Information Hiding: A New Approach in Text Steganography

L. Y. POR¹, B. Delina² Faculty of Computer Science and Information Technology, University of Malaya, 50603, Kuala Lumpur, MALAYSIA

Abstract:- Sending encrypted messages frequently will draw the attention of third parties, i.e. crackers and hackers, perhaps causing attempts to break and reveal the original messages. In a digital world, steganography is introduced to hide the existence of the communication by concealing a secret message inside another unsuspicious message. Steganography is often being used together with cryptography and offers an acceptable amount of privacy and security over the communication channel. This paper presents an overview of text steganography and a brief history of steganography along with various existing text-based steganography techniques. Highlighted are some of the problems inherent in text steganography as well as issues with existing solutions. A new approach is proposed in information hiding using inter-word spacing and inter-paragraph spacing as a hybrid method. Our method offers dynamic generated stego-text with six options of maximum capacity according to the length of the secret message. This paper also analyzed the significant drawbacks of each existing method and how our new approach could be recommended as a solution.

Key-Words: - Stegonography, Text Stegonography, Information Hiding, Security, Suspicious.

1 Introduction

Information hiding is a general term encompassing many subdisciplines. One of the most important subdisciplines is steganography [22] as shown in Fig.1.

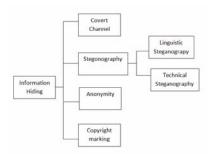


Fig.1: A Classification of Information Hiding techniques

Steganography, is derived from a work by Johannes Trithemus (1462-1516) entitled "Steganographia" and comes from the Greek ($\sigma\tau\epsilon\gamma\alpha\nu\delta-\varsigma$, $\gamma\rho\alpha\phi-\epsilon\iota\nu$) defined as "covered writing" [21]. It is an ancient art of hiding information in ways a message is hidden in an innocent-looking cover media so that will not arouse an eavesdropper's suspicion.

The goal of steganography is to transmit a message through some innocuous carrier i.e text, image, audio and video over a communication channel where the existence of the message is concealed. Based on Figure 1, steganography is one of the information hiding techniques and which can be categorized into linguistic steganography and technical steganography. Linguistic steganography defined by Chapman et al. [25] as "the art of using written natural language to conceal secret messages". A more specific definition by Krista Bennet in [21] explaining linguistic steganography as a medium which required not only the steganographic cover that is composed of natural language text, but the text itself can be either generated to have a cohesive linguistic structure, or the cover text that begin with natural language. On the other hand, technical steganography is explained as a carrier rather than a text which can be presented, as any other physical medium such as microdots and invisible inks.

During World War II, invisible inks offered a common form of invisible writing. With the invisible ink, a seemingly innocent letter could contain a very different message written between the lines. Therefore, the document text can conceal a hidden message through the use of null ciphers (unencrypted message), which perfectly camouflage the real message in an ordinary letter. Open-coded messages in which are plain text passages, but they are shown in only ordinary occurrence. The suspect communication can be detected by mail filters while "innocent" messages are allowed to flow through. In [13], there is an example on one of the most significant null cipher messages sent by a Nazi spy:

Apparently neutral's protest is thoroughly discounted and ignored. Isman hard hit. Blockade issue affects pretext for embargo on by-products, ejecting suets and vegetable oils [13].

By extracting the second letter from each word, this hidden message can be decoded as:

Pershing sails from NY June 1. [13]

The development of new digital technologies has given an opportunity to improve message detection that can pass more information and even be less conspicuous in transmission such as the microdots technology developed by the Germans. Microdots [16] uses microscopic shrink technique to hide pictures of text which can only be read using a microscope. German spies used them in many different ways like messages hidden in letters, on the face of watches and even on spotted ties as shown in [16].

The principle of information hiding is pioneered and documented in On the Criteria to be Used in Decomposing Systems Into Modules in 1972 [1], whereby Parnas designed a software system and each module's "interface of definition was chosen to reveal as little as possible about its inner workings". Many researchers are trying to carry out research by applying this concept in information hiding. There are three aspects in information hiding systems contend with each other: capacity, security and robustness [17]. Capacity refers to the amount of information that is able to be hidden in the medium, whereas security is important when a secret communication is kept to be secret and undetectable by eavesdroppers. Lastly, robustness can be explained as the amount of modification the stegomedium can withstand before an adversary can destroy hidden information.

Figure 2 shows the basic text steganography mechanism. Firstly, a secret message (or an embedded data) will be concealed in a cover-text by applying an embedding algorithm to produce a stego-text. The stego-text will then be transmitted by a communication channel, e.g. Internet or mobile device to a receiver. For recovering the secret which sent by the sender, the receiver needs to use a recovering algorithm which is parameterised by a stego-key to extract the secret message. A stego-key is used to control the hiding process so as to restrict detection and/or recovery of the embedded data to parties who know it [18].

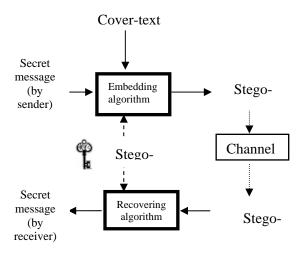


Fig. 2: The Mechanism of Text Steganography

Text steganography can be classified in three basic categories [21] - format-based, random and statistical generation and linguistic method.

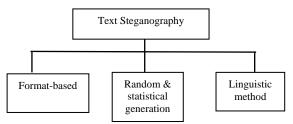


Fig.3: Three basic categories of text steganography

Format-based methods used physical text formatting of text as a place in which to hide information. Generally, this method modifies existing text in order to hide the steganographic text. Insertion of spaces, deliberate misspellings distributed throughout the text, resizing the fonts are some of the many format-based methods being used in text steganography. However, Bennett has stated that those format-based methods managed to trick most of the human eyes but it cannot trick once computer systems have been used.

Random and statistical generation is generating cover text according to the statistical properties. This method is based on character sequences and words sequences. The hiding of information within character sequences is embedding the information to be appeared in random sequence of characters. This sequence must appear to be random to anyone who intercepts the message. A second approach to character generation is to take the statistical properties of word-length and letter frequency in order to create "words" (without lexical value) which will appear to have the same statistical properties as actual words in a given language. The hiding of information within word sequences, the actual dictionary items can be used to encode one or more bits of information per word using a codebook of mappings between lexical items and bit sequences, or words themselves can encode the hidden information.

The final category is linguistic method which specifically considers the linguistic properties of generated and modified text, frequently uses linguistic structure as a place for hidden messages. In fact, steganographic data can be hidden within the syntactic structure itself.

In this paper, a new approach is proposed for text steganography by creating a hybrid method in utilising whitespaces between words and paragraphs in right-justification of text. This method can be an improvement of open space method [2] because it is not using a sole method of encoding data as what has been mentioned in [2]. By combining both methods which is inter-word spacing and interparagraph spacing into an embedding algorithm, a larger capacity for embedding hidden bits is provided.

The proposed scheme is inspired by Bender's open space method and a non-commercial used program namely SNOW by Matthew Kwan. Instead of using one method for every embedding mechanism, we propose to create a hybrid method in manipulation of whitespaces so that it is able to hide the secret bits in a dynamic generated cover text to produce a seemingly innocent stego-text.

This paper is divided into three sections. In the first section, it provides an overview for information hiding and text steganography and discuss about the new proposed approach. The second section analyzes this hybrid scheme by identifying the improvements that have been made over other approaches of text steganography. In the last section, we draw some conclusions and discuss the future work.

2 Inter-word Spacing and Interparagraph Spacing Approach

We are proposing a new approach on hiding

information in manipulation of whitespaces between words and paragraph as a hybrid method. The proposed method is able to provide more capacity for hiding more bits of data into a cover-text.

In the literature [2], hiding information within spaces seems to be potential as people hardly can know about the existence of the hidden bits. Bender et al. had shown that one space is interpreted as "0" whereas two spaces are interpreted as "1". This embedding scheme was applied in the space which appears between the words. The major drawback of Bender's method is that it requires a great deal of space to encode few bits. For example, a character is equivalent of 8 bits, and it requires approximately 8 inter-spaces to encode one character. Thus, this problem can be solved if we compress the character in the secret message from 8 bits to 3 bits or less in the proposed method. By combining with interparagraph in hiding the secret bits can effectively utilizing most of the whitespaces in a text document. In the embedding method, two spaces encode one bit per line, four encode two, eight encode three, etc., gradually increasing the amount of information we want to encode.

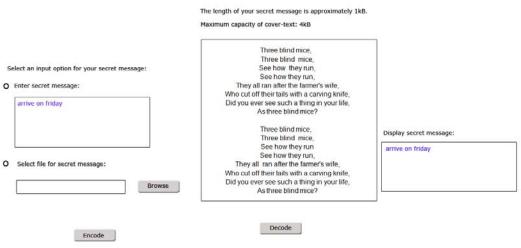
Currently, manipulation of whitespaces seems beneficial and has its potential in information hiding because whitespaces appear in a text documents more than the appearance of words. It is even an advantage when no one will know that a blank piece of document is actually vital secret information.

The cover-text will be dynamically generated according to the length of the secret message. The maximum capacity of hidden bits is determined by the system whether the length of the secret message can be accommodated in 4kB, 16kB, 32kb, 64kb, 128kb or 256kb. 4kB is used as a lower bound capacity among the options because the minimum input of text by a user is considered (e.g. a short secret message may contains only a few characters). By adapting the source of the generated text from any lyrics of the nursery rhymes, the stego-text will definitely present an innocuous and naive appearance. Besides, the chorus of the lyrics can be duplicated and reused to generate a longer covertext. There is even more advantageous in using center-alignment because the appearance of the spacing occurs naturally without purposely arising suspicion.

The significance of this research is by merging the two concepts and creating an algorithm for the data embedding system. The capacity of the cover-text is depending on the length of the secret message. Based on Figure 4, a user is required to enter the secret message as provided in the text field or by selecting a file that contains secret. The system will calculate the length of the message and generate a cover-text which can suitably encode the secret message. As illustrated in Figure 4, the length of the secret is approximately 1kB and can fit into the cover-text which contains 4kB of capacity. Then the secret message is entered and encoded. Eventually, the stego-text is generated as shown in Figure 5. When another user receives this message, the hidden bits can be retrieved by going through the recovering algorithm.

Size of the text (kB)	
Secret	Cover
<4	4
<16	16
<32	32
<64	64
<128	128
<256	256

Table 1: Cover-text is generated according to the
length of secret text



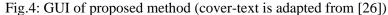




Fig.5: Original Cover Text (Left), Generated stego-text with hidden data (Right)

3 Analysis of Other Related Work

Peter Wayner [14] proposed a mimicry algorithm that aimed at text in his book Mimic Functions, Cryptologia XVI-3. His approach is to produce mimicked text that looks similar to the real structure of the original text. Peter Wayner used a set of grammatical rules to generate stegotext and the choice of each word determines how secret message bits are encoded. The grammatical rules are based on static grammar which means the grammars must be designed before the algorithm can be used. Apparently, the algorithm generates context-free structures. The system's user must design a grammar that he wishes the text to mimic. The quality of the resulting stegotext directly depends on the quality of the grammar. The grammar acts as the key for hiding data. Spammimic [12] is another website demonstrating mimicry. This implementation of Wayner's system employs a grammar that mimics the appearance of spam. Based on the author, the advantage of this application is a secret message can be encoded into something innocent looking, in a form of spam where nobody will notice there is a secret message being concealed. In fact, our proposed method keeps this advantage but changes the content of the cover-text which will be extracted from innocent nursery rhymes. Unlike Spammimic, it generates cover-text which contains of commercial purposes.

 Start
 →
 noun verb

 Noun
 →
 Fred || Barney

 Verb
 →
 went fishing where || went bowling where

 Where
 →
 in direction lowa. || in direction Minnesota.

 Direction
 →
 northerm || southerm

Fig.6: Example of Context-Free Grammar

Brassil et al. [3][4][5] gave the initial idea of document coding methods in his paper by proposing life-shift coding, word-shift coding and feature coding (character coding) to discourage illicit dissemination of document distributed by computer network. Line-shift coding is a method of altering a document by vertically shifting the locations of text lines to uniquely encode the document. Word-shift coding is a method to alter a document by horizontally shifting the locations of words within text lines to uniquely encode the document. Character coding or feature specific coding is a coding method that is applied only to the bitmap image of the document and can be examined for chosen character features, and those features are altered, or not altered, depending on the codeword. A document is marked in an indiscernible way by a codeword identifying the registered owner to whom the document is sent. If a document copy is found that is suspected to have been illicitly disseminated, that copy can be decoded and the registered owner identified.

Bender et al. [2] considered three methods of encoding data when they carried out a study on the techniques for data hiding in text. The three methods are open space methods that encode through manipulation of whitespace, syntactic methods that utilize punctuation and semantic methods that encode using manipulation of the synonymous words. Open space methods derive inter-sentence spacing, end-of-line spaces and inter-word spacing in justified text. Inter-sentence spacing method is to encode a binary message into a text by placing either one or two spaces at the end of each termination character. It encodes a "0" by adding a single space, encodes "1" by adding two spaces. This method works, the only disadvantage is inefficiency because it requires a large amount of text to encode a very few bits. One bit per sentence

equates to a data rate of approximately one bit per 160 bytes assuming sentences are on average two 80-character lines of text. This method fully depends on the structure of the text. End-of-line space method exploits whitespaces at the end of each line. Data encoded using a predetermined number of spaces at the end of each line. For example, two spaces will encode one bit, four spaces will encode two bits and eight spaces will encode three bits and so on. It works better than the inter-space method because increasing the number of spaces can hide more data. The third method of using whitespaces to encode data involves the right-justification of text can also be used to encode data within text files. Data is encoded by controlling where the extra spaces are placed. One space between words is interpreted as a "0". Two spaces are interpreted as a "1". It is found that not every inter-word space can be used as data due to the constraints upon justification. Bender et al. employed a Machester-like encoding method to determine which of the inter-word spaces represent hidden data bits and which is part of the original text. "01" is interpreted as "1" and "10" as "0". The bit strings "00" and "11" are null.

Information hiding in whitespaces has inspired Australian programmer, Matthew Kwan to invent a non-commercial use program – SNOW [23]. The encoding scheme used by SNOW relies on the fact that spaces and tabs (knows as whitespace), when invisibly appearing at the end of lines. Since the trailing spaces and tabs occasionally occur naturally, their existence should not be sufficient to immediately alert an observer who stumbles across them.

The SNOW program runs in two modes, first is message concealment and the second one is message extraction. In fact there are others optional modes such as built-in compression and encryption. Similarly, the extraction part will be reversing the process, i.e. extracting data from text, with optional decryption and decompression. Besides, this program is able to tell the user on how much data it can fit in the cover file.

■ infile3 - Notepad File Edk Format Yew Help Snow is a program for concealing messages in text files by appending tabs. ■ The data is concealed in the text file by appending sequences. Snow provide rudimentary compression.

Fig.7: An example of the contents in a cover file

Issuing the command to calculate the storage capacity can indicate the estimation of available bits so that we know whether the embedded data can fit within the cover file. Therefore, Figure 7 is an example that indicates that it can store between 68 bits to 74 bits of data when the storage calculation command is issued. The size of the embedded data can be fit into the cover text when it is less that the available storage capacity.

In the embedding scheme, 3 bits are usually coded in 8 columns of text and given the default line length is 80 characters, this allows 30 bits to be stored on empty lines. If a message could not fit into the available cover text, empty lines will be appended and used to accommodate the overflow. The disadvantage of SNOW is the embedded message could be easily destroyed by simply removing the extra spaces at the end of the line.

Chen Chao et al. expanded the idea of using TeX, a popular typesetting tool created by Donald E. Knuth [8]. This tool is particularly useful in generating scientific and technical documents with professional page layout. The Chinese researchers proposed that inter-word spaces in a text document can be modified to carry secret information. Thus the embedded data are extracted from the document image. With stego-encoding, space changes are scattered, leading to stealth improvement of the stego-text.

4 Conclusion

We have presented a new approach of text steganography method using inter-word and interparagraph spacing for hiding information. The unique feature about the method is to generate a cover-text dynamically by offering six options for user according to their length of the secret message.

The future work should be focused towards optimizing the robustness of the decoding algorithm. This is because the hidden data will be destroyed once the spaces are deleted by some word processing software. Besides that, it is important to improve the capacity of the embedded scheme by taking other compression method into consideration.

References:

[1] D. Parnas, "On the Criteria to Be Used in Decomposing Systems Into Modules", *Communication of the ACM*, vol. 15, no. 12, December 1972, pp. 1053-1058.

- [2] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding", *IBM syst. J.*, vol. 35, nos. 3 4, 1996, pp. 313 336.
- [3] J. Brassil, S. Low, N. F. Maxemchuk, and L. O'Garman., "Marking Text Features of Document Images to Deter Illicit Dissemination", IEEE, 1994, pp. 315-319.
- [4] J. Brassil, S. Low, N. F. Maxemchuk, and L. O'Garman, "Electronic Marking and Identification Techniques to Discourage Document Copying", IEEE, Oct. 1994, pp. 1278-1287.
- [5] J. Brassil, S. Low, N. F. Maxemchuk, and L. O'Garman. "Electronic Marking and Identification Techniques to Discourage Document Copying". *IEEE Journal on Selected Areas In Communications*, Vol. 13, Oct. 1995, pp. 1495-1504.
- [6] J. Brassil, S. Low, N. F. Maxemchuk, and L. O'Garman, "Copyright Protection for the Electronic Distribution of Text Documents", *Proceedings of IEEE*, Vol. 87, July 1999.
- [7] J. Brassil, S. Low, N. F. Maxemchuk, and L. O'Garman. "Document Marking and Identification using Both Line and Word Shifting," Proc. Infocom95. *IEEE CS Press*, Los Alamitos, Calif., 1995.
- [8] Chen Chao, Wang Shuozhong, Zhang Xinpeng. "Information Hiding In Text Using Typesetting Tools with Stego-Encoding", *ICICIC*, 2006.
- [9] P. Wayner, "Mimic functions", Cryptologia archive Volume XVI, Issue 3, July 1992, pp.193 – 214.
- [10] P. Wayner. "Strong Theoretical Steganography". Cryptologia, XIX(3), July 1995, pp. 285-299.
- [11] M. Chapman, G. Davida. "Hiding the Hidden: A Software System for Concealing Ciphertext as Innocuous Text". Master Thesis, Milwaukee: University of Wisconsin-Milwaukee, 1998.
- [12] "Spammimic", 2000. [Online] Available: http://www.spammimic.com [Accessed March 8 2008]
- [13] N. F. Johnson, S. Jajodia, "Exploring Steganography: Seeing the Unseen," *IEEE Computer*, February 1998, pp.26–34.
- [14] P. Wayner, Disappearing Cryptography: Being and Nothingness on the Net, Academic Press, Inc., 1996.
- [15] Richard Bergmair, "Towards Linguistic Steganography: A Systematic Investigation of Approaches", Systems, and Issues, *Technical Report*, Nov 2004.

- [16] "Spy Gadgets in World War II: Microdots", 2007. [Online]. Available: http://www.mi5.gov.uk/output/Page303.html [Accessed Feb. 15, 2008].
- [17] N. Provos, P. Honeyman, "Hide and Seek: An Introduction to Steganography", *The IEEE Computer Security*, 2003.
- [18] Fabien A. P. Petitcolas, Ross J. Anderson and Markus G. Kuhn. "Information Hiding – A Survey", Proceedings of the IEEE, special issue on protection of multimedia content, July 1999, pp. 1062 – 1078.
- [19] Bret Dunbar, "A Detailed Look At Steganographic Techniques and Their Use in Open-Systems Environment", SANS Institute, 2002.
- [20] B. Pfiztzmann, "Information Hiding Terminology." pp. 347{350, ISBN 3-540-61996-8, results of an informal plenary meeting and additional proposals, 1996.
- [21] K. Bennett, "Linguistic Steganography: Survey, Analysis, and Robustness Concerns for Hiding Information in Text", Purdue University, *CERIAS Tech. Report*, 2004.
- [22] Dr. Mohammed Al-Mualla and Prof. Hussain Al-Ahmad, "Information Hiding:

Steganography and Watermarking". [Online]. Available:

http://www.emirates.org/ieee/information_hidi ng.pdf [Accessed: March 12, 2008].

- [23] Matthew Kwan, "The SNOW Home Page", 1998.[Online].Available: http://www.darkside.com.au/snow/ [Accessed : March 12, 2008].
- [24] Sabu M. Thampi. "Information Hiding Techniques: A Tutorial Review", *ISTE-STTP* on Betwork Security & Cryptography, LBSCE 2004.
- [25] M. Chapman, G. Davida, and M. Rennhard, "A Practical and Effective Approach to Large-Scale Automated Linguistic Steganography", *Proceedings of the Information Security Conference*, October 2001, pp. 156-165.
- [26] [26] Nursery Rhymes lyrics and origins. [Online].Available:http://www.famousquotes. me.uk/nursery_rhymes/nursery_rhymes_index. htm [Accessed: March 31, 2008]
- [27] S. Katzenbeisser, Fabien A.P. Petitcolas, Information Techniques for Steganography and Digital Watermarking, Artech House, Boston, London, 2000.