Detection of Steganographic Information in Tags of Webpage

Work-in-Progress

Huang Hua-jun, Sun Xing-ming, Sun Guang, Huang Jun-wei
{hhj0906, sunnudt, simon5115, albert.holmes}@163.com
School of Computer & Communication, Hunan University, Changsha 410082, P.R. China

Abstract

Secret messages can be embedded in a webpage by switching the uppercase-lowercase states of letters in tags. In this paper, a novel steganalytic approach called Tag-Mismatch analysis for detection of hidden information embedded in tags is presented.

1. Introduction

Covert communication can use a text file or webpage as cover objects [1-2]. The existing steganographic methods embedded secret messages in a webpage by embedding invisible characters [1], such as blanks and tabs, at the end of every line or by switching the uppercase-lowercase states of letters in tags [2].

Steganalysis mainly discusses the frangibility of the steganography algorithm, detects the stego-objects, and extracts the secret messages embedded in the objects. Up to now, there are some papers discussing about steganalysis of texts [3-7]. Edward proposed a universal steganalysis method based on language models and support vector machines to differentiate sentences modified by a lexical steganography algorithm from unmodified sentences [3]. Kot implemented a technique for the steganalysis of electronic text documents based on the similarity between the same characters or symbols [4]. Sun designed and implemented a system for text steganalysis based on the noise of the stego-text [5]. In reference [6] and reference [7], Sui proposed two steganalysis methods to detect hidden information in text file.

To the best of our knowledge, no studies exist on detecting hidden information in webpage. In this paper, we present a novel steganalytic approach called Tag-Mismatch analysis for detection of hidden information embedded in tags of a webpage.

2. Formal description of switching the uppercase-lowercase method

In this paper, we pay close attention to the embedding method that switching the uppercase-lowercase states of letters in tags. The formal description of the embedding method is given as following. Let set C={A(x)|x∈{26 uppercase letters in the English alphabet}}be the ASCII codes of the uppercase state letters, and c={A(x)|x∈{26 lowercase letters in the English alphabet}}is the set of the ASCII codes of the lowercase state letters, where the function A(•) obtain the ASCII code of an English alphabet. Define a function $f_1(x)=x-32$, where $x \in c$ and $f_1(x) \in C$, which changes the lowercase state of a letter in tags to uppercase. Define another function $f_0(x)=x$, where $x \in c$ and $f_0(x) \in c$, which does not change the written state a letter in tags. The embedding progress is given as follows. Let a sequence $m=\{0,1\}^n$ denotes the n watermark bits, and $m_i \in m$ denotes *i-th* watermark bit in m, where $0 \le i \le n-1$. If the watermark bit $m_i=0$, the written state of the corresponding letter in a tag is changed according to the function $f_0(x)$. If the watermark bit $m_i=1$, the written state of the corresponding letter in a tag is changed according to the function $f_1(x)$. The embedding process will not end until $i=n-1$.

3. Tag-Mismatch Analysis

In this section, we describe a new statistical attack called Tag-Mismatch analysis for the embedding method for a webpage which embeds secret messages in the uppercase-lowercase states of letters in tags. The new method is simple and fast and can estimate the secret message length with relatively high precision. Before we describe the principle of Tag-Mismatch Analysis, we analyze the impact of the embedding algorithm above on a cover webpage.

Let $T(x_1, x_2, ..., x_n)$ be a tag in a webpage. Where x_i is the *i-th* letter in the tag, $1 \le i \le n$ and n is the number of letters in the tag. In general, the tag shows stable structure that the written states of all letters are in the same. For all x_i in the tag, $A(x_i) \in C$ or $A(x_i) \in c$. The embedding process will disturb this structure. The written state of each letter is modified with probability $1/2$ according to the watermark bit. Some letters are uppercase and some are lowercase. Finally, when the maximal length message has been embedded in the cover

INFOSCALE 2007, June 6-8, Suzhou, China

DOI 10.4108/infoscale.2007.959

webpage (1 bit per letter), the structures of all tags have been disturbed.

Having presented our arguments in the paragraph above, we describe our new detection method. We call two tags as a tag-pair, which one tag is the start tag and the other is the end tag. For example the two tags <*body*> and <*/body*> is a tag-pair. A null-tag refers to a tag has neither a start tag nor an end tag. For example a tag <*br*> or <*/p*> is called a null-tag. The offset of a tag $T(x_1, x_2, ..., x_n)$ is calculated with the discrimination function F:

$$F(T(x_1, x_2, ..., x_n)) = \sum_{i=1}^{n-1} \left| A(x_{i+1}) - A(x_i) \right|,$$

$$A(x_i) \in c \cup C . \quad (1)$$

The purpose of the offset can quantify the structure of a tag. Next, the Tag-Mismatch is defined, which the offset of a start tag is unequal to the end tag's or the same null-tags have different offset after the secret message embedded. For example, <HtmL> and </hTml> is a tag-pair mismatch,
 and
 is a null-tag mismatch. We record the number of tag-pairs mismatch and null-tags mismatch to estimate the secret message length.

Theorem 1. Let s_1 be the number of tag-pairs mismatch of a webpage, s_2 is the number of the entire null-tags mismatch and k is the average number of letters in all of those tags in a webpage. Then, the estimated secret message length s:

$$s \approx k \times (2 \times s_1 + s_2) . \quad (2)$$

Let X be the embedded rate, which is the ratio of the secret message length to the file size, α is the decision threshold which determine the false positive and false negative of the detection algorithm. The resulting detection algorithm consists of the following:

(1): Determine the tag-pairs and null-tags in a webpage, and calculate the number of tag-pairs mismatch s_1 and the number of null-tags mismatch s_2;

(2): Calculate k and the webpage file size and denote as $filesize$;

(3): Calculate the estimated embedded rate $X \approx k \times (2 \times s_1 + s_2) / filesize$;

(4): if $X \geq \alpha$, output "a stego-webpage" and the estimated embedded rate X, or else, output "a regular webpage".

4. Conclusion

With the popular and easily-applied steganographic tools appearing on the Internet, it is possible that terrorists use tools to communicate with their accomplices, and transmit blueprints of the next terrorist attack. How to control and destroy the activity is an urgent problem to the government, security department, and army.

In this paper, we propose an algorithm to detect the hidden information embedded in letters of tag in a webpage.

Acknowledgments

This work is Supported by the National Grand Fundamental Research 973 Program of China under Grant No. 2006CB303000; the National Natural Science Foundation of China under Grant No.60373062, 60573045; the National Research Foundation for the Doctoral Program of Higher Education of China No.20050532007.

References

[1] QJ ZHAO, HT LU, XH JIANG. Web page watermarking for tamper-proof. Journal of Shanghai Jiaotong University (Science), 2005, 3 (E-10), pp: 280~284.

[2] XG SUI, H LUO. A new steganography method based on hypertext. In: Proc of Asia-Pacific Radio Science Conference, 2004, 181~184.

[3] CM TASKIRAN, U TOPKARA, M TOPKARA, et al. Attacks on lexical natural language steganography systems. Proceedings of the SPIE, 2006, pp: 97-105.

[4] J CHENG, CK ALEX, J LIU, et al. Steganalysis of data hiding in binary text images. In: Proc. of 2005 IEEE International Symposium on Circuits and Systems, Kobe, 2005, pp: 4405~4408.

[5] G LUO, XM SUN, YL LIU. Research on steganalysis of stegotext based on noise detecting. Journal of Hunan University (Natural Sciences), 2005, 32(6), pp: 181~184.

[6] XG SUI H LUO. A Steganalysis Method Based on the Distribution of Space Characters. in: Proc. of the 2006 International of Communications, Circuits and Systems, 2006, pp: 54~56.

[7] XG SUI, H LUO, ZL ZHU. A Steganalysis Method Based on the Distribution of First Letters of Words. in: Proc. of the 2006 International on Intelligent Information Hiding and Multimedia Signal, 2006, pp: 369~372.