Steganography in Text by Using MS Word Symbols

Ammar Odeh, Khaled Elleithy, Miad Faezipour
Computer Science & Engineering
University of Bridgeport
Bridgeport, CT06604, USA
aodeh@bridgeport.edu, elleithy@bridgeport.edu, mfaezipo@bridgeport.edu

Abstract – The massive amount of data transfer over internet raises different challenges such as channel types, transmission time and data security. In this paper, we present a novel secure algorithm to hide the data inside document files, where four symbols are used to embed the data inside the carrier file. The main process depends on a key to produce a symbol table and match the data to be hidden with the representative symbols. This method can be extended to any language and does not change the file format. In addition, the capacity ratio of the presented algorithm is high compared to other algorithms.

Keywords: Carrier file, Zero width chracter, Information Hiding,Stego Key.

I. INTRODUCTION

A. Background

Different strategies are used to protect transmitted data from eavesdroppers. Traditionally, cryptography is used, which is defined as data protection by converting a readable message into cipher form, preventing any middle users to read the original message [1]. Cryptography may face brute force attacks to analyze the encrypted message and conclude the secret information [2]. Alternatively, other approaches hide the secret message inside a public carrier file while manipulating it to insert the secret message [3]. In the regard, steganography attempts to avoid any suspicions by avoiding user file analysis. Thus everyone can read the carrier file but only authorized users can extract the hidden data.

Information hiding mainly consists of two branches, Digital Watermarking and Steganography. Steganography is an art of sending invisible messages. The word Steganography is derived from Greek words; “Stego” means “cover” and “Grapha” means “writing” [4]. Most historical stories about steganography are recorded back to 440 BC. One story says that Greek shave the prisoners head and wrote secret messages on his scalp. When his hair grew back, the king would send him to the other side where no one could read that message [5]. Other famous stories indicate that words were used to write secret messages and were covered by wax. The cover tablet was then sent to the receiver who would remove the wax and read the hidden message [6].

B.Motivation

Nowadays Steganography uses digital media to cover the secret message. Stego carrier files are classified into four categories as shown in Figure 1.

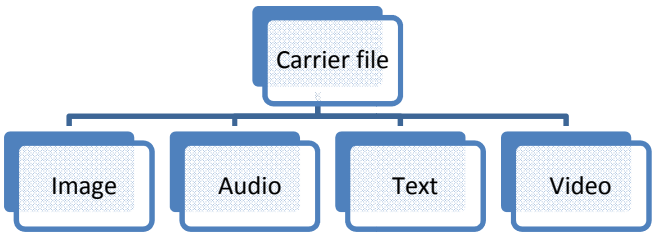

Figure 1.Steganography Carrier files

Image represents a popular carrier for secret messages, especially in RGB format. The image can be changed through the least significant bit in each pixel to substitute secret data on it [7]. Other algorithms use audio files as carriers by using frequency domain control limits for upper and lower frequencies. Video represents merging image and audio properties to hide data [8].

Text represents the hardest carrier file to hide data where it contains small redundant data compared to image and audio files [9, 10]. On the other hand, some text steganography algorithms depend on language properties which could restrict the algorithm applications to those specific languages. Text steganography can be divided into 3 categories:-

1. Format Based: - by changing format of the carrier file, we can pass our secret message. Format strategies depend on language properties. Thus, some algorithms can be applied to specific languages and cannot be applied to other languages. Some methods are generically enough to be applied to any text regardless of the carrier file language [11].

2. Random and Statistical Generation Methods: in this strategy, a cover text is generated depending on the statistical properties of the language. Probabilistic context-

978-1-4799-5233-5/14/$31.00 ©2014 IEEE

free grammar (PCFG) is the most common strategy used to produce the cover file. Other strategies employ word statistics such as letter frequency and word length [10, 12].

3. Linguistic Methods: these methods can be divided into two groups. The first group is the syntactic methods that depend on some punctuation signs to hide the data. The second category creates a synonym dictionary and replaces the interactive word by some carrier file word to pass the hidden bits [12, 13].

C. Main Contribution and Paper Organization

A novel text steganography algorithm is presented in this paper. The main idea is to use word symbols that enable us to hide 4 bits and avoid intruder suspicions.

The rest of this paper is organized as follows. In Section II we discuss previous text Steganography techniques. Symbols insertion algorithm is discussed in Section III. Section IV discusses and analyzes the presented algorithm. Finally, conclusions are offered in Section V.

II. PRIOR WORK

Word Synonym [14-16] is classified as one of the semantic steganography methods. This method focuses on replacing some of the words by their synonyms. In this technique, the hidden data will be transmitted without being suspicions to the attackers. However, in this method the data size is considered small compared to the other methods but it could change the sentence meaning.

Another method uses punctuations like (.) and (;) to represent hidden text. For example, “NY, CT, and NJ" is similar to "NY, CT and NJ" where the extra comma is used to represent 1 or to represent 0. The amount of hidden data in this method is very small in comparison to the amount of the cover media. Inconsistence use of punctuation might be noticeable from a Stegoanalysis perspective [16].

Line shifting involves vertically shifting a line a little to hide information to create a unique shape of the text. Unfortunately, line shifting can be detected by character recognition programs. Moreover, retyping the document will remove all the hidden data [14].

Two other Text Steganography algorithms were introduced in [17], where the space character was added after words and two bits were encoded. Depending on the number of word letters, and the number of space characters after that word, one of the values in the set {00, 01, 10, 11} would be passed. The second scenario suggests a new spacing method, where single spaces were used to pass 0, and double spaces were used to pass 1. The previous two methods have a problem since a word processor can highlight the additional spaces.

a [18]new method was introduced to hide data inside Telugu text by horizontally shifting inherent vowel signs. The main advantage of this method is that huge amount of data can be hidden inside the text file. Another algorithm was introduced in [19] by merging between three languages Chinese, Arabic, and English. In this approach, the authors create two tables; the first one is used for storing Arabic Diacritics and the other table is used for storing English letters. By translating Chinese text into English sentences, each English letter would correspond to two Arabic Diacritics. Then, the Arabic text is created which contained selected Diacritics.

III. PROPOSED ALGORITHM

The algorithm presented in this paper hides data inside a word file without inferring any changes in the file properties like file size, content and format. The proposed algorithm employs some invisible symbols to hide four bits between letters, which improves the hidden capacity ratio compared to other algorithms. Moreover, no changes in the word format or letter shape would be made. Furthermore, suggested algorithm avoids suspicions and any stegoanalyzer noticeability, which will in turn, improve the algorithm robustness. Inserting one of the table variation symbols after each letter enables us to hide four bits. Mainly, we use Right remark (200E), Left remark (200F), Zero width joiner (200D), and Zero width non-joiner (200C) by embedding any of these symbols to Steganography carrier file data.

TABLE I. SAMPLE OF HIDDEN BITS BY USING WORD SYMBOLS

Right Remark	Left Remark	ZWJ	ZWNJ	Hidden code
X	X	X	X	0000
X	X	X		0001
X		X		0101
X	X			0011
X				0111
X	X	X	X	1111
X	X		X	1101

In Table I we present some of the hidden codes when inserting the word symbols. For instance, if we insert all four symbols then the passing bits code is 0000. In this technique, different variations can be used to represent hidden bits for a total of 16 different codes.

Figure 2 represents the data hiding scenario/steps when using three inputs; the carrier file, hidden data, and Stego key. The main purpose of Stego key is to change the symbols bit representation. In other words, 0 represents bit absence while the other state represents a 1. In the next step, a symbols table is created depending on the Stego key; we

insert four bits from the hidden data after each letter in carrier file.

The capacity of the carrier file is computed as follows:

$$Capacity\ of\ carrier\ file = Number\ of\ letters \times 4 \quad (1)$$

So the hidden capacity of our algorithm is:

$$Capacity\ Ratio = (Number\ of\ letters \times 4) / carrier\ file\ size \quad (2)$$

The receiver can extract the hidden data by reading the carrier file and using the Stego key to build the symbols table. Reading the symbols after each letter and matching them with the symbols table would enable the receiver to extract the hidden data.

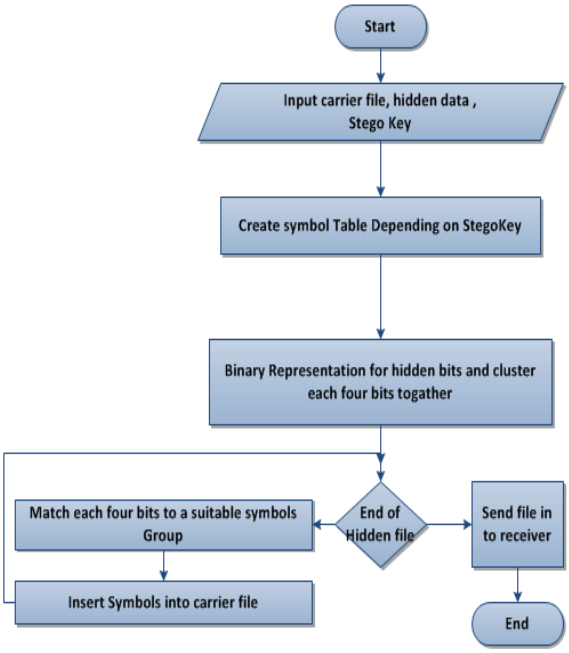

Figure 2. Data Hiding Algorithm

IV. Analytical Discussion

Table II shows the capacity for the new algorithm and two other algorithms which are applied to different visited web sites. Text steganography was used in those pages to evaluate the hidden capacity. Figure 3 shows a comparison histogram for the three algorithms.

The algorithm has many advantages over other algorithms. For example, this algorithm can be applied to any language regardless of if it is Unicode or ASCII codes, where other algorithms such as [11, 18, 19] can be applied to only some Unicode languages. Moreover, there is no need for special software or equipment to hide the data and extract it. The algorithm does not change the file format since the used symbols do not affect the format of the letters. Consequently, this algorithm improves transparency feature which is one of key Steganography objectives.

V. Conclusion

Different algorithms have been presented in literature to hide data inside text files. Some of these methods were designed to be applied to specific languages, while others are generic and can be applied to any language. In this paper, we introduced a novel algorithm that can be used to hide data inside document files of any language by using word symbols. Our technique employed Remarks (Right Remark, Left Remark, ZWJ, and ZWNJ) symbols which can be used in any language and at any position in the words. These scenarios enable the user to pass 4 bits between any two letters. In addition, the algorithm has been enhanced by using a Stego key to create symbols table representation

.

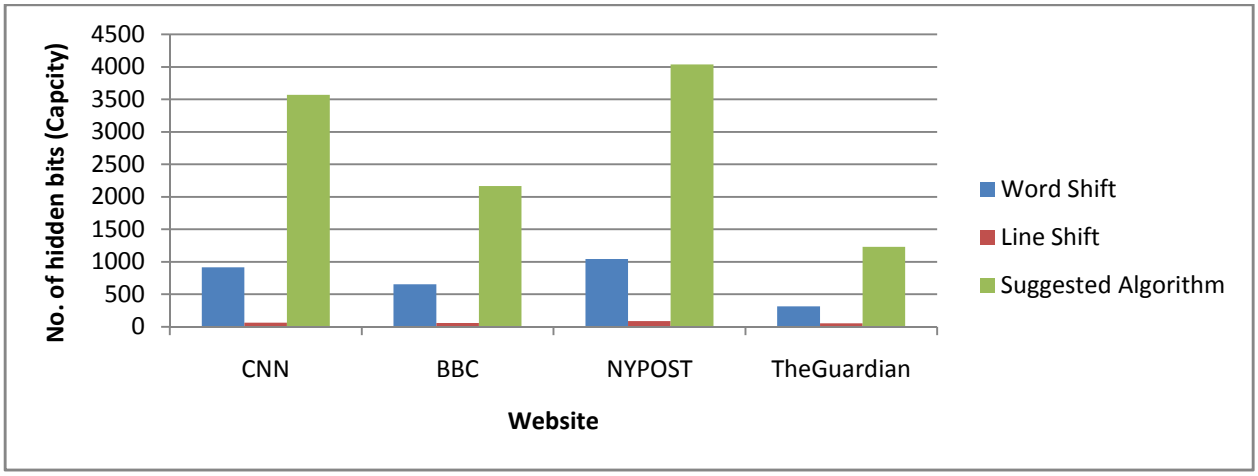

Figure 3. Comparison between three algorithms

TABLE II. REPRESENT THE SIMULATION RESULT OF FILE SIZE AND NUMBER OF BITS CAN BE INSERT IN TO CARRIES WEB PAGES

	Web site	Size (K.B)	Number of lines	Number of words	Number of letters	Our Algorithm	Line shift algorithm	Word shift algorithm
1	www.cnn.com	19.8	74	763	4592	928	4	39
2	www.bbc.com	19.3	67	749	4065	842	3	39
3	www.nypost.com	19.8	48	634	3532	714	2	32
4	www.guardian.co.uk	21	64	935	5625	1071	3	45
5	www.ctpost.com	20.5	51	640	3652	713	2	31

REFERENCES

[1] M. Shirali-Shahreza, "Pseudo-space Persian/Arabic text steganography," in *Computers and Communications, 2008. ISCC 2008. IEEE Symposium on*, 2008, pp. 864-868.

[2] W. A. Arbaugh, N. Shankar, Y. J. Wan, and K. Zhang, "Your 80211 wireless network has no clothes," *Wireless Communications, IEEE,* vol. 9, pp. 44-51, 2002.

[3] M. Shirali-Shahreza and S. Shirali-Shahreza, "Steganography in TeX documents," in *Intelligent System and Knowledge Engineering, 2008. ISKE 2008. 3rd International Conference on*, 2008, pp. 1363-1366.

[4] R. Krenn, "Steganography and steganalysis," *Retrieved September,* vol. 8, p. 2007, 2004.

[5] J. Silman, "Steganography and steganalysis: an overview," *SANS Institute,* vol. 3, pp. 61-76, 2001.

[6] B. Dunbar, "A detailed look at Steganographic Techniques and their use in an Open-Systems Environment," *SANS Institute,* 2002.

[7] N. F. Johnson and S. Jajodia, "Exploring steganography: Seeing the unseen," *IEEE computer,* vol. 31, pp. 26-34, 1998.

[8] F. Djebbar, B. Ayad, H. Hamam, and K. Abed-Meraim, "A view on latest audio steganography techniques," in *Innovations in Information Technology (IIT), 2011 International Conference on*, 2011, pp. 409-414.

[9] V. Potdar and E. Chang, "Visibly Invisible: Ciphertext as a Steganographic Carrier," in *Proceedings of the 4th International Network Conference (INC2004)*, 2004, pp. 385-391.

[10] S. Bhattacharyya, I. Banerjee, and G. Sanyal, "A novel approach of secure text based steganography model using word mapping method (WMM)," *Journal on "International Journal of Computer and Information Engineering,* vol. 4, p. 2, 2010.

[11] R. Prasad and K. Alla, "A new approach to Telugu text steganography," in *Wireless Technology and Applications (ISWTA), 2011 IEEE Symposium on*, 2011, pp. 60-65.

[12] V. N. Rao and D. D. Shulman, *Classical Telugu poetry: an anthology*: University of California Press, 2002.

[13] K. Bennett, "Linguistic steganography: Survey, analysis, and robustness concerns for hiding information in text," *CERIAS Technical Report 3, Purdue University,* pp. 1-30, 2004.

[14] M. H. Shirali-Shahreza and M. Shirali-Shahreza, "A new approach to Persian/Arabic text steganography," in *Computer and Information Science, 2006 and 2006 1st IEEE/ACIS International Workshop on Component-Based Software Engineering, Software Architecture and Reuse. ICIS-COMSAR 2006. 5th IEEE/ACIS International Conference on*, 2006, pp. 310-315.

[15] M. Nosrati, R. Karimi, and M. Hariri, "An introduction to steganography methods," *World Applied Programming,* vol. 1, pp. 191-195, 2011.

[16] M. H. Shirali-Shahreza and M. Shirali-Shahreza, "Text steganography in chat," in *Internet, 2007. ICI 2007. 3rd IEEE/IFIP International Conference in Central Asia on*, 2007, pp. 1-5.

[17] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," *IBM systems journal,* vol. 35, pp. 313-336, 1996.

[18] S. ALAMETI, A. POTHALAIAH, and A. BABU, "A New Approach to Telugu Text Steganography by Shifting Inherent Vowel," *International Journal of Engineering Science and Technology,* vol. 2, pp. 7203-7214, 2010.

[19] A. C. Shakir, G. Xuemai, and J. Min, "Chinese Language Steganography using the Arabic Diacritics as a Covered Media," *International Journal of Computer Applications IJCA,* vol. 11, pp. 24-28, 2010.

Ammar Odeh is a PhD. Student in University of Bridgeport. He earned the M.S. degree in Computer Science College of King Abdullah II School for Information Technology (KASIT) at the University of Jordan in Dec. 2005 and the B.Sc. in Computer Science from the Hashemite University. He has worked as a Lab Supervisor in Philadelphia University (Jordan) and Lecturer in Philadelphia University for the ICDL courses and as technical support for online examinations for two years. He served as a Lecturer at the IT, (ACS,CIS ,CS) Department of Philadelphia University in Jordan,

and also worked at the Ministry of Higher Education (Oman, Sur College of Applied Science) for two years. Ammar joined the University of Bridgeport as a PhD student of Computer Science and Engineering in August 2011. His area of concentration is reverse software engineering, computer security, and wireless networks. Specifically, he is working on the enhancement of computer security for data transmission over wireless networks. He is also actively involved in academic community, outreach activities and student recruiting and advising.

Dr. Khaled Elleithy is the Associate Dean for Graduate Studies in the School of Engineering at the University of Bridgeport. He has research interests are in the areas of network security, mobile communications, and formal approaches for design and verification. He has published more than two hundred fifty research papers in international journals and conferences in his areas of expertise. Dr. Elleithy is the co-chair of the International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE). CISSE is the first Engineering/Computing and Systems Research e-Conference in the world to be completely conducted online in real-time via the internet and was successfully running for four years. Dr. Elleithy is the editor or co-editor of 10 books published by Springer for advances on Innovations and Advanced Techniques in Systems, Computing Sciences and Software.

Dr. Miad Faezipour is an Assistant Professor in the Computer Science and Engineering program at the University of Bridgeport and the director of the D-BEST Lab since July 2011. Prior to joining UB, she has been a Post-Doctoral Research Associate at the University of Texas at Dallas collaborating with the Center for Integrated Circuits and Systems and the Quality of Life Technology laboratories. She received the B.Sc. in Electrical Engineering from the University of Tehran, Tehran, Iran and the M.Sc. and Ph.D. in Electrical Engineering from the University of Texas at Dallas. Her research interests lie in the broad area of biomedical signal processing and behavior analysis techniques, high-speed packet processing architectures, and digital/embedded systems. Dr. Faezipour is a member of IEEE and IEEE women in engineering.