Steganalysis Methods

STEVENS
Institute of Technology

Outline

- Perfect stego system
- Steganalysis
  - Categories of steganalysis methods
  - Sample pair based steganalysis method for LSB embedding algorithm
  - A wavelet based universal steganalysis example

Perfect Secrecy of Stegosystems

- In order to define secrecy of a stegosystems we need to consider
- probability distribution P_C on the set C of covertexts;
- probability distribution P_M on the set M of secret messages;
- probability distribution P_K on the set K of keys;
- probability distribution P_S on the set $\{ E_K(c, m, k), \mid c \in C, m \in M, k \in K \}$ of stegotexts.
- The basic related concept is that of the relative entropy $D\,(P_1 \,||\, P_2)$ of two probability distributions P_1 and P_2 defined on a set Q by

$$D(P_1 \| P_2) = \sum_{q \in Q} P_1(q) \lg \frac{P_1(q)}{P_2(q)},$$

- which measures the inefficiency of assuming that the distribution on Q is P_2 if it is really P_1.

Definition Let S be a stegosystem, P_C the probability distribution on covertexts C and P_S the probability distribution of the stegotexts and $\varepsilon > 0$. S is called – ε-secure against passive attackers, if

$$D\,(P_C \,||\, P_S) \leq \varepsilon$$

and perfectly secure if $\varepsilon = 0$.

Perfect Secrecy of Stegosystems cont'd

- A perfectly secure stegosystem can be constructed out of ONE TIME-PAD CRYPTOSYSTEM

- Theorem There exist perfectly secure stegosystems.

  Proof. Let n be an integer, $C_n = \{0,1\}^n$ and P_C be the uniform distribution on C_n, and let $m \in C_n$ be a secret message.

  The sender selects randomly $c \in C_n$, computes $c \oplus m = s$. The resulting stegotexts are uniformly distributed on C_n and therefore $P_C = P_S$ from what it follows that

$$D(P_{Cn} \mid\mid P_S) = 0.$$

  In the extraction process, the message m can be extracted from s by computation

$$m = s \oplus c.$$

Detecting Secret Messages

- The main goal of a passive attacker is to decide whether data sent to Bob by Alice contain a secret message or not.

- The above task can be formalized as a statistical hypothesis-testing problem with the test function $f: C \to \{0,1\}$:

- $$f(c) = \begin{cases} 1, & \text{if } c \text{ contains a secret message;} \\ 0, & \text{otherwise} \end{cases}$$

- There are two types of errors possible:

- Type-I error - a secret message is detected in data with no secret message;

- Type-II error - a hidden secret message is not detected

- Practical steganography tries to minimize probability that passive attackers make type-II error. In the case of ε-secure stegosystems there is well know relation between the probability β of the type II error and probability α of the type I error.

- Theorem Let S be a stegosystem which is ε-secure against passive attackers and let β be the probability that the attacker does not detect a hidden message and α be the probability that the attacker falsely detects a hidden message. Then $d(\alpha,\beta) \le \varepsilon$, where $d(\alpha,\beta)$ is the binary relative entropy defined by

$$d(\alpha,\beta) = \alpha \lg \frac{\alpha}{1-\beta} + (1-\alpha)\lg \frac{1-\alpha}{\beta}.$$

Detecting Secret Messages

Definition Let S be a stegosystem and P be a class of mappings $C \rightarrow C$. S is P-robust, if for all $p \in P$

$$D_K(p(E_K(c, m, k)), k) = D_K(E_K(c, m, k), k) = m$$

in the case of a secret-key stegosystem and

$$D(p(E(c, m))) = D(E(c, m)) = m$$

in the case of pure stegosystem, for any m, c, k.

- There is a clear tradeoff between *security* and *robustness*.
- Some stegosystems are designed to be robust against a specific class of mappings (for example JPEG compression/decompression).
- There are two basic approaches to make stegosystems robust:

  - By foreseeing possible cover modifications, the embedding process can be robust so that possible modifications do not entirely destroy embedded information.

  - Reversing operations that has been made by an active attacker.

ACTIVE and MALICIOUS ATTACKS

- At the design of stegosystems special attention has to be paid to the presence of active and malicious attackers.
- Active attackers can change cover during the communication process.
- An attacker is malicious if he forges messages or initiates a steganography protocol under the name of one communicating party.
- In the presence of a malicious attacker, it is not enough that stegosystem is robust.
- If the embedding method does not depend on a key shared by the sender and receiver, then an attacker can forge messages, since the recipient is not able to verify sender's identity.

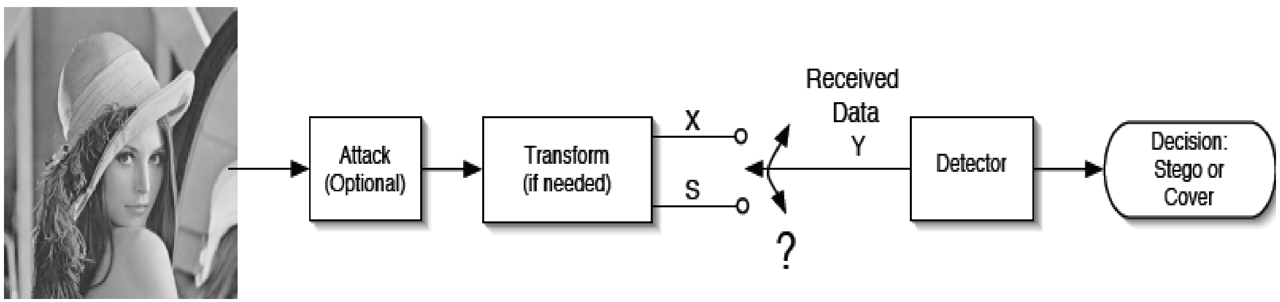
SECURITY of STEGOSYSTEMS

- Definition A steganographic algorithm is called secure if
- Messages are hidden using a public algorithm and a secret key. The secret key must identify the sender uniquely.
- Only the holder of the secret key can detect, extract and prove the existence of the hidden message. (Nobody else should be able to find any statistical evidence of a message's existence.)
- Even if an enemy gets the contents of one hidden message, he should have no chance of detecting others.
- It is computationally infeasible to detect hidden messages.

Steganalysis

- Definition
  - Searching for the existence of hidden messages or Stego-content in a given medium.
- Stego-only: only stego-medium is available for analysis
- Known cover: both original cover media and stego-media are used
- Known message: hidden message is revealed to facilitate review of media in preparation for future attacks

Goals

- Passive steganalysis
  - 

- Active Steganalysis
  - Estimate the message length and location
  - Determine the algorithm/Stego tool
  - Estimate the Secret Key in embedding
  - Extract the message

Types of Steganalysis

- Model based steganalyis: only works for specific embedding algorithm
  - Good detection accuracy for the specific technique
- Universal Steganalysis: work for multiple steganography methods
  - Less accurate in detection
  - Usable on new embedding techniques

Universal Steganalysis Techniques

- Techniques which are independent of the embedding technique
- Identify certain image features that reflect hidden message presence.
- Two problems
  - Calculate features which are sensitive to the embedding process
  - Finding strong classification algorithms which are able to classify the images using the calculated features

Supervised learning based Steganalysis

- Supervised learning methods construct a classifier to differentiate between stego and non-stego images using training examples.
- Some features are first extracted and given as training inputs to a learning machine. These examples include both stego as well as non-stego examples.
- The learning classifier iteratively updates its classification rule based on its prediction and the ground truth. Upon convergence the final stego classifier is obtained.

Statistical detection based Steganalysis

- a) For completely known statistics case, the parametric models for stego-image & cover image.
- b) For partially known statistics case, the parametric probability models are available but, not the exact parameter models. These parameters are estimated
- c) For completely unknown case, Bayesian prior models are assumed and detectors are developed.

Sample Pair Analysis for LSB Steganography

- The sample pair is ***P*** *(s_i,s_j)* drawn from the signal sequence sample s_1, s_2, …s_N, and their values are in the range $0…2^b$-1.

- $\boldsymbol{D_n}$ is the submultiset of ***P*** that consists of sample pairs of the form *(u, u+n)* or *(u+n, u)*, where *u* is value of one sample in ***P*** and $0 \le u, n \le 2^b - 1$.

- $\boldsymbol{C_m}$ is the submultiset of ***P*** that consists of the sample pairs whose value differ by *m* ($0 \le m \le 2^{b-1} - 1$) in the first *(b-1)* bits (i.e., by right shifting one bit and then measuring the difference)

Modification Pattern

- For each modification pattern $\pi \in \{00,10,01,11\}$, and any submultiset $A \subseteq P$, denote by $\rho(\pi, A)$ the probability that the sample pairs of *A* are modified with pattern as a result of the embedding.

- Let *p* be the length of the embedded message in bits divided by the total number of the samples. Then, the fraction of the samples modified by the LSB embedding is *p/2*.Then, then submultisets after embedding are as follows:

$$\rho(00, P) = (1 - p/2)^2$$

$$\rho(01, P) = \rho(10, P) = p/2 * (1 - p/2)$$

$$\rho(11, P) = (p/2)^2$$

Finite state machine between submultisets

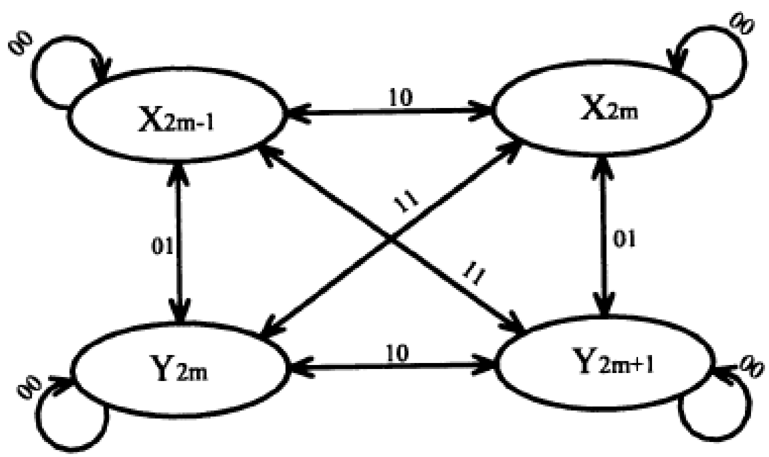


Fig. 1. Finite-state machine whose states are trace multisets of C_m. Note that C_m is closed under LSB steganography, but its four subsets are not.

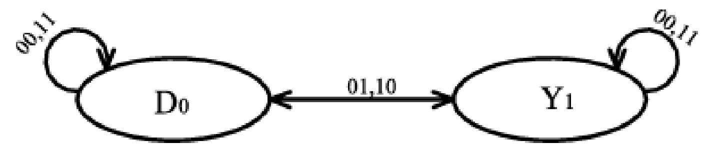


Fig. 2. Finite-state machine associated with C_0.

- X_{2m} is the pixel value pair: (2k-2m,2k) or (2k+1, 2k-2m+1)
- X_{2m+1} is the pixel value pair: (2k-2m-1,2k) or (2k,2k-2m-1)
- Y_{2m} is the pixel value pair: (2k-2m+1,2k+1) or (2k,2k-2m)
- Y_{2m+1} is the pixel value pair: (2k-2m,2k+1) or (2k+1,2k-2m)

The relationships between the cardinality of subsets

$$|X_{2m-1}|(1-p)^2 = \frac{p^2}{4}|C_m| - \frac{p}{2}\left(|D'_{2m}| + 2|X'_{2m-1}|\right) + |X'_{2m-1}| \quad (1)$$

$$|Y_{2m+1}|(1-p)^2 = \frac{p^2}{4}|C_m| - \frac{p}{2}\left(|D'_{2m}| + 2|Y'_{2m+1}|\right) + |Y'_{2m+1}| \quad (2)$$

$$|Y_1|(1-p)^2 = |C_0|\frac{p^2}{2} - \frac{p}{2}\left(2|D'_0| + 2|Y'_1|\right) + |Y'_1| \quad (3)$$

The quadratic equations of value p

Based on the experimental observation:

$$E\{|X_{2m+1}|\} = E\{|Y_{2m+1}|\} \qquad \textbf{(4)}$$

Combining with (1) – (4) yields:

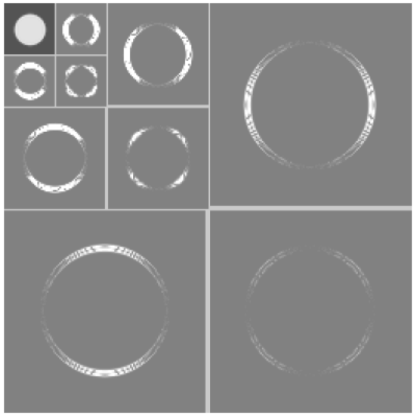
$$\frac{(|C_m| - |C_{m+1}|)\,p^2}{4} - \frac{(|D'_{2m}| - |D'_{2m+2}| + 2|Y'_{2m+1}| - 2|X'_{2m+1}|)p}{2} + |Y'_{2m+1}| - |X'_{2m+1}| = 0, \qquad m \geq 1 \qquad \textbf{(5)}$$

$$\frac{(2|C_0| - |C_1|)p^2}{4} - \frac{(2|D'_0| - |D'_2| + 2|Y'_1| - 2|X'_1|)p}{2} + |Y'_1| - |X'_1| = 0, \qquad m = 0. \qquad \textbf{(6)}$$

- By finding the smallest root of equation (5), or (6), the percentage *p* of embedding data can be estimated.

Wavelet-based Universal Steganalysis

- Wavelet transform is used to obtain the features.
- The mean, variance, skewness and kurtosis of the sub band coefficients at each location, scale and color channel forms features. i.e. 12(n-1) features per color. n: Number of scales.
- usually 4 scales are used.
- therefore 36 features per color channel.



Wavelet-based Universal Steganalysis

- In order to capture higher order statistical correlations second set of 36 features per color are found based on the errors in a linear predictor of coefficient magnitude.
- For green channel at scale i ,

$$\begin{aligned} |V_i^g(x,y)| &= w_1|V_i^g(x-1,y)| + w_2|V_i^g(x+1,y)| + w_3|V_i^g(x,y-1)| + w_4|V_i^g(x,y+1)| \\ &+ w_5|V_i^g(x/2,y/2)| + w_6|D_i^g(x,y)| + w_7|D_{i+1}^g(x/2,y/2)| + w_8|V_i^r(x,y)| + w_9|V_i^b(x,y)| \end{aligned}$$

- This can be written in the matrix form as,

$$\vec{v} = Q\vec{w},$$

- $\vec{w}$ is found by minimizing, $E(\vec{w}) \;=\; [\vec{v} - Q\vec{w}]^2$

Wavelet-based Universal Steganalysis

- Therefore $\vec{w}$ is found by solving

$$\frac{dE(\vec{w})}{d\vec{w}} = 2Q^T(\vec{v} - Q\vec{w}) = 0$$

Which yields, $\vec{w} = (Q^TQ)^{-1}Q^T\vec{v}.$

- The log error between the actual & predicted coefficients is,

$$\vec{p} = \log(\vec{v}) - \log(|Q\vec{w}|)$$

- Then the mean, variance mean, variance, skewness and kurtosis of this log error is used as another 36 features per color.