The International Journal Research Publication's
ISSN:2251 1563
Research Journal of Science & IT Management

A Review on Text Based Steganography

Prateek Dobriyal , Joy Yadav , Jatin Jain
UnderGraduate Student , Dept. of CSE, H.M.R Institute Of Technology,
New Delhi, India

ABSTRACT
With the rapid growth of networking mechanisms, where large amount of data can be transferred between users over different media, the necessity of secure systems to maintain data privacy increases significantly. Steganography is a significant means by which secret information is embedded into cover data imperceptibly for transmission, so that information cannot be easily aware by others.
In this paper we review different Text Based Steganography already in-use for hiding the data from the research done till now in this field to give a brief overview about it and provide analyse of different methods used today.
There is a lot of advantages using Text Steganography over other techniques as it is low in redundancy and related to different natural language rules which leads to limit manipulation of text, so they are both great challenges to conceal message in text properly and to detect such concealment.

1.INTRODUCTION

Though the security is nothing new, the way that security has become an essential part of our life is unprecedented. As the people are getting connected to one another across the globe through technology and this technology is becoming an important part of our life. With these technology comes responsibility to protect the users from that technology from any unwanted or unfavourable third party. Today we use security in technological sphere starting from pass codes that we use to enter our own highly secure houses ,to retina-scanning technology that identifies us as we enter our office buildings, to scanners in airports, we have made security an essential part of our life. We are also surrounded by a world of secret communication, where people of all types are transmitting information.
Though this scheme that makes secret communication possible is not new. Julius Caesar used cryptography to encode political directives[10]. During World War 2, German spies used them in many different ways like messages hidden in letters, on the face of watches and even on spotted ties. Microdots uses microscopic shrink technique to hide pictures of text which can only be read using a microscope[11]. Though these techniques are old but its use in Internet, high-speed computer and transmission technology makes this a unique technique in history for covert communication.
Steganography is derived from Greek words Steganous meaning "covered" and graphy meaning "writing". So it is known as "covered writing". Steganography is a technique which is used to hide the message and prevent the detection of hidden message. From our review we found - Unlike Cryptography, which merely ensures the confidentiality of the message content , Steganography adds another layer of secrecy by adding another layer by keeping confidential even the fact that secret communication takes place. Steganography is an ancient art of embedding private messages in seemingly innocuous messages in such a way that prevents the detection of the secret messages by a third party. In other words, steganography means establishing covert channels. A covert channel is a secret communication channel used for transmitting information. Figure-1 shows the process of schematic diagram of a typical steganography system.

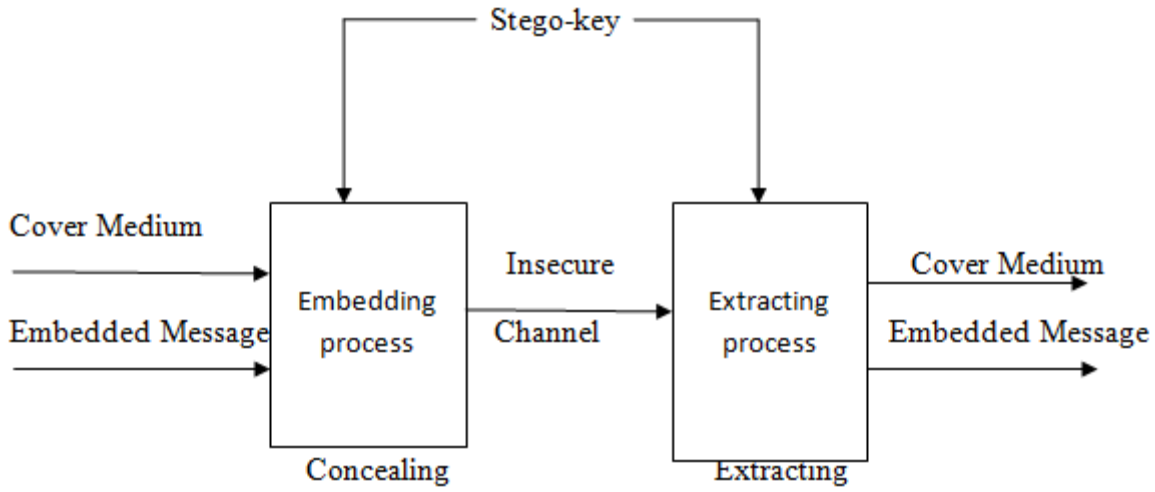


Fig-1: The General Steganography System

Modern steganography is generally understood to deal with electronic media rather than physical objects and texts. In steganography, the text to be concealed is called embedded data. An innocuous medium, such as text, image, audio, or video file; which is used to hide embedded data is called cover. The key (optional) used in embedding process is called stego-key. A stego-key is used to control the hiding process so as to restrict detection and/or recovery of embedded data to the parties who know it. The stego object is an object we get after hiding the embedded data in a cover medium.

2. TEXT STEGANOGRAPHY

Steganography can be classified into image, text, audio and video steganography depending on the cover media used to embed secret data. Text steganography can involve anything from changing the formatting of an existing text, to changing words within a text, to generating random character sequences or using context-free grammars to generate readable texts.

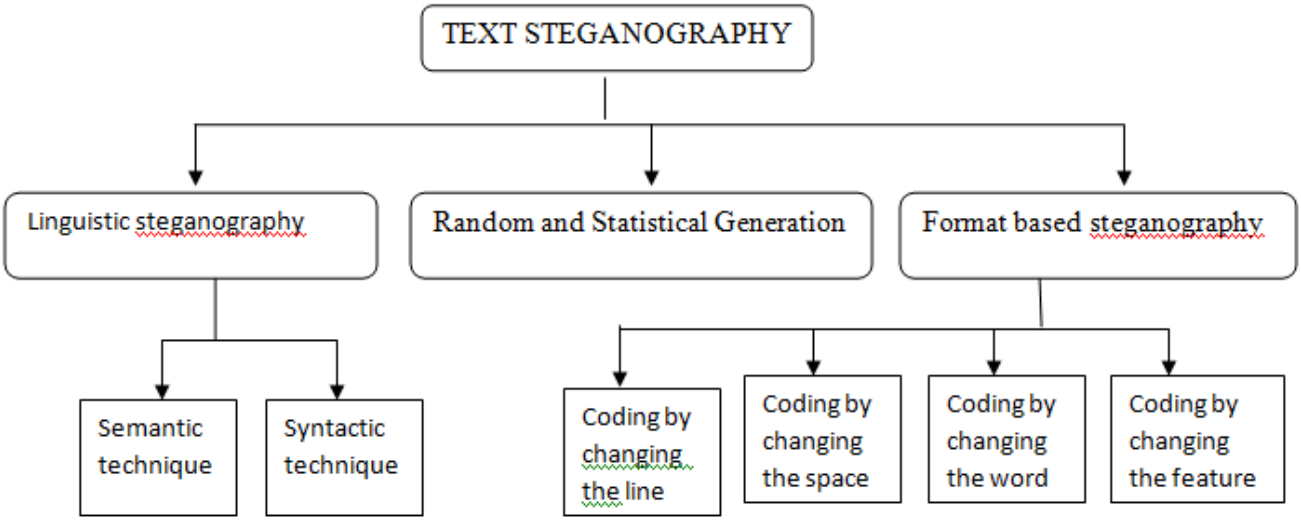
Text steganography is the most difficult kind of steganography; this is due to lack of redundant information in a text file as compared with a picture or a sound file [3]. The structure of a text document is identical with what we observe, while in other types of documents such as pictures, the structure of document is different from what we observe. Therefore, in such documents, we can hide information by introducing changes in the structure of the document without making a notable change in the concerned output. Unperceivable changes can be made to an image or an audio file, but, in text files, even an additional letter or punctuation can be marked by a casual reader. Storing text file require less memory and its faster as well as easier communication makes it preferable to other types of steganographic methods

One of the first methods of text steganography is the use of a system that extracts message automatically or hides it in nth characters or hides coded message by changing distance after or between characters. Of course, all of these methods had a low security and some of them have attracted hackers' attention and or the coded message was removed by rewriting the original text.

Another published method of steganography was to put data as noise in a covering media or change the available text format and style. However, such changes attracted hackers' attention and facilitated detection of coded text. With time and advancement of steganography, the use of covering media accompanied by a coded key was prevalent, which had a higher degree of security.

Despite great difficulties of this work, enormous efforts were made to design text steganography methods in English, Chinese, Persian, Arabic, and so forth that can generally be divided into three main categories as shown in figure 1.

The format-based method of steganography includes changing the words, lines, spaces and features of the original text and the linguistic methods is divided into syntactic and semantic classes.
Text steganography can be broadly classified into three types- Linguistic -based, random and statistical generation and format based.



2.1 FORMAT BASED
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[11].

2.2 RANDOM & 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. 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.

2.3 LINGUISTIC METHOD
Linguistic method which specifically considers the linguistic properties of generated and modified text, frequently uses linguistic structure as a place for hidden messages. It is the combination of syntax and semantics. Syntactic steganalysis ensure the correct structure as the text is generated from grammar. In semantic method value is assigned to synonyms and data can be encoded to the actual word of text[4].

2.4. EXISTING APPROACHES
In this sub-section, we present some of the popular approaches of text steganography.

2.4.1. LINE SHIFT
In this method, secret message is hidden by vertically shifting the text lines to some degree. A line marked has two unmarked control lines one on either side of it for detecting the direction of movement of the marked line . To hide bit 0, a line is shifted up and to hide bit 1, the line is shifted down . Determination of whether the line has been shifted up or down is done by measuring the distance of the centroid of marked line and its control lines. If the text is retyped or if a character recognition program (OCR) is used, the hidden information would get destroyed. Also, the distances can be observed by using special instruments of distance assessment.

2.4.2. WORD SHIFT

In this method, secret message is hidden by shifting the words horizontally, i.e. left or right to represent bit 0 or 1 respectively . Words shift are detected using correlation method that treats a profile as a waveform and decides whether it originated from a waveform whose middle block has been shifted left or right . This method can be identified less, because change of distance between words to fill a line is quite common .

2.4.3. SYNTATIC METHOD

This technique uses punctuation marks such as full stop (.), comma (,), etc. to hide bits 0 and 1[1]. But problem with this method is that it requires identification of correct places to insert punctuation marks . Therefore, care should be taken in using this method as readers can notice improper use of the punctuations .

2.4.4 SEMANTIC METHOD

This method is very similar to word spelling method. Rather than encoding binary data by exploiting ambiguity of form, these methods assign two synonyms primary or secondary value.For example, the word "big" could be considered primary and "large" secondary.

Whether a word has primary or secondary value bears no relevance to how often it will be used, but , when decoding, primary words will be read as ones, secondary words as zeros.

2.4.5. WHITE STEG OR OPEN SPACE METHOD

This technique uses white spaces for hiding a secret message. There are three methods of hiding data using white spaces. In **Inter Sentence Spacing**, we place single space to hide bit 0 and two spaces to hide bit 1 at the end of each terminating character . In **End of Line Spaces**, fixed number of spaces is inserted at the end of each line. For example, two spaces to encode one bit per line, four spaces to encode two bits and so on. In **Inter Word Spacing** technique, one space after a word represents bit 0 and two spaces after a word represents bit 1. But, inconsistent use of white space is not transparent[10]

2.4.6. SPAM TEST

HTML and XML files can also be used to hide bits. If there are different starting and closing tags, bit 0 is interpreted and if single tag is used for starting and closing it, then bit 1 is interpreted . In another technique, bit 0 is represented by a lack of space in a tag and bit 1 is represented by placing a space inside a tag[1] .

2.4.7. SMS-TEXTING

When SMS is used for sending text messages it utilizes the various algorithms of text steganography .We can send binary images through SMS and use this image as cover data for hiding text data into SMS which is new concept in steganography. Further an enhanced method for SMS steganography using SMS-texting language, by removing the static nature of word-abbreviation list and introducing computationally light weighted XOR encryption[8]. The dynamic level arrangement of 'word-abbreviation list', if used alone, provides moderate level of security and makes it difficult for an adversary to instantly extract zeros' or ones' out of the SMS-text by knowing only the algorithm.

2.4.8. FEATURE CODING

In feature coding, secret message is hidden by altering one or more features of the text. A parser examines a document and picks out all the features that it can use to hide the information[8] .

For example, point in letters i and j can be displaced, length of strike in letters f and t can be changed, or by extending or shortening height of letters b, d, h, etc. .

2.4.9. SSCE (Secret Steganographic Code for Embedding)

This technique first encrypts a message using SSCE table and then embeds the cipher text in a cover file by inserting articles a or an with the non specific nouns in English language using a certain

mapping technique . The embedding positions are encrypted using the same SSCE table and saved in another file which is transmitted to the receiver securely along with the stego file[1].

2.4.10 WORD MAPPING

This technique encrypts a secret message using genetic operator crossover and then embeds the resulting cipher text, taking two bits at a time, in a cover file by inserting blank spaces between words of even or odd length using a certain mapping technique . The embedding positions are saved in another file and transmitted to the receiver along with the stego object[6].

2.4.11. CSS (Cascading Style Sheet)

This technique encrypts a message using RSA public key cryptosystem and cipher text is then embedded in a Cascading Style Sheet (CSS) by using End of Line on each CSS style properties, exactly after a semicolon. A space after a semicolon embeds bit 0 and a tab after a semicolon embeds bit 1[1] .

PERFORAMANCE ANALYSIS

	Advantage	Disadvantage
Line Shifting	This method is suitable only for printed text. In printed text OCR(character recognition) never used.	When OCR is applied, the hidden information is destroyed.
Word Shifting	This identify less because of change of distance between words to fill line is quite common.	If anyone aware about the algorithm that related to word shifted distance then easily get hide data .It can be destroyed by using OCR.
Syntactic Method	The amount of information to hidden the method is trivial.	Smart reader can find hidden data easily
Semantic	This method is better that above methods syntactic ,line and word shifting because that cannot detect by retyping or using OCR programs.	Smart readers which has huge knowledge of words can easily discover it.
SMS-Texting	The cover media is in forms of SMS either text or binary image is very cost effective service.	Limited size of the embedded data.
Feature Coding	We can change case of every word to proper case or lower case depending on the secret bit.	If an OCR program is used or if re-typing is done, the hidden content would get destroyed.
Spam Text(HTML and XML)	Any html document has a considerable number of tags and attributes. Thus the capacity of the hiding process to hide secret messages is also high in the proposed technique	This scheme is has not used a cryptographic techniques to encrypt the message so that if the intruder knows the techniques of the steganographic, then the message can be solved.
White Steg or Open Space method	Changing the number of trailing spaces has little chance of changing the meaning of a phrase or sentence and casual reader is unlikely to take notice.	A word processor may inadvertently change the number of spaces, destroying the hidden data
CSS (Cascading Style Sheet)	Since CSS stored on the web servers so it is not possible to changing the data by any third parties	The limited amount of characters that can be embedded, it is depending of the available semi-colon amount.

SSCE (Secret Steganographic Code for Embedding)	It generate the encrypted form of the message in order to achieve high level of security. This approach is capable of secure transfer of the message compared to earlier techniques.	Though there are no shortcomings in this method, but work should be done to improve the capacity of the embedding scheme by incorporating some compression technique on the secret message.

3. CONCLUSION

The three main performance parameters to be considered when studying steganographic systems are: **capacity, security and robustness**.
Capacity refers to the ability of cover media to conceal secret information.
Security refers to the ability of an eavesdropper to figure out the hidden information easily.
Robustness refers to the ability of protecting unseen data from modification.

As steganography becomes more widely used in computing there are issues that need to be resolved. A wide variety of different techniques or approaches are discussed in present paper with their advantages and disadvantages. Many are not robust enough to prevent detection and removal of embedded data.
Steganography although conceals the existence of a message but is not completely secure. It is not meant to supersede cryptography but to supplement it. Here, approaches ranging from simple line shifting method to usage of html tags and attributes is being used to hide the data. A lot of work needs to be done in the field of steganography and how it can complement the cryptographic method of text-based data hiding.

4.ACKNOWLEDGEMENT

The authors are thankful to Mr. Hitesh Singh **,**Assistant Professor, Dept. of CSE, H.M.R Institute Of Technology , New Delhi, India, for his constant encouragement and valuable
suggestions and encouragement.

5.REFERENCES

[1]Monika Agarwal," TEXT STEGANOGRAPHIC APPROACHES: A COMPARISON," *International Journal of Network Security & Its Applications (IJNSA), Vol.5, No.1,2013*
[2] Mohit Garg, "A Novel Text Steganography Technique Based on Html Documents, " *International Journal of Advanced Science and Technology Vol. 35, October, 2011*
[3] Ei Nyein Chan Wai and May Aye Khine , "Syntactic Bank-based Linguistic Steganography Approach," *2011 International Conference on Information Communication and Management IPCSIT vol.16 ,2011*
[4] Ammar Odeh1 , Khaled Elleithy and Miad Faezipour,"Text Steganography Using
Language Remarks,"
[5] Tohari Ahmad, Melvin S. Z. Marbun, Hudan Studiawan, Waskitho Wibisono, and Royyana M. Ijtihadie, "A Novel Random Email-Based Steganography," *International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 4, No. 2, April 2014*
[6] 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*
[7] Hitesh Singh and Kriti Saroha," Reconstruction of Printed Document using Text Based Steganography"
[8] Chandrakant Badgaiyan, Ashish Kumar Dewangan and Bhupesh Kumar Pandey , " A SURVEY PAPER ON SMS BASED STEGANOGRAPHY," *International Journal of Advanced Computer and Mathematical Sciences ,ISSN 2230-9624. Vol 3, Issue 4, 2012, pp 441-445*
[9] Swati Gupta and Deepti Gupta, "Text -Steganography:

Review Study & Comparative Analysis," *(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (5) , 2011, 2060-2062*

[10] Hitesh Singh, Pradeep Kumar Singh and Kriti Saroha , "A Survey On Text Based Steganography," *Proceedings of the 3rd National Conference;INDIACom-2009 Computing for Nation Development , February 26 – 27,2009*

[11] L. Y. POR and B. Delina ," Information Hiding: A New Approach in Text Steganography," *7th WSEAS Int. Conf. on APPLIED COMPUTER & APPLIED COMPUTATIONAL SCIENCE (ACACOS '08), Hangzhou, China, April 6-8, 2008*

[12] K B Shiva Kumar 1, K B Raja2, R K Chhotaray3, Sabyasachi Pattnaik4," Performance Comparison of Robust Steganography Based on Multiple Transformation Techniques," *K B Shiva Kumar et al, Int. J. Comp. Tech. Appl., Vol 2 (4), 1035-1047*