International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-3, Issue-11)

Research Article | **November 2014**

Different Approaches to Text Steganography: A Comparison

Vidhya Saraswathi, Mrs. Sumathy Kingslin, (Associate Professor)
Computer Science, Quaid-Eh Millath College
Tamil Nadu, India

Abstract—

Steganography is the art of hiding the fact that communication is taking place, by hiding information in other information. In this paper, we discuss different approaches of text steganography. There exists a large variety of text Steganography techniques, some are more complex than others and all of them have respective strong and weak points. Different applications have different requirements of the text steganography technique used. We describe a common application that needs both absolute invisibility and large secret data get hidden. We analyze some basic approaches of text based steganography and compare it according their performance based on the advantages and dis advantages.

Keywords— data hiding, Hiding Data in Paragraphs, Text -Steganography, security.

I. INTRODUCTION TO STEGANOGRAPHY

Although people have hidden secrets in plain sight—now called steganography—throughout the ages, the recent growth in computational power and technology has propelled it to the forefront of today's security techniques. Steganography is the art and science of hiding communication; a steganographic system thus embeds hidden content in unremarkable cover media so as not to arouse an eaves dropper's suspicion. In the past, people used hidden tattoos or invisible ink to convey steganographic content. Today, computer and network technologies provide easy-to-use communication channels for steganography. Essentially, the information-hiding process in a steganographic system starts by identifying a cover medium's redundant bits (those that can be modified without destroying that medium's integrity) [7].
Steganography can be split into two types:
Fragile: This steganography involves embedding information into a file which is destroyed if the file is modified.
Robust: Robust marking aims to embed information into a file which cannot easily be destroyed. [16]
Steganography is derived from the Greek for covered writing and essentially means "to hide in plain sight". 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. Steganography techniques can be categorized into linguistic steganography and technical steganography. Linguistic steganography defined by Chapman et al. [11] as "the art of using written natural language to conceal secret messages". 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. [6]

II. TEXT STEGANOGRAPHY

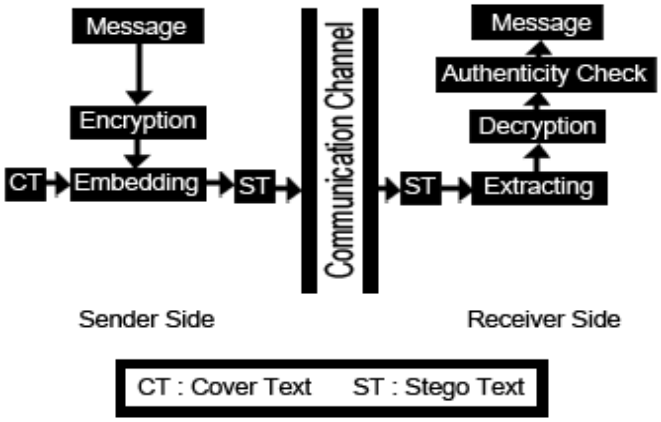
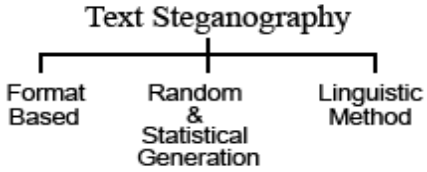


Fig: Text Steganography Model

The text steganography is a method of using written natural language to conceal a secret message as defined by Chapman et al [12]. A message is embedded in a text (cover text) through an embedding algorithm. The resulting stego text is transmitted over a channel to the receiver where it is processed by the extraction algorithm with the help of a secret key. During transmission the stego text, it can be monitored by unauthenticated viewers who will only notice the transmission of an innocuous-text without discovering the existence of the hidden message in it. Text steganography can be broadly classified into three types- format-based, random and statistical generations and Linguistic method [13] [10] [4] [1].



i. **Format-based methods [15]:** This method uses the physical formatting of text as a space in which to hide information. Insertion of spaces or non-displayed characters, careful errors tinny throughout the text and resizing of fonts are some of the many format-based methods used in text steganography. Some of these methods, such as deliberate misspellings and space insertion, might fool human readers who ignore occasional misspellings, but can often be easily detected by a computer.
ii. **Random and statistical generation method [15]:** Random and statistical generation methods are used to generate cover-text automatically according to the statistical properties of language. These methods use example grammars to produce cover-text in a certain natural language. A probabilistic context-free grammar (PCFG) is a commonly used language model where each transformation rule of a context- free grammar has a probability associated with it [15].
iii. **Linguistic methods [15]:** Linguistic steganography specifically considers the linguistic properties of generated and modified text, and in many cases, uses linguistic structure as the space in which messages are hidden [2].

III. STATE-OF-THE-ART IN TEXT STEGANOGRAPHY

Hiding information in plain text can be done in many different ways. This section sheds the light on the various techniques used in text steganography including line shifting[16], word shifting[16], hiding using whitespace[14], semantic-based hiding[5], syntax based hiding[7], abbreviation-based hiding[9] techniques and hiding data in paragraphs [13].

3.1 Format-based methods

A format-based text steganography method is open space method [21]. In this method extra white spaces are added into the text to hide information. A single space is interpreted as"0" and two consecutive spaces are interpreted as"1". Another two format-based methods are **word shifting and line shifting**.

3.1.1 Word shifting method: In this method, by shifting words horizontally and by changing distance between words, information is hidden in the text. This method is acceptable for texts where the distance between words is varying. This method can be identified less, because change of distance between words to fill a line is quite common.

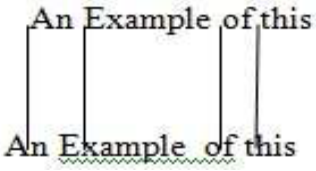


Fig: Word Shifting

But if somebody was aware of the algorithm of distances, he can compare the present text with the algorithm and extract the hidden information by using the difference. The text image can be also closely studied to identify the changed distances. Although this method is very time consuming, there is a high probability of finding information hidden in the text. Retyping of the text or using OCR programs destroys the hidden information [2] [15] [17].

3.1.2 Line shifting: In this method, the lines of the text are vertically shifted to some degree (for example, each line is shifted 1/300 inch up or down) and information are hidden by creating a unique shape of the text. This method is suitable for printed texts.

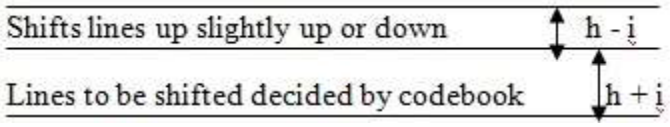


Fig: Line Shifting

However, in this method, the distances can be observed by using special instruments of distance assessment and necessary changes can be introduced to destroy the hidden information. Also if the text is retyped or if character recognition programs (OCR) are used, the hidden information would get destroyed. This method hides information by shifting the text lines to some degree to represent binary bits of secret information [8] [11] [17].

3.1.3 Feature coding: In feature coding method, some of the features of the text are altered. For example, the end part of some characters such as h, d, b or so on, are elongated or shortened a little thereby hiding information in the text. In this method, a large volume of information can be hidden in the text without making the reader aware of the existence of such information in the text. By placing characters in a fixed shape, the information is lost. Retying the text or using OCR program destroys the hidden information [17].

3.2Random and statistical generation:

Random and statistical generation is generating cover text according to the statistical properties. This method is based on character sequences and words sequences [1].

3.2.1Character sequences**:** The hiding of information within character sequences is character based information that is available and transmitted over networks. One approach to text [6] steganography might hide information in what appears to be a random sequence of characters. Of course, to both the person sending and receiving the message, this sequence is far from random, but it must appear to be random to anyone who intercepts the message.

3.2.2Word sequences: To solve the problem of detection of non-lexical sequences, actual dictionary items can be used to encode one or more bits of information per word. This might involve a code-book of mappings between lexical items and bit sequences, or the words themselves (length, letters, etc.) might encode the hidden information. [15]

3.2.3Statistical generation of sequences: text mimicking: In addition to using the statistical frequency of letters or words in order to generate cover text, Wayner proposed a clever method of generating text which can be fairly convincing lexically and syntactically (and often semantically).One string is picked at random to start the steganographic text, and the next letter in the text is chosen by looking at a window of the last $n - 1$ characters in the steganographic text and finding all of the strings in the table which start with those characters. The next letter is chosen from the last letter of these strings by using the data structures from the Huffman compression algorithm in reverse; the statistical frequency of all of the possible next letters that end the strings that start with the desired $n - 1$ characters is used to generate a tree which uses the frequency of each of the selected strings to organize the last letters into an encoding tree. [8]

3.3 Linguistic methods:

Linguistic method is a combination of syntax and semantics methods. Syntactic steganalysis is to ensure that structures are syntactically correct. In Semantic Method you can assign the value to synonyms and data can be encoded into actual words of text.

3.3.1 Syntactic method: By placing some punctuation marks such as full stop (.) and comma (,) in proper places, one can hide information in a text file. This method requires identifying proper places for putting punctuation marks. The amount of information to hide in this method is trivial.

3.3.2 Semantic method**:** This method uses the synonym of certain words thereby hiding information in the text. The synonym substitution may represent a single or multiple bit combination for the secret information. However, this method may alter the meaning of the text [3].

IV. COMPARISON OF TECHNIQUES IN TEXT STEGANOGRAPHY

The various techniques used in text steganography including their pros and cons as follows:

TABLE-I

Text Steganography methods	**Advantage**	**Dis-Advantage**
Line shifting	This method is suitable only for printed text.	When OCR (character recognition program) applied the hidden information gets destroyed.
Word shifting	Word shifting method identify less because of change of distance between words to fill line is quite common.	The algorithm that related to word shifted distance, easily can get hidden data.
Syntactic Method	The amount of information to hidden the method is trivial.	Smart reader can find hidden data easily.
Semantic-based hiding	This method is better than above methods, syntactic, line shifting and word shifting because that cannot detect by retyping or using OCR programs.	Smart reader which has huge knowledge of words their synonyms or antonyms can discover it.
Abbreviation based hiding	This method is because it's a kind of any abbreviation present and we built also.	It is limited only for small data means out of large data. Only small part of data can be hidden.
Hiding Data using white spaces	One way of hiding data in text is to use white space. Due to the fact that in practically all text editors, extra white space at the end of lines is skipped over, it won't be noticed by the casual viewer.	In a large piece of text, this can result in enough room to hide a few lines of text or some secret codes.
Hiding Data in Paragraphs	The approach works by hiding a message using start and end letter of the words of a cover file. A word having same start and end letter is skipped. Since no change is made to the cover, the cover file and its corresponding stego file are exactly the same.	The volume of data hiding in the paragraph would be very less. The capacity of hiding the large volume of data leads to the challenge

English Text Based Steganography with Indian Root	This approach gives flexibility and freedom from the point view of the sentence construction	This approach can be of increase in computational complexity.

V. CONCLUSION

Among all the Text Steganography methods, each method has respective capability to hide data in text. By using line shifting method, we can hide huge amount of data, but line shifting method only capable for printed text because in this method, other than printed text character reorganization program(OCR) is used and hidden information get destroyed. Word shifting method is quite useful to hide data. In this method key term is algorithm made & used for word shifting. If this algorithm found by someone else than also security destroyed. Syntactic method used in wars to send very important information and hide very small amount of data. Like, we can use (.) and (,) in a poem and hide data in (0) as (.) and (1) as (,) and we can get data by extracted it, But if smart detector examine then hidden data find out and security get break. Semantic method is efficient and its security is higher than previous method because there are no such easy method to detect hidden data present but reader know huge knowledge of antonym and synonym, detect it. In abbreviation method, we can use abbreviation that already made & we built it on our own. It is a safe method as compare to above methods of data out of large data get hidden, because we cannot make abbreviation of all data. It takes large time also. Hiding Data in Paragraphs being the most efficient can be used to transmit confidential data securely over the Internet. Also, the concept of this approach can be applied to hide large amount of data with less amount of time.

The last approach (English text based steganography with Indian root) is one among the techniques which can be used for does not give any separate importance to vowels and consonants providing better flexibility in hiding data in case of English language.

REFERENCES

[1] Youssef Bassil, LACSC – Lebanese Association for Computational Sciences Registered under No. 957, 2011,Beirut, Lebanon, "A Generation-based Text Steganography Method using SQL Queries", International Journal of Computer Applications (0975 – 8887) Volume 57– No.12, November 2012.

[2] Arvind Kumar, Km. Pooja, "Steganography- A Data Hiding Technique", International Journal of Computer Applications (0975 – 8887), Volume 9– No.7, November 2010.

[3] Souvik Bhattacharyya , Indradip Banerjee and Gautam Sanyal, "A Novel Approach of Secure Text Based Steganography Model using Word Mapping Method(WMM)", International Journal of Computer and Information Engineering 4:2 2010.

[4] Souvik Bhattacharyya, and Gautam Sanyal. Study of secure steganography model. In Proceedings of International Conference on AdvancedComputing and Communication Technologies (ICACCT-2008), Panipath,India, 2008.

[5] M. H. Shirali-Shahreza, M. Shirali-Shahreza, 2008. ANew Synonym Text Steganography, IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing.

[6] W. Bender, D. Gruel, N. Morimoto, A. Lu, 1996.Techniques for data hiding IBM Systems Journal, vol. 35, no 3, pp. 313-336.

[7] Adnan Gutub and ManalFattani, 2007.A Novel Arabic Text Steganography Method Using Letter Points and Extensions, World Academy of Science, Engineering and Technology, Vol. 27.

[8] SudeepGhosh, 2007. StegHTML: A message hiding mechanism inHTML tags.

[9] M.Shirali-Shahreza,M.H.Shirali-Shahreza, 2007. Text Steganography in Chat, 3rd IEEE/IFIP International Conference in Central Asia on Internet.

[10] JHP Eloff. T Mrkel and MS Olivier. An overview of image steganography. In Proceedings of the fifth annual Information Security South Africa Conference., 2005.

[11] 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.

[12] G. Davida M. Chapman and M. Rennhard. A practical and effective approach to large-scale automated linguistic steganography. In Proceedings of the Information Security Conference, pages 156–165, October 2001.

[13] N.F.Johnson. and S. Jajodia. Steganography: seeing the unseen. IEEE Computer, 16:26–34, 1998.

[14] 14W. Bender, D. Gruhl, N. Morimoto, A. Lu,1996.Techniques for data hiding IBM Systems Journal, vol. 35, no 3, pp. 313-336.

[15] S. Low, N.Maxemchuk, J.Brassil, L. O'Gorman, 1995.Document marking and identification using both line and word shifting, Proceedings of the 14th Annual Joint Conference of the IEEE Computer and Communications Societies, INFOCOM 95.

[16] Niels provos and peter honeyman, University of Michigan, "Hide and Seek: An Introduction to Steganography", IEEE Security & Privacy.

[17] Prem Singh, Rajat Chaudhary and Ambika Agarwal. "A Novel Approach of Text Steganography based on null spaces", IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661 Volume 3, Issue 4, PP 11-17.

[18] Souvik Roya,*, P.Venkateswaran, "A Text based Steganography Technique with Indian Root", International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013, Procedia Technology 10 (2013) 167 – 171.

