The Ultimate Steganalysis Benchmark?

Andrew Ker
adk@comlab.ox.ac.uk

Royal Society University Research Fellow
Oxford University Computing Laboratory

ACM Multimedia & Security Workshop
21 September 2007
The Ultimate Steganalysis Benchmark?

Outline

• Currently-used benchmarks not ideal
• New benchmark based on KL divergence
• Difficulties estimating the benchmark value
• Examples
Binary Steganalysis

payload size $p$

cover $\rightarrow$ embedding $\rightarrow$ cover or stego $\rightarrow$ decision

"positive" or "negative"

steganographer steganalyst
Common Benchmarks

- ROC curve
  \textit{difficult to rank; too much information}

- Area under ROC

- Minimize sum of false positive & negative
  \textit{assumes false positive and false negatives are equivalent}

- False negative rate at fixed false positive
- False positive rate at fixed false negative
  \textit{impossible to justify numbers objectively}
Common Benchmarks

- ROC curve
  
  difficulties to rank; too much information

- Area under ROC

- Minimize sum of false positive & negative
  
  assumes false positive and false negatives are equivalent

- False negative rate at fixed false positive

- False positive rate at fixed false negative

  impossible to justify numbers objectively

- Minimum payload detectable at fixed false positive & false negative rate

  impossible to justify numbers objectively
Distribution Differences

steganography

payload size $p$

covers $\rightarrow$

embedding

feature extraction

dimension reduction

threshold test

"positive" or "negative"

steganalysis

covers $\rightarrow$

feature extraction

dimension reduction

threshold test

"positive" or "negative"
Distribution Differences

**steganography**

- payload size $p$
- covers -> embedding
- covers -> feature extraction
- covers -> dimension reduction
- covers -> threshold test

**steganalysis**

- $C_p$
- $D_{KL}(C_0, C_p)$
- “positive” or “negative”

- “positive” or “negative”
New Benchmark

- Based on $D_{KL}(C_0, C_p)$, where $C_p$ is the univariate distribution produced just before threshold test.

From steganalysis/info theory literature

*If steganography is repeated at a fixed embedding rate, the probability of detection tends to 1.*  
  
  [Cachin; Moulin; Ker; ...]

- For long-run performance we should concentrate on payload sizes tending to zero.

A theorem by S. Kullback

*Let $F_p$ be a family of distributions satisfying certain regularity conditions. Then*  
\[
\lim_{p \to 0} \frac{D_{KL}(F_0, F_p)}{p^2}
\]
*exists and is nonzero.*  
  
  [adapted from Kullback, 1968]

- If we believe that the regularity conditions are satisfied, then $D_{KL}(C_0, C_p)$ is, locally to zero, a multiple of $p^2$. 
New Benchmark

The quantity \( Q = \lim_{p \to 0} \frac{D_{KL}(C_0, C_p)}{p^2} \) tells us how quickly “evidence” accumulates. This is the proposed benchmark.

\textit{Note:}

- “Payload size” should be measured by number of embedding changes
- Then \( Q \) is measured in “nats per embedding change squared”
Experimental Results

Estimate KL divergence* by [Wang, Kulkarni, & Verdu, 2005]
Experimental Results

- 10000 cover images
- LSB replacement embedding, 50 payload sizes, repeated 10 times each
- "Triples" steganalysis
Experimental Results

Optimal method for estimating $Q$?

Current heuristics depend on hand-picked “sensible” choice of $p$.

- 10000 cover images
- LSB replacement embedding, 50 payload sizes, repeated 10 times each
- “Triples” steganalysis
Experimental Results

- 20000 cover images
- LSB matching (±1) embedding, 90 payload sizes, repeated 10 times each
- “Calibrated HCF COM” steganalysis
Conclusions

• There is a need for an application-independent benchmark.

• The new “Q-factor” benchmark measures how quickly information is accumulated as payload increases.

• More work needed for good empirical estimation of “Q”:
  – Currently seems to need a very large experimental base
  – Test objects should be the same size
  – Optimal estimation?
Conclusions

- There is a need for an application-independent benchmark.
- The new “Q-factor” benchmark measures how quickly information is accumulated as payload increases.
- More work needed for good empirical estimation of “Q”:
  - *Currently seems to need a very large experimental base*
  - *Test objects should be the same size*
  - *Optimal estimation?*

<table>
<thead>
<tr>
<th>Steganalysis</th>
<th>3000 grayscale bitmap covers</th>
<th>10000 colour JPEG covers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SPA</strong></td>
<td>16.1</td>
<td>28.3</td>
</tr>
<tr>
<td>[Dumitrescu et al, IHW 2002]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SPA/LSM</strong></td>
<td>12.1</td>
<td>161</td>
</tr>
<tr>
<td>[Lu et al, IHW 2004]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Triples</strong></td>
<td>20.7</td>
<td>1500</td>
</tr>
<tr>
<td>[Ker, IHW 2005]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Triples/WLSM</strong></td>
<td>16.1</td>
<td>1500</td>
</tr>
<tr>
<td>[Ker, SPIE EI 2007]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*nanonats per embedding change squared*
End