Digital Steganography for ASCII Text Documents
Ph.D. RESEARCH PROPOSAL

Khan Farhan Rafat
International Islamic University
Islamabad
+923005154788
farhan.phdcs35@iiu.edu.pk

Muhammad Sher
International Islamic University
Islamabad
+92519019527
m.sher@iiu.edu.pk

ABSTRACT
The digitization of analog signals has inadvertently opened doors for Covert Channel communication which is being exploited as an innocent carrier of Secret information. A number of techniques have been proposed and are in use to attain confidentiality, integrity and authentication for on-line transaction/exchange of messages of which Cryptography and Steganography stands ahead.

Cryptography is focused on changing the contents in a manner difficult for the adversary to interpret. Steganography, on the other hand emphasizes on hiding the existence of secret information i.e. making these appear as non-existent.

The ability to express more in fewer bits and less printing cost involved has made Text file format as an ideal candidate for hiding priceless secrets inside its body.

This Ph.D. research proposal aims at evolving a new steganographic technique for hiding information in digital ASCII Text documents, an area of research considered as the most difficult [1] because digital ASCII Text documents are devoid of any extra overhead to hide information within it. This is followed by software implementation / simulation of the concept together with proposing new / enhancing some existing text-based Steganographic techniques / methods.

General Terms
Algorithms, Documentation, Performance, Design, Reliability, Experimentation, Security, Human Factors, Theory, Verification.

Keywords
Steganography, Steganology, Steganalysis, Information Hiding, Conceal, Cover Channel, Conceal.

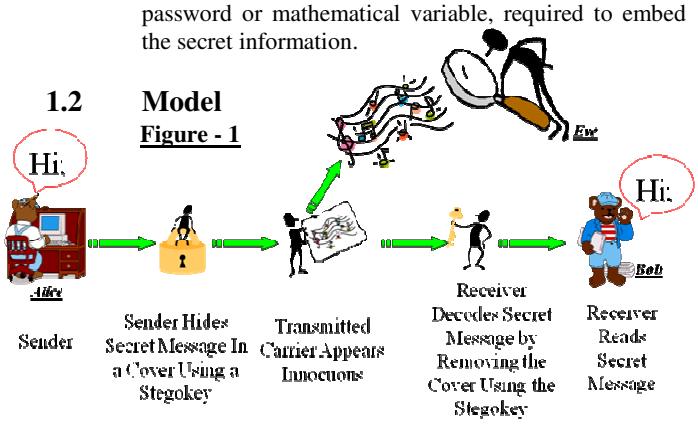
1. INTRODUCTION
Edgar Allen Poe in 'A Few Words on Secret Writing' (1841), has written that "we can scarcely imagine a time when there did not exist a necessity, or at least a desire, of transmitting information from one individual to another in such a manner as to elude general comprehension" (Rosenheim 1997, 171).

Despite of the fact that an enormous amount of our daily routine work is being routed electronically via e-mails, blogs, e-Commerce, e-fax etc., only a few people are aware of the security risks involved in such types of communication or of security precautions necessary to protect information from falling into the hands of the hostile.

Steganography is derived from the Greek words *Steganos* (meaning: covered or roofed) and *Graphos (meaning: writing),* i.e., στεγανός , γραφ-ειν : "Covered Writing [2].

1.1 Terminology
i. **Cover:** Generally, innocent looking carriers, e.g., pictures, audio, video, text, etc. that hold the hidden information
ii. **Stego_Object:** The combination of hidden data-plus-cover is known as the *stego-object*
iii. **Stegokey:** An additional piece of information, such as a password or mathematical variable, required to embed the secret information.

1.2 Model
Figure - 1



1.3 Brief History:
The recent interest in steganography (with the publication of the NEWS that steganography is being used by terrorist to hide their communications from the law enforcement agencies in **USA Today** however, should not be miss-interpreted as something being new because the use of steganography dates back to the time of ancient Greeks where in the fifth century B.C. it was practiced by the political prisoners of King Darius. Another well known attribute to steganography is the tattooing of a secret message on the shaved scalp of a slave that vanished as his hair grew. Steganography has also been effectively used by Germans in the World War II. They introduced the technique of 'microdot' where secret messages were photographed and reduced to a size of period (full stop).

FIT'09, December 16–18, 2009, CIIT, Abbottabad, Pakistan.

1.4 Choice of Cover Media: Introduction of digital technology has opened up new opportunities for steganography to flourish where a number of media are available for use as cover for hiding secret information i.e. Image, Audio, Video, Text etc. according to the changing security needs of time [3][4][5]. Unconventional way of writing messages with invisible ink using onion juice, alum, ammonia salts and other such materials which glow dark when held over a flame, also remained popular. This technique was extensively used by the British and Americans during American Revolution.

1.5 Limitations of Text Steganography: Text steganography is the most difficult amongst all of its counterparts because a digital ASCII text document is written, saved and retrieved by a computer in exactly the same way as it appears before naked-eye in contrast to other file formats such as picture where the observed and saved contents are different and are used for the purpose of hiding secret information within its structure. The challenges concerned with hiding data in text files include:

i. Changing a single bit of a byte results in an all together different ACII code that may/may not have any relevancy with the text contents.

ii. Inserting additional spaces to represent information results in an increase in stego-cover object size.

1.6 How this Proposal is organized: The organization of the Ph.D. thesis proposal is as follows: General introduction is followed by brief introduction to steganography which is narrowed down to the area where digital text documents are being used for hiding information highlighting elaborating on their advantages, disadvantages and challenges to meet. At the end, a proposed model and techniques to be used to attain the thesis objectives are discussed.

2. State of the Art: This section gives a literature review of some of the text-based steganographic techniques such as using acronyms, synonyms; semantics (Sub Para 2.1 – 2.3 etc., refers).

2.1 Acronym: Mohammad Sirali-Shahreza and M. Hassan Shirali-Shahreza from Iran have suggested the use of substitution of words with their respective abbreviations or *viza viz* in [6] to hide bits of secret message. Their suggested method operates as follows: Table - 1

Acronym (0)	Translation (1)
2l8	Too late
ASAP	As Soon As Possible
C	See
CM	Call Me
F2F	Face to face

Methodology_*A table of two columns is organized with a pre-selected list of words and their corresponding acronyms in such a way that the column under which words or its translation will appear is labeled as '1' while that containing respective acronym is labeled as '0'.The message/information required to be hidden is converted into its equivalent binary (Table 1 refers for detail).*

2.2 Change of Spelling: In his research paper [7] Mohammad Shirali-Shahreza presented a method to exploit same words which are spelled differently in British and American English for hiding secret message bits. The concealment methodology elaborated below, where the words spelled in British and American English are arranged in separate columns; is identical to that explained in preceding sub-para 2.1.

Methodology_*The column labeled '1' contain words with British spellings while that containing same words of American spelling is given a label '0'.The secret message is converted into its equivalent binary. The cover message is then iterated to find words that match to those available in pre-defined list. On finding a matching word, the secret message bit is mapped to the column headings and the word at the cross-section of that column and matched word's row is substituted in the cover message for the matched word.*

Whole of the cover message is iterated to find matching words in the list followed by substitution of word under column indicated by secret message bit.

2.3 Semantic Method: The authors Mohammad Sirali-Shahreza and M. Hassan Shirali-Shahreza in [8] have used synonym words substitution for hiding secret message bits on the anology of 2.1-2.2.

2.4 Miscellaneous techniques: A large number of idiosyncrasies methods have been given by the authors in [9] that may be used for hiding secret message bits inside a cover text e.g., by introducing modification or injecting deliberate grammatical word/sentence errors with in text. Some of the suggested techniques / procedures given in this context include:

(i) *Typographical errors - "tehre" rather than "there".*
(ii) *Using abbreviations / acronyms - "yr" for "your" / "TC" in place of "Take Care".*
(iii) *Transliterations – "gr8" rather than "great".*

2.5 MS Word Document: Change tracking technique of MS Word have been use by the author at [10] for hiding secret information, by making the stego-object appeared as work of collaborated writing.

Methodology_*The message bits to be hidden are first embedded in the degenerated segments of the cover document, followed by the revision of degenerated text, imitating it as being an edited piece of work.*

2.6 Hiding data within white spaces [11]: The steganographic technique of representing a bit '1' with space and bit '0' as two spaces or viza viz is equally applicable on web pages. Placing spaces between HTML TAGS does not affect the visibility of the web contents nor does viewing the source, as previously explained, gives an instant clue to user.
The draw back associated with this technique is that the size of the document gets increased considerably.

2.7 Hiding data by changing case of TAG - [12][13]: As already stated, HTML Tags and associated members are case insensitive e.g., <html>, <HTML> or <hTmL> will have the same impact on the document's outlook. Bits are hidden in TAGS by changing the case of the alphabets based on the bit as either '0' or '1'.
The draw back is that the changes are frequent and besides eye catching these can also be easily decoded to extract the hidden information in the absence of any pre-agreed stego-key and usage details.

2.8 Hide data in HTML comments [14]:

Comments in HTML are placeholders that convey information as add to the developer's memory or to make a note of something that will help in subsequent development / modification or to explain the purpose of the contents that follows comment.

3. Advantages and Disadvantages:

Table 2 summarizes the advantages and disadvantages of the existing text-based data hiding methods and proposed enhancement.

4. Challenges:

i. **Perceptibility/Quality:** Issues concerning "distortion" of cover medium by embedding process i.e., minimum requirement = visually acceptable (More of subjective nature).

5. Research Methodology:

All text based steganographic techniques make use of insertion and / or substitution method for hiding secret bits as changing a single bit of a byte results in a different ASCII code that may / may not have any relevancy with the word/ phrase / sentence. **Our proposed solution will be based on any or both of the aforementioned techniques**.

6. Proposed Model:

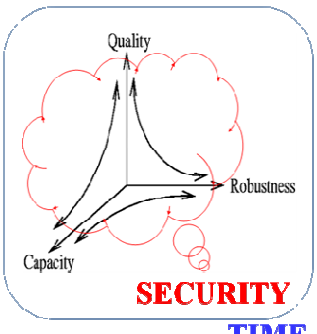
Our implementation model is as shown in Figure - 3. Although our primary emphasis is on evolution of new text-based steganographic technique but we are also aiming towards providing a blended solution comprising of Encryption and Steganography dully added by compression before encryption. To focus more on the subject and proposing a closest possible feasible secure communication solution, we shall

Table – 2 Advantages & Disadvantages of EXISTING and ENHANCED METHOD(s)

TECHNIQUES	ADVANTAGES	DISADVANTAGES
Acronym	Speed Flexible i.e., word/acronym list can grow dynamically as desired. Can easily be implemented in a variety of fields like Science, medicine, etc.	The main drawback lies in the static word/acronym substitution. Anyone who knows the algorithm can easily extract the hidden bits of information and decode the message i.e., non-adherence to Kerckhoff's Principle.
Change of Spelling	Speed	Language specific Static substitution methodology Does not follow Kerckhoff's Principle
Semantic Methods	Speed	Language specific Does not follow Kerckhoff's Principle Only one synonym is taken for words having multiple synonyms
Miscellaneous Techniques	Variant data hiding techniques	Eye Catching
MS Word	Easy to use as most users are versed with MS Word	Easily detectable due to MS Word built-in spell-checker and Artificial Intelligence (AI) features
HTML	Spacious	Increase in Stego-cover file size Does not follow Kerckhoff's Principle
XML	Steg-Analysis is difficult as bulk of data is exchanged over the next via XML	Eye catching Increased Stego-cover file size Does not follow Kerckhoff's Principle
White Space	Can pass by undetected through human eye	Does not follow Kerckhoff's Principle
Line Shift	Difficult to detect in the absence of original text	Looses format if the document is saved as text
Feature Coding	More variations for hiding information	Eye catching

PROPOSED ENHANCED ALGORITHM	
ADVANTAGES	DISADVANTAGE
Compression, Encryption, Dynamic substitution, Adherence to Kerckhoff's Principle	Fractionally Slower than existing technique

Figure - 2



i. **Capacity:** Tradeoff between perceptibility and embedded information (Matters concerning to Information Theory).

ii. **Robustness:** The proposed embedded solution should be strong enough so that an intruder can not extract the secret information out of the stego-cover.

iii. **Security & Time:** The security of embedded information should remain intact at all time.

be using the best available compression (Paq8p), hash (SHA-2) and encryption algorithms/techniques (AES), whose statistics have already been studied /gathered, if we may not come up with our own algorithms.

7. Publication:

We so far, on our credit, have one international research publication titled: Enhanced text steganography in SMS, Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on 17-18 Feb. 2009, Digital Object Identifier 10.1109/IC4.2009.4909228.

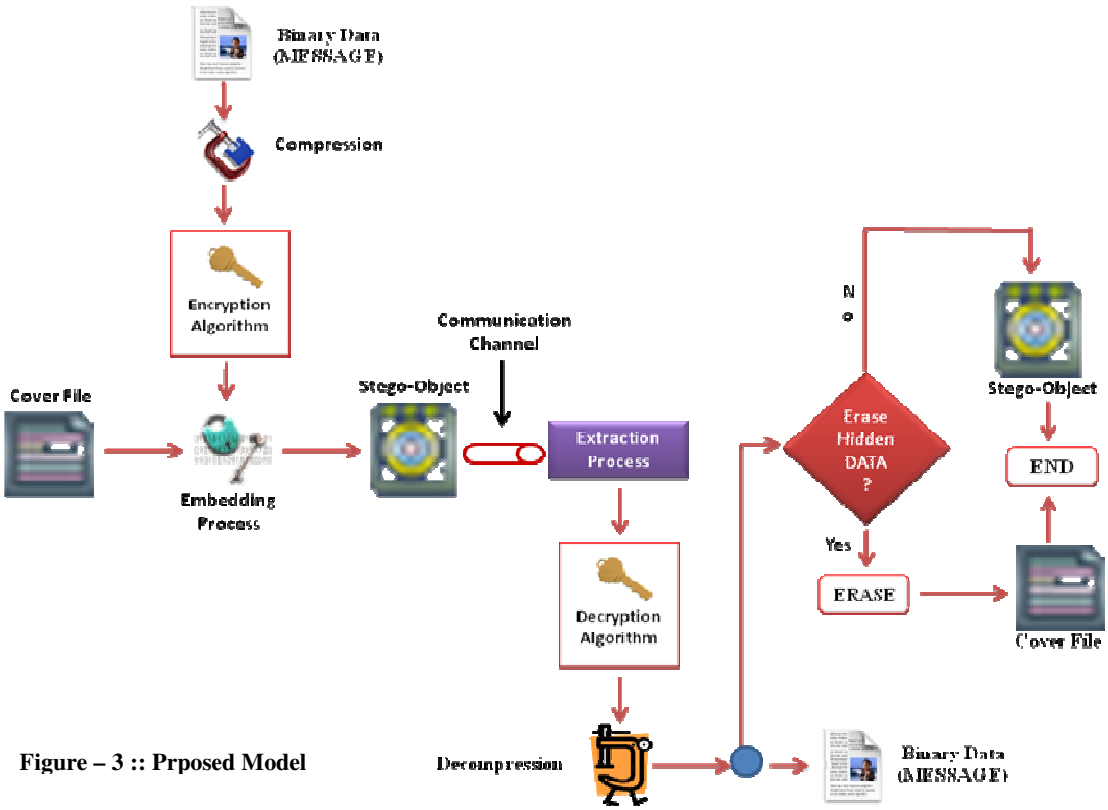


Figure – 3 :: Prposed Model

References

[1] **W. Bender, D. Gruhl, N. Morimoto, and A. Lu**, *Techniques for data hiding IBM Systems Journal*, Vol. 35, Issues 3&4, pp. 313-336, 1996.

[2] **Dave Kleiman (Technical Editor), Kevin Cardwell, Timothy Clinton, Michael Cross, Michael Gregg, Jesse Varsalone**, *The Official CHFI Study Guide (Exam 312-49) for Computer Hacking Forensic Investigators,* Published by: Syngress Publishing, Inc., Elsevier, Inc., 30 Corporate Drive, Burlington, MA 01803, Craig Wright

[3] **Stefan Katzenbeisser, Fabien A. P. Petitcolas,** *Information Hiding Techniques for Steganography and Digital Watermarking*, Artech House, Boston – London

[4] **Nedeljko Cvejic**, *Algorithms For Audio Watermarking And-Steganography*, Department of Electrical and Information engineering, Information Processing Laboratory, University of Oulu, 2004.

[5] **Jessica Fridrich, Tomáš Pevný, Jan Kodovský,** *Statistically Undetectable JPEG Steganography: Dead Ends, Challenges, and Opportunities*, Copyright 2007 ACM 978-1-59593-857-2/07/0009 ...$5.00.

[6] **Mohammad Sirali-Shahreza, M. Hassan Shirali- Shahreza**, *Text Steganography in Chat,* 1-4244-1007/07 © 2007 IEEE

[7] **Mohammad Shirali-Shahreza**, *Text Steganography by Changing Words Spelling*, ISBN 978-89-5519-136-3, Feb. 17-20, 2008 ICACT 2008

[8] **M. Hassan Shirali-Shahreza, Mohammad Shirali-Shahreza**, *A New Synonym Text Steganography* ,International Conference on Intelligent Information Hiding and Multimedia Signal Processing 978-0-7695-3278-3/08 © 2008 IEEE

[9] **Mercan Topkara, Umut Topkara, Mikhail J. Atallah**, *Information Hiding Through Errors: A Confusing Approach*, Purdue University

[10] **Tsung-Yuan Liu, Wen-Hsiang Tsai,and Senior Member**, *A New Steganographic Method for Data Hiding in Microsoft Word Documents by a Change Tracking Technique*, 1556-6013 © 2007 IEEE

[11] **NeoByte Solutions**, "*Invisible Secrets 4*", http://www.invisiblesecrets.com/index.html

[12] **Mohammad Shirali Shahreza**, *A New Method for Steganography in HTML Files*, Computer, Information, and Systems Sciences, and Engineering, Proceedings of IETA 2005, TeNe 2005, EIAE 2005, 247-251, Springer

[13] **K. Bennett**, "*Linguistic Steganography: Survey, Analysis, and Robustness Concerns for Hiding Information in Text*", Purdue University, CERIAS Tech. Report 2004-13

[14] HIPS Systems, "Shadow Text", http://home.apu.edu/~jcox/projects/HtmlStegol