

# Detecting Covert Communications on the Internet: Some Challenges and Solutions

R. Chandramouli ("Mouli")  
Stevens Institute of Technology

DFRWS 2003



Research sponsored by :U.S. Air Force Research Laboratory  
National Science Foundation



## Covert Channel Definition

---

Any communication channel that can be exploited by a process to transfer information in a manner that violates the system security policy.

--- U.S.D.O.D. 1985, Trusted computer system evaluation criteria.

# Example Internet Covert Channels

- TCP/IP Protocol
  - » Unused header fields.
  - » Encoding information in sequence numbers.
- Timing Channel
  - » Encode covert info. in the rate at which jobs are sent to a time-shared server.
  - » Measuring response times to jobs gives noisy version of message.
- Digital media (image, video...) on the web

# Example Covert Messages

---

- Spy programs
- Company proprietary information.
- Computer virus...

# Issues in Intercepting Covert Messages on Internet (I)

- What to look for in the Internet?
- Where to look? How to identify web links that could potentially contain covert messages?
  - » Side information may be available.
  - » No side information at all.
    - random search may be futile.
    - metrics used by current web search engines may not work: e.g., popular sites.

# Issues in Intercepting Covert Messages on Internet (II)

- Message carrying website may have not a public link
  - » Use http traffic request in the back bone to identify these hidden links?
- Websites are created, moved, and destroyed randomly on a daily basis
  - » Continually monitor websites of interest?
  - » How often to monitor?
- A web-page like e-bay could contain thousands of images
  - » Efficient search techniques.

## Some Approaches

- Candidate websites for investigation could be chosen as follows:
  - » External info. such as email trace, phone tapping, etc.
  - » Eliminate certain sites such as .mil, .gov...
  - » Past history.
  - » Religious cult's website?
  - » Websites of groups with radically politically opposed views?
  - » Info. from network forensic tools.

# Steganography Covert Channel Requirement

- Maximize stealth
  - » Detection via steganalysis is "difficult."
  - » Perceptually transparent
- Maximize capacity
  - » Maximum embedded message length such that steganalysis detection is "difficult."
- Efficient encoding/decoding

# Intercepting Steganographic Messages (I)

- **Steganalysis**
  - » Analyze digital data to determine presence of secret messages.
- **Passive Steganalysis**
  - » Steganalyst/hacker/interceptor tries to find if a secret message is present.
  - » Identify the embedding algorithm/software used.
  - » Removal of secret message is not an aim.
  - » Little or no *a priori* information available.
- **Active Steganalysis**
  - » Estimate the secret key, message length, etc.
  - » Estimate the secret message (grand goal!).

# Intercepting Steganographic Messages (II)

- Theoretical issues.

- » What must a steganalysis algorithm look for?
- » What are the "give aways" in current published steganography algorithms/software?
- » What is the minimum message length that can be detected?
- » What about false alarm and miss probabilities?
- » Mathematical tools from probability and statistics.

- Scalability.

- » Investigating every web site is not possible.
- » How often web sites are to be investigated?
- » Can we identify "high risk" sites in some sense?
- » Number of possible embedding algorithms could be large.
- » Message sizes could be small.

# Intercepting Steganographic Messages (III)

- Is steganalysis realistic?
  - » Current approaches seem to be extreme:
    - tuned to work for one particular embedding algorithm or use large training data set.
  - » What if the embedding algorithm is not published publicly?
  - » Need: steganalysis that works for a "class" of embedding algorithms.

# Steganalysis Current Trends (I)

- Classifier/statistical learning based
  - » Train steganalysis classifiers on large training sets
  - » Use host data features.
- Pros
  - » Well understood classifier theory
  - » Works reasonably well in practice
- Cons
  - » May not work well for data that are significantly different from training set
  - » Overfitting problems
  - » How to choose training set? How large a training set? Mostly heuristics involved here.

## Steganalysis Current Trends (II)

- Blind statistical system identification based
  - » Use individual host data features.
  - » No training set.
- Pros
  - » Sound theoretical analysis possible.
  - » Covert message extraction demonstrated.
- Cons
  - » Stochastic non-stationarity of digital data, e.g., images.
  - » New tricks needed to make it work in practice.

# Optimal Web Search for Covert Message: A Mathematical Model

- Let,
  - » Total number of web sites to be searched =  $W$ .
  - »  $P(j) = \text{Pr}(\text{website "j" contains the covert message})$ .
  - »  $b(j,t) = \text{Pr}(\text{detecting covert message after spending } t \text{ time units in site } j \mid \text{message in site } j)$ .
  - »  $C(j,f(t)) = \text{cost of searching site } j \text{ with a time/resource allocation of } f(t)$ .
  - »  $P(f) = \sum_j P(j)b(j,f(t))$  is average probability of finding covert message.
  - »  $C(f) = \sum_j C(j,f(t))$  is total cost.

## Possible Scenarios

- Case 1:  $\{P(j); j=1,2,\dots,W\}$  completely known.
  - » Subjectively chosen.
- Case 2: Only ordering of probabilities known, i.e.,  $P(1) > P(2) > \dots > P(W)$ 
  - » More realistic.
- Case 3:  $\{P(j)\}$  completely unknown.
  - » When no side info. available.

- Web search strategy for Case 1:
  - »  $P(f^*) = \max P(f)$  subject to  $C(f) < T$ .
- $f^*(t)$  is the optimal allocation of time to search each of the web sites.
- Suppose  $b(j,t) = 1 - e^{-t}$ ,  $t > 0$  then,  $f^*(t)$ :

$$t_j = \max(0, \ln(p_j/K)); j=1,2,\dots,W$$

where  $K = [\prod p_j]^{1/W} e^{-T/W}$

$$1 - e^{-T/M} \leq P(f^*) \leq 1 - T [\prod p_j]^{1/W} e^{-T/W}$$

- » For a desired covert message detection accuracy, bound on total required resource  $T$  can be computed.

## Web Search Strategies (I)

- Let probability of detecting covert website/channel =  $q$
- Probability of success statistically independent from one attempt to another.
- Possible search strategies:
  - » Co-ordinated strategy.
  - » Randomized strategy.

## Web Search Strategies (II)

- Co-ordinated search
  - » Results of previous search results stored. That is, "memory" is built into searching.
  - » Web links previously searched, images previously investigated, etc. are stored.
  - » Avoid these links/data in future searches.
  - » Pros: Optimal strategy because of the memory.
  - » Cons: Large storage needed, cached data could become outdated...

## Web Search Strategies (III)

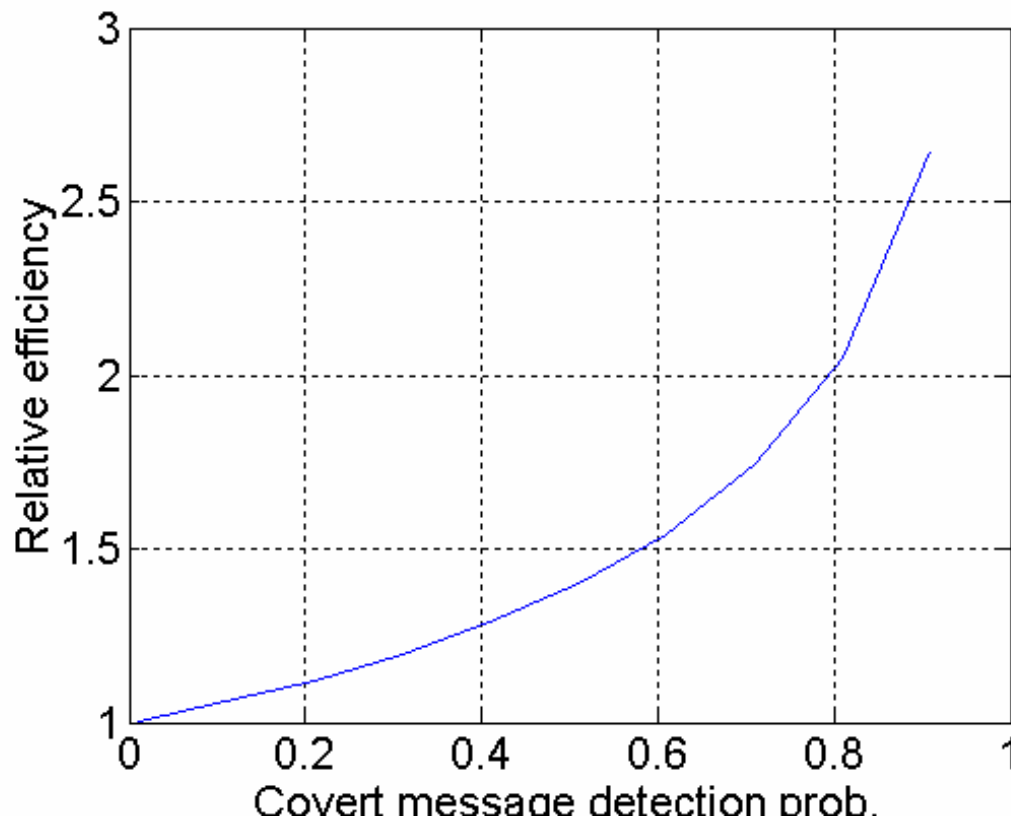
- Randomized search
  - » No memory built.
  - » Search web sites randomly.
  - » Pros: Large storage is not needed.
  - » Cons: Does not exploit memory.

## Search Relative Efficiency (I)

- $N_r$  = no. of times a web site is searched for detecting a covert message using randomized search.
- $N_c$  = no. of times a web site is searched for detecting a covert message using co-ordinated search.
- $d_r$  = prob. Of detecting covert message using randomize search.
- $d_c$  = prob. Of detecting covert message using randomize search.

## Search Relative Efficiency (II)

- If  $d_r = d_c$  then, rel. eff. of the random search w.r.t. co-ord. =  $-\ln(1-q)/q$ .



## Key Observation

---

- If covert message detection reliability is low, then co-ordinated and randomized searches are nearly equally efficient.

# Additional Information

---

<http://www.ece.stevens-tech.edu/~mouli>

email: [mouli@stevens-tech.edu](mailto:mouli@stevens-tech.edu)