c)\) with some exceptions from an oracle.
6. Eve makes a guess \(\hat{b} \in \{0, 1\}\), and wins if \(\hat{b} = b\).

Security assessment in general

1. Hypothetical attacker (Eve)
   - Do some research
   - Make a guess
   - A random guess (\(\hat{b}\) is uniformly distributed) wins 50% of the time.
   - Note that \(b\) is uniformly distributed.
   - If Eve is unable to win significantly more than 50% of the time, then the system is secure.

Classification by information available

1. Tentative steganogram to be analysed (always)
2. System and algorithm (always by Kerckhoffs’ principle)
3. Old steganograms (same key) with additional information
   - Known message
   - Known covertext
   - Chosen message
   - Chosen covertext
   - Chosen steganogram
4. Old steganograms with different key
   - same message or different message?
5. Targeted steganalysis
   - Specialised to detect stego-text from a single system
6. Uniclass steganalysis (fully blind)
   - Not related any particular system
7. Multiclass steganalysis (quasi-blind)
   - Can identify a range of different systems
**Blind and targeted steganalysis**

- **Targeted steganalysis**
  - Taylor-made for a specific stego-system
  - Extremely accurate
  - Completely inflexible
  - Important step in the evaluation of stego-systems
- **Blind steganalysis**
  - Intended to work against any stego-system
  - Can (potentially) identify the stego-system used
  - Rarely as accurate as targeted techniques

**Multiclass steganalysis**

- Targeted attacks may be added for extra scrutiny
- Less blind than uniclass categorisation
- More complicated training process

**Blind steganalysis**

- **Blind steganalysis = classification problem**
- **Common approach is machine learning**
  - Define (many) statistics
  - Train on objects from each class
  - Empirical data → choose threshold
  - Threshold used for decision in new cases
- **Uniclass steganalysis**
  - Cover-image or stegogramme
  - Train on cover-images
- **Multiclass steganalysis**
  - One class per known stego-system + cover-images
  - Train on cover-images and stegogrammes

**Limitations**

- Sensitive to image type
  - Cover-images are widely different
  - Photos vs. drawing; Scanning vs. digital camera; etc.
  - Known techniques are sensitive to different digital cameras(!)
  - Known steganalysis techniques seem to detect double compression rather than embedding.
  - (the bug in the F5 software)
- Sometimes known as **Targeted blind steganalysis**
  - Uses information about the image source (e.g. camera make)
  - No technique is known to be truly blind.
Features

- Features are statistics calculated from the image
  - pixel variance
  - DCT coefficient mean
  - $\chi^2$ statistic (pairs of values)
- Some may be good steganalytic statistics independently
- Others show marginal variation between stego- and normal images
- Machine Learning allows us to aggregate statistics
  - Many weak statistics make a strong one

Feature vectors

- Feature vector: the list of all the extracted features
  \[ \vec{v} = (v_1, v_2, \ldots, v_m) \]
- Farid (2002) used 72 features
- More recent works use more
- We could use a weighted average
  \[ V = \sum_{i=1}^{m} w_i v_i \]
- … but how do we find the weights $w_i$?
- Analytic solution is too complex
  - Machine Learning can work

An example

The abstraction

- Machine learning algorithms work on feature vectors
  - They are oblivious to the images
- The feature vector is a sequence of numbers
  - No semantics
- Advantage of abstraction
  - Reuse algorithms for other applications
  - … typically medicine, bioinformatics, et c.
Machine Learning Techniques

- Classification
  - Steganogram vs. natural image
- Estimation
  - Message length

The training

- Machine learning comprises two steps
  1. Training
  2. Use: Classification or Estimation
- Training requires a large sample
  - Where the machine knows the correct answer
  - The machine builds up experience about the different classes
    - In practical use, it depends on past experience

Training

- Cover Images
- Steganograms
- Machine
- Weights
- Threshold

Analysis

- Weights
- Threshold
- Machine
- Yes/No
- Image
**Different Machine Learning Techniques**

- FLD
- SVM
- Neural Networks

**The conditional probabilities**

- Three sets of features, for 
  
  \((x, y, z) = (a_1, a_2, a_3, b_1, b_2, b_3, c_1, c_2, c_3)\)

- Consider, the conditional probabilities
  
  - \(P(x > y | y > z), P(x > y | y < z), P(x > y | y = z)\)
  - \(P(x < y | y > z), P(x < y | y < z), P(x < y | y = z)\)
  - \(P(x = y | y > z), P(x = y | y < z), P(x < y | y = z)\)

- The probabilities may be estimated
  
  - counting relative frequencies in the JPEG image

- The estimated probabilities are our features
  
  - i.e. \(3 \times 9 = 27\) features

**The coefficients**

- Based on JPEG
- Uses (primarily) 9 coefficients
- due to Ainuddin Abdul Wahab

**Performance**

- Tests shows detection accuracy of 97-99%
- Slightly better than other techniques for medium-size message
  
  - Similar accuracy on long messages
- Computationally fast feature extraction
  
  - Significantly faster than previous techniques
Conclusion

- Steganalysis is not really well understood yet
- There are many error sources
  - Good results can only be guaranteed with accurate knowledge of cover sources
- Machine learning and AI techniques appears to have great potential
  - ignoring the limitations above ...