

# Web Search Steganalysis: Some Challenges and Approaches

R. Chandramouli

Multimedia Systems, Networking, and Communications (MSyNC) Laboratory

Department of Electrical and Computer Engineering

Stevens Institute of Technology

Email:mouli@stevens-tech.edu

**Abstract**—This paper presents some important issues in searching the Internet for covert messages. Several related theoretical and practical problems are discussed. Web search is formulated as a mathematical optimization problem and solutions are proposed. Two algorithms: *coordinated search* and *random search* are discussed within this optimization framework along with their pros and cons.

## I. INTRODUCTION

A covert channel is defined as: *any communication channel that can be exploited by a process to transfer information in a manner that violates the systems security policy* [1]. It seems that the Internet has opened up new avenues for covert communication. Some of them are the following:

- **Digital data as covert channel:** digital data such as image, video, and audio are easily created, manipulated, and distributed across the Internet with the use of current technologies and devices. These are good channels for hiding information with medium to high message carrying capacities. There are several techniques available today that can embed messages within these data types without causing significant perceptual distortion. For example, the least significant bits of a digital image can be replaced by message bits. The resultant image can then be placed on a website for a receiving party to download and extract the hidden message. More about this topic can be found in [2].
- **TCP/IP protocol as covert channel:** Some features of the TCP/IP protocol suite can be used to send covert messages as discussed in [3]. Encrypted or non-encrypted information can be encapsulated within otherwise normal TCP/IP packets. The TCP/IP header information can also be modified to encode secret messages. There are some fields in

the packet header that are not used by the current communication networks. These fields can be used as message carriers.

- **Timing channel:** An user of a time-shared computing server can transmit covertly by varying the rate at which it sends jobs for processing. Since the response time of the computing server depends on its instantaneous load, other users can get a *noisy* version of the covert information by measuring the response time to their own jobs. One of the earliest work on this topic is by Lipner [4].

These covert channels are an immense cause of security concern because they can be used to pass malicious messages. These messages could be in the form of computer viruses, spy programs, terrorist messages, etc. Therefore, detecting these covert channels is an important issue that needs to be addressed.

Information hiding/embedding also known as *steganography* is a popular research area currently. Hiding information in digital medium such as a digital image has been receiving a lot of attention these days. Steganalysis is the counterpart of steganography that deals with detecting data that could contain hidden messages. Current steganalysis algorithms use statistical methods to determine the presence/absence of a hidden message. In this paper we assume image steganography without loss of generality. Also, we focus on detecting digital medium as covert channels in the Internet. We believe that the techniques presented here could be applied in detecting some other types of covert channels also. A work closely related to this paper is by Bloom [5].

First, we remark that every website and every digital image on these sites is a potential covert channel. However, it is clear that we cannot search every website for covert messages. In

addition to the sheer problem of the large number of websites, we also face the following issues:

- Some covert message carrying websites may not have public links.
- Websites are created, moved, and destroyed on a daily basis, perhaps even randomly.
- Websites carrying covert messages may not be found by current search engines because of the web search metrics used by them. For example, some search engines show webpages that are linked to by other webpages. We have no reason to believe that a webpage containing images with covert messages will have several links pointing to it. This beats the purpose of covert communication!
- A webpage like e-bay could contain thousands of images thus making it computationally very difficult to detect hidden messages.

Since exhaustive search of the Internet is infeasible, some natural follow-up questions are the following:

- How can the search efficiency be improved?
- How can *a priori* information about the Internet sites be exploited?
- What are the options for an efficient resource allocation for covert message search?

In this paper, we attempt to address some of these questions and provided a mathematical formulation of this problem. The solution to this mathematical formulation gives us some insights on how to design algorithms for web search to detect covert communication. The paper is organized as follows. Web search steganalysis is posed as an optimization problem in Section II followed by concluding remarks in Section III.

## II. WEB SEARCH STEGANALYSIS AS AN OPTIMIZATION PROBLEM

Consider the following web search steganalysis problem formulation. Let us say that a covert message  $X$  may be present in one of  $M$  webpages with a priori probabilities  $p_j$ ,  $j = 1, 2, \dots, M$ . Note that the set of  $M$  webpages of interest can be chosen in a variety of ways, such as:

- An external intelligence information such as email trace, tapping phone conversations, etc. could raise suspicion about certain websites.
- As discussed in [5], some websites can be safely eliminated before the search begins due its security level. For instance, there is little reason to believe that government

websites (.mil, .gov) may contain images with hidden information.

- Websites of organizations with radical political or religious views may be a good candidate for search. Such a suspicion could also be strongly supported by the text content of the website.
- Tracking http requests in the backbone network [5], we can find out about sites that do not have a publicly available link that may be a cause for concern.

Once some kind of a criterion or side information is used to short list the candidate websites for searching, the next issue is to compute the probabilities  $p_j$ ,  $j = 1, 2, \dots, M$ . There are three possible scenarios:

- **(a) complete information case:** the probabilities  $\{p_j\}$  are completely known.
- **(b) partial information case:** only an ordering of these probabilities are known, i.e., it is known that  $p_1 \geq p_2 \geq \dots \geq p_M$ , but not their exact values.
- **(c) no information case:**  $\{p_j\}$  completely unknown.

Each of these three cases occur could occur in practice. Complete information about  $\{p_j\}$  may be available to government intelligence and law & order agencies. Partial information about the probabilities may be obtained with the help of some side information from law enforcement agencies, monitoring some suspect web sites, Internet chat rooms, tracing http requests, etc. The last case of course occurs when a blind web search is conducted for stego information. Nevertheless, in this paper we provide detailed analysis of case (b) only. We however note that this analysis is extendable to the other two cases also.

If  $J$  denotes the set of websites in which the message  $X$  may be found, then, the message location distribution on  $J$  is  $p : J \rightarrow [0, 1]$  and  $\sum_{j \in J} p_j = 1$ . Since the websearch is limited by total search time/effort due to a variety of physical and logical limitations, it may be necessary to locate  $X$ 's website with the minimal amount of time/effort. Let  $b : J \times [0, \infty) \rightarrow [0, 1]$  be the *location function* such that  $b(j, z)$  denotes the conditional probability of locating  $X$  with the amount of search effort spent in site  $j$ ,  $z \geq 0$ , given that  $X$  is in site  $j$ . Note that the value of  $b$  is affected by a number of factors such as the quality of the steganalysis algorithms being used to analyze the images in a website, computational resources available, quality and amount of side information about the embedding algorithm, etc. Then

the total probability of locating the message  $X$  is given by,

$$\sum_{j \in J} p_j b(j, f(j)), \quad (1)$$

where  $\sum_{j \in J} f(j)$  is the total effort. A cost function on  $J$  is a function  $c : J \times [0, \infty) \rightarrow [0, \infty)$  such that  $c(j, z)$  gives the cost of applying search effort  $z$  in site  $j$ . The function  $f : J \rightarrow [0, \infty)$  gives the amount of effort spent in each site, called an *allocation on  $J$* . Then,

$$P[f] = \sum_{j \in J} p_j b(j, f(j)), \quad C[f] = \sum_{j \in J} c(j, f(j)) \quad (2)$$

denotes the probability of locating the covert message and the cost resulting due to the allocation  $f$ , respectively. As an example, if  $f(j) = T_j$  denotes the amount of search time spent to localize  $X$  then  $c(j, \cdot) = T_j$  denotes the amount of time spent in searching site  $j$ . Note that, other allocation functions can also be used depending on the resource constraint. Suppose  $F$  denotes the allocations and  $T$  is the upper bound on the total time cost in the localization search process, then, the proposed search problem is to find  $f^* \in F$  by solving the following constrained optimization problem,

$$C[f^*] \leq T, \quad \text{and} \quad P[f^*] = \max \{P[f] : f \in F \text{ and } C[f] \leq T\}. \quad (3)$$

Then  $f^*$  is the optimal message localization search allocation for total time cost  $T$ . Note the resemblance of this formulation with optimal search theory [6].

Suppose  $b(j, T_j) = 1 - e^{-T_j}$ ,  $j = 1, 2, \dots, M$ , that is, the probability of locating a message in site  $j$  increases exponentially with the amount of time spent in searching that site. At present, we only have an intuitive justification for this detection probability model. However, we note that this model may not be far from a practical model. In practice, suppose there are  $L$  ( $L$  large) number of steganalysis algorithms that we wish to run on all the images downloaded from a website. It is reasonable to assume that as the number of steganalysis algorithms run on the images increases, the probability of the covert message not being detected by any one of these will decrease. Of course, here we assume that the individual steganalysis algorithms are such that the rate of decrease of the miss probability is exponential. The proposed approach can also be extended to other models for  $b(\cdot, \cdot)$ .

Let  $0 < \sum_j T_j \leq T$  be the total time constraint for web-search steganalysis. Then, following an analysis presented in [6] we show here without proof that the probability of locating

the message with optimal search effort allocation satisfies:

$$1 - e^{-T/M} \leq P[f^*] \leq 1 - T \left[ \prod_{j=1}^M p_j \right]^{\frac{1}{M}} e^{-T/M}, \quad (4)$$

and, if  $\sum_j T_j = T$  then the optimal search time that must be allocated for searching site  $j$  is given by,

$$T_j = \max(0, \ln \frac{p_j}{\lambda}), \quad j = 1, 2, \dots, M \quad (5)$$

$$\lambda = \left[ \prod_{j=1}^M p_j \right]^{\frac{1}{M}} e^{-T/M}. \quad (6)$$

We observe that, from Eq. (4) it is possible to estimate the total time bound  $T$  for a desired message localization accuracy.

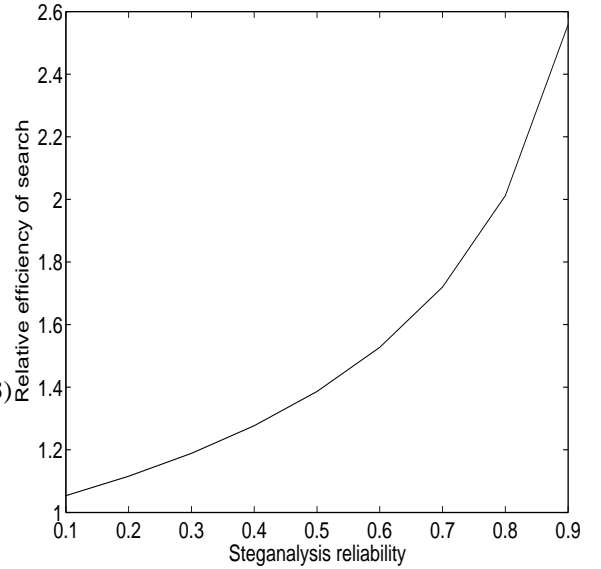


Fig. 1. Relative efficiency of random search w.r.t. coordinated search. Higher value of relative efficiency implies coordinated search strategy is more efficient.

Next, we consider a second model for web search. Let us say a particular website is being searched to locate a message. Due to practical constraints such as computational constraints, time constraints, modelling errors, and reliability of steganalysis algorithms being used, it may only be possible to detect the presence of the covert message with probability  $q$  in each search attempt independent of previous searches. Also, assume that false alarms are negligible. This is justified because most of the steganalysis algorithms allow us to put an upper bound on the false alarm probability. Now, there are two possible ways to search the website, namely, *coordinated search* and *random search*. In a coordinated search, the images within a site that were analyzed previously for a covert message and did not yield a positive result are completely stored within the searching al-

gorithm's/computer's memory. These images are avoided in future searches. On the other hand, in a random search, the search algorithm does not maintain the list of previously searched images thus eliminating the need for large built-in memory requirement. Instead, it simply searches the entire website in a random fashion. Clearly, the coordinated search seems to be more effective than random search; however, its complexity and memory requirement are higher. The basic question we ask is: when is choosing the random search better than coordinated search, if at all? The answer to this question will give us valuable information in trading off efficiency for cost and simplicity. If  $N_c$  is the total number of times the website is searched, then the probability of detecting the covert message is  $d_c = 1 - (1 - q)^{N_c}$ . For the random search strategy, by adopting the technique given in [6], we can see that the detection probability is  $d_r = 1 - e^{-qN_r}$ . Suppose we want  $d_c = d_r$ , then we see that it produces the following relative efficiency factor of the random search message detection w.r.t. coordinated search,

$$\Lambda = \frac{N_r}{N_c} = -\ln(1 - q)/q, \quad 0 < q < 1. \quad (7)$$

From Fig. 1 we see that if the steganalysis accuracy is not high, then random search performs almost as good as coordinated search. In addition, no extra memory cost or coordination is necessary.  $\Lambda$  in Eq. (7) also serves as a useful measure to compare the performance of any other web search algorithm w.r.t. the coordinated search which is optimal.

### III. CONCLUSION

From the analysis presented here, we conclude that it is possible to design efficient web search algorithms to detect covert messages. The proposed mathematical web search model admits a wide variety of resource constraints. Depending on the application, implementation, hardware, and steganalysis probability of error constraints, a suitable resource model can be used to derive an optimal web search strategy using the proposed technique. Depending on the reliability of the steganalysis algorithms employed and the storage constraint one of two strategies, namely, coordinated search or random search can be chosen. It is seen that for a certain range of steganalysis reliability, both these methods give comparable performance.

### ACKNOWLEDGEMENT

This material is based on research sponsored by Air Force Research Laboratory under agreement number F306020-02-2-

0193 and NSF DAS 0242417. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Air Force Research Laboratory or the U.S. Government.

### REFERENCES

- [1] U. S. D. O. D. 1985., "Trusted computer system evaluation criteria."
- [2] I. Cox, J. Bloom, and M. Miller, *Digital Watermarking: Principles & Practice*. Morgan Kaufmann, 2001.
- [3] C. Rowland, "[http://www.firstmonday.dk/issues/issue2\\_5/rowland/](http://www.firstmonday.dk/issues/issue2_5/rowland/)."
- [4] S. Lipner, "A comment on the confinement problem," *Fifth symposium on Operating systems principles*, pp. 192–197, Nov. 1975.
- [5] J. Bloom, "Smartsearch steganalysis," *SPIE Conf. on Security and Watermarking of Multimedia Contents*, vol. 5020, pp. 167–172, 2003.
- [6] L. Stone, *Theory of optimal search*. Academic Press, 1975.