Volume 3, No. 5, May 2012
Journal of Global Research in Computer Science
ISSN-2229-371X
REVIEW ARTICLE
Available Online at www.jgrcs.info

AN OVERVIEW OF VARIOUS FORMS OF LINGUISTIC STEGANOGRAPHY AND THEIR APPLICATIONS IN PROTECTING DATA

M. K. Kaleem
ME (Canada), MS (IT), B-Tech.
Wollega University, Nekemte, Ethiopia.

***Abstract*—** Steganography is defined as the science and art of writing hidden messages in such a way that no one apart from the intended recipient knows of the message existence. Due to these reasons it has become absolutely essential to keep the systems free from different internet attacks and take steps to eliminate any risks. Linguistic steganography, a type of steganography is defined as a collection of techniques and methods that allows the hiding of any digital information within texts based on some linguistic knowledge. Linguistic steganography is of various forms and have varied applications. This paper makes a study on various forms of linguistic steganography and their applications in protecting data.

***Index Terms*—** Linguistic Steganography, Steganography application, data protection.

INTRODUCTION TO STEGANOGRAPHY

Cross and Shinder (2008, p 525) described that steganography is referred to a method of hiding data not just observing its concepts as encryption does but observing its very existence. Steganography is used in conjunction with encryption for additional security of sensitive data. This method improves one of the largest issues of encrypting data the fact that it is encrypted grabs the attention of people who are viewing for sensitive or confidential information.

Steganography is defined as the science and art of writing hidden messages in such a way that no one apart from the intended recipient knows of the message existence. In steganography the message itself may not be critical to decode but the majority will not gain the presence of it. Steganography hides a message inside another message and views like a normal sound, graphic or other file. The aim of steganography is to deliver messages under cover observing the very existence of information exchange (Last and Kandel, 2010, p 95).

Contrary to that Gupta and Sharman (2008) described that steganography means encryption by means of hiding information. It consists of hiding information in any kind of data such as audio, video or image files. This refers to hiding a message in what views like an ordinary text piece.

Similarly Kizza (2010, p 273) defined that steganography is the art of information hiding in ways that avoid its detection. Steganography has viewed a rebirth with the onset of computer technology with steganographic techniques based on computer that stores information in the form of binary files, images or text by putting a message within a bigger one in such a way that others cannot predict the contents or presence of the hidden message. Steganography prevents drawing suspicion to the exchange of a hidden message. Forensic analysts pay special attention to this type of hidden data and the steganalysis uses utilities that invents and renders useless such covert messages.

Alam, Siddiqui and Seeja (2009, p 528) described that Steganography is derived from the Greek words Graptos and Steganos. Thus steganography is a process that hides sensitive or private information. Inside something that exists to be nothing out of the usual. In other words the steganography is a process that hides a file inside another file i.e. videos, pictures or audio files. When file or a information is hidden inside a carrier file usually the data is encrypted. Steganography is always confused with cryptography because both of them are common in the way that both are used to secure the information. But in steganography the file or the information is not modified and is just enclosed into the cover file. The below figure shows the steganography general model:

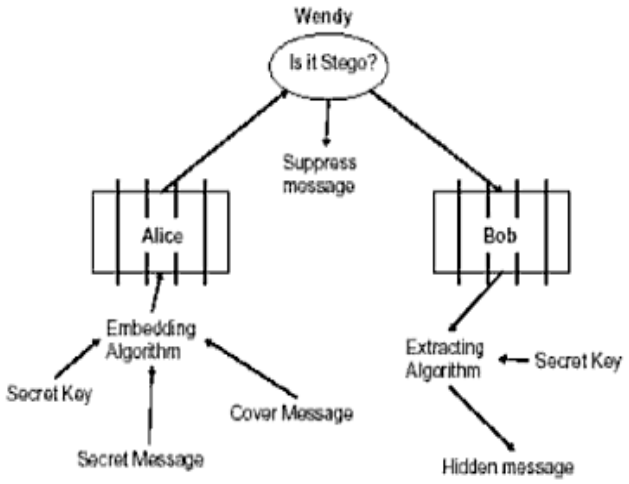

Figure 1: Steganography General Model

Source: Alam M A, Siddiqui T and Seeja K R (2009), Recent Developments in Computing And Its Applications, I K International publishing House, New Delhi, p 528

Steganography is concerned with avoiding unauthorized users from accessing plaintext data. However the basic considerations behind the steganography use are rather varied from those of cryptography. Steganography is necessarily the study of hiding information. The major goal of steganography is for a sender to exchange a plaintext to a receiver in such a way that only the receiver can retrieve the plaintext because only the receiver knows that a hidden plaintext occurs in the first place and how to view for it. In steganography an inceptor will be unaware that observed data consists of hidden information (Martin, 2012).

APPLICATIONS OF STEGANOGRAPHY

EC-Council (2010, p 1-2) have described that the steganography can be used for different illegal and legal uses and also can be used for the following needs such as:

Medical Records:

Steganography is used in medical records to avoid any mix up of records of patients. Each patient has an electronic patient record which has examinations and other medical records stored in it.

Terrorism:

Specific extremist web sites have been referred to use text and pictures to communicate messages to terrorist cells secretly performing around the world. Computers and servers around the world offer a new twist on this cover task.

Digital Music:

Steganography is also used to secure music from being copied by introducing subtle alterations into a music file that act as a digital signature. BlueSpike Technology removes a few chosen tones in a narrow band. Verance adds signals that are out of the range of frequency predictable by human ear. Others adjust the sound by altering the frequency slightly. Digital audio files can also be modified to carry a huge number of information. Some files simply denote that the content is under copyright. More sophisticated versions of steganography can consist of information about the artist.

Workplace Communication:

Steganography can be used as an effective method for employees who miss privacy in the workplace to bypass normal communication channels. In this area Steganography can be an obstacle to network security.

Movie Industry:

Steganography can also be used as copyright security for VCDs and DVDs. The DVD copy security program is configured to support a copy generation management system. Second generation DVD players with capabilities of digital video recording continues to be introduced in the black market. To secure itself against piracy, the movie industry requires to copyright DVDs.

WHAT IS LINGUSTIC STEGANOGRAPHY?

According to Fridrich (2004, p 180) linguistic steganography is defined as a collection of techniques and methods that allows the hiding of any digital information within texts based on some linguistic knowledge. To hide the very fact of hiding the out coming text must not only remain inconspicuous i.e. exist to be ordinary text with orthography, fonts, morphology, syntax, lexicon and word order outwardly corresponding to its meaning but also conserve semantic cohesion and grammatical correctness.

FORMS OF LINGUISTIC STEGANOGRAPHY

The below figure shows the types of linguistic steganography:

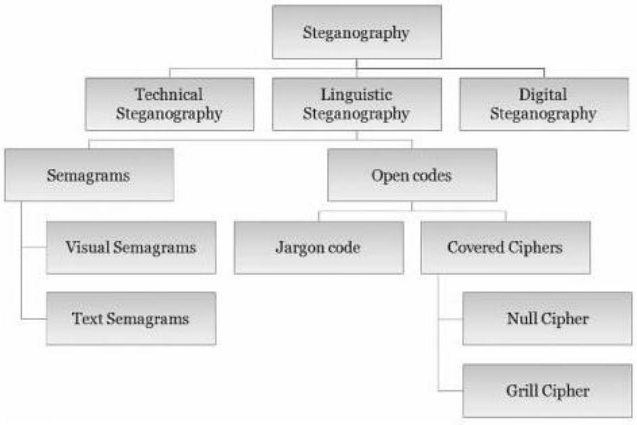

Figure 2: Types of Linguistic Steganography

Source: EC-Council (2010), Computer Forensics: Investigating Data and Image Files, Cengage Learning, USA, p 1-4

The steganography is categorized into 3 types:

i. Technical steganography
ii. Linguistic Steganography,
iii. Digital Steganography.

Here the linguistic steganography is described briefly:

The linguistic steganography hides the message in the carrier in some non-obvious ways and is further classified as open codes and semagrams.

a) Semagrams: The semagrams hides information by the use of signs or symbols. A visual semagrams uses everyday physical objects or innocent viewing to convey a message such as doodles or the components positioning on a website or desk. In this technique a painting, music, drawing, letter or any other symbol is used to hide the information. A text semagrams hides a message by modifying the carrier text appearance such as subtle alteration in font type or size, adding extra spaces or varied flourishes in handwritten text or letters (Goje, Gornale and Yannawar, 2007, p 201).

b) Open codes: Open codes hide a message with a legitimate carrier message in ways that are obvious to an unsuspecting observer. Sometimes the carrier message is referred as the overt communication while the hidden message is the covert communication. Open codes make use of openly readable text. This text consists of sentences or words that can be hidden in a vertical or reversed order. The letter must be in chosen selected place of the text. This classification is subdivided into covered ciphers and jargon codes (Shih, 2010, p 477).

c) ***Jargon codes:*** According to Zelkowitz (2011, p 54) a jargon code uses languages that is understood by a group of people but is meaningless to other people. Jargon codes consists of warchalking i.e. the symbols used to indicate the type and presence of wireless network signal undergoing terminology or an innocent conversation that conveys special meaning because of known facts only to the speakers. The jargon codes are common to a substitution cipher in several respects but rather than exchanging individual letters the words themselves are altered. A subset of jargon codes is cue codes where specific prearranged phrases convey meaning. The cue codes use a possibly brief carrier message to signal the existence of an event whose semantics have been prearranged.

d) ***Covered Ciphers:*** Arnold, Schmuker and Wolthusen (2003, p 17) described that covered or concealment ciphers hides a message openly in the carrier medium so that it can be recovered by anyone who knows the secret for how it was concealed. The covered ciphers can be categorized into grill and null ciphers.

e) ***Grill ciphers:*** A grill cipher employs a template that is used to cover the carrier message and the words that exist in the template openings are the hidden message. It is possible to encrypt plaintext by writing it onto a sheet of paper through separate pierced cardboard or paper sheet. When a recognized pierced sheet is placed on the message the actual text can be read. The grill system is critical to decipher and crack as only the person with the grill can decipher the hidden message.

f) ***Null Ciphers:*** According to some prearranged set of norms null cipher hides the message such as read every 5th word or view at the 3rd character in each word. The null cipher hides the message within a huge number of useless data. The actual data may be mixed with the unused data in any order permitting only the person who knows the order to understand it.

PROTECTING DATA USING LINGUISTIC STEGANOGRAPHY

The power of computation is highly capable to analyze several critical linguistic structures. While the produced texts may approximate some of the legitimate text appearance "simulating the statistical text properties" must increasingly manage properties of linguistic steganography as well as orthographic and lexical distribution. Particularly linguistic steganography assumes the linguistic properties of modified and produced text, and in several cases, uses linguistic structure as the space in which the messages are hidden. This explains linguistic steganography occurring methods as background for the linguistic problems.

One solution to predict ungrammatical sequence of lexical items because syntactic analysis can be used and one solution prediction by syntactic steganalysis is to assure that structures are syntactically proper from the initiation. In fact, the data of steganography can be hidden within the syntactic structure itself. Wayner proposed that context-free grammars can be used as a basis for steganographic texts production (Wayner 1992, 2002). Because the text is produced directly from the grammar, unless the grammar is flawed syntactically and the text is guaranteed to be syntactically correct. Furthermore, the tree framework created by context free grammars is a natural tool to use for encoding bits. The tree structures are used as optimizing data structures in several computer science areas from compilers to sort algorithms. At any point where there is a branch in syntax tree in the simplest scheme, the right branch might mean '1' and left branch might mean '0'. This section explains occurring context free grammar methods to produce proper steganographic text appropriately.

Mann and Thompson (1988) described that the simplest way to produce correct texts syntactically is to produce them from syntax itself. This seems obvious, but predicting a way to hide information within text may not be importance. Wayner proposes a custom-made context free grammar in which any optional branch in a production denotes a sequence of bits.

Wayner's method considers that the grammar is in Greibach Normal Form in order to avoid left recursion in parsing i.e. non-terminals come only at productions end. For examples consider the following grammar:

```
Start := subject predicate
subject := propernoun || The noun
predicate := is adjective || can't be adjective
adjective := asleep || crusty
noun := moose || ball of cheese
propernoun := Oscar the Grouch || Trogdor the Burninator
```

By determining that for each stage in grammar where there is an option the first option is 0 and the second option is 1, the bit sequences can be encoded and when parsing text created from this grammar, then extract the bit sequences. Note that if this grammar is represented as a tree according to pre order traversal the bits are encoded. Thus at a non terminal node when a determination is made the bit from that determination is encoded, followed by whole decisions from the left child's sub tree and then its right child sub tree. To encode the bit string 0110 a pre-order grammar traversal is used. First the start symbol is processed from left to right; then the first decision needs 0 bit, so a proper noun is selected as a subject and 0 is encoded. Then 1 is required when selecting proper noun, next 2nd choice Trogdor the Burninator is selected. Though all the subject sub trees have been processed and its sub trees and predicate are also processed. Though the next bit, 1 is needed in the sequence next the 2nd choice for predicate, is chosen which is not an adjective (Johnson, Neil, 2000). Finally o is required and the 1st choice for adjective is selected which is asleep. So to encode the bit string 0110 with this grammar the output string will be "Trogdor the Burninator can't be asleep."

Outside the fact that actual phrases and indeed whole sentences must be repeated in constructed grammars

carelessly it is also true that unless big grammars are built syntactic structures may be repeatedly detectable (Chapman et al 2001 159). Developing grammars of this size can be a difficult activity. Even if the sentences produced by these grammars are readable by humans, the "writing style" that arouses out of such generation may be strange that both computers and humans may be capable to predict the syntactic constructions, register, vocabulary usage anomalies and so forth. None of the methods so far has any semantics of words concept within a syntactic framework.

One step toward such an approach comes from Davida and Chapman who developed the NICETEXT system is to do 2 things: the first thing is that it targets to create "interesting parts of speech sequences", and second ting is that they target to classify such parts-of-speech by "kinds" (which are significantly semantic classifications) and to unite these kinds into syntactic structures used for production to make the text more believable to human reader (Chapman and Davida, 1997). This was done by enclosing big code dictionaries which classifies words by these kinds and by creating style sources which acts as templates of syntactic in which the words must be inserted (Chapman 1998: 6). Dictionaries are capable with style sources if they consist of all kinds required by the source templates of style. These code dictionaries is having words classified by type and also consist of bit values which denotes every word for the steganographic data encoding.

The style sources initiates with a sentence model, which is a syntactic template with extra information for sentence formatting. When producing a text the syntactic framework is selected from a sentence model table, which is a collection of structures of syntactic sentences which has similar semantic classifications as the vocabulary words in the created dictionary and one of the words which matches the present semantic type and part of speech with the desired value of bit is inserted. Such texts have a unique "style" which is derived by creating the sentence model table through huge corpus analysis. Information is not encoded in the grammar but by the word options which is inserted into the present sentence model. When the corpora is used as a sentence model tables source from obscure or technical sources of language and there is a greater chance that the produced text will be human readers because the style varies greatly from different syntactic frameworks (Provos, Niels, 2001).

PROPOSED SYSTEM

Steganography is a technique in which media files such as image or audio will be used to represent the original information in such a way that no one but the sender and recipient understands them.

While most system support either one form by name image or audio, this system will support two forms of encryption of messages such as Steganography with audio as well as Steganography with image. As the name suggests Steganography with audio hides data in an audio format and Steganography with image hides the data in the form of image. This system will encrypt the message together with an image and audio and send to recipient's end such that the data is very secure. The following figure explains the working of the system:

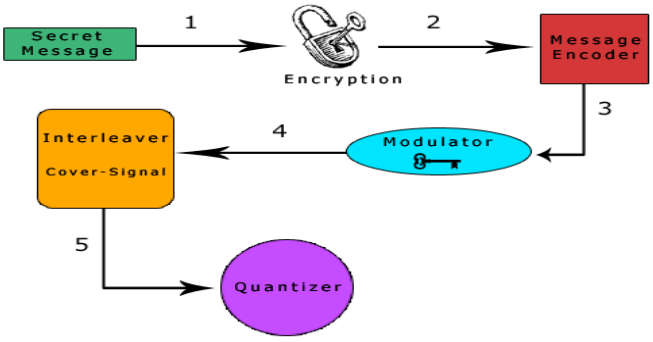

Figure 3: Steganography with Image and Audio.

The following are the steps of working of this system:

a. A symmetric key (K1) is used by the system to encrypt the secret message.
b. The encrypted data is encoded with an error-correcting code that is of low rate in order to increase the system's overall robustness.
c. A pseudorandom signal is then generated with the help of the second key (K2) and is used to modulate the encoded information.
d. The resultant signal is then interleaved with the help of a cover signal.
e. This signal is additionally quantized in order to create a brand new digital audio file that has in itself the original data or message and send to recipient.
f. The recipient reverses this process and extracts the original message.

This system prevents the data from regular hacking and is more secure when compared with conventional steganographic systems.

MODULES

a. Login: - Login is the basic module of the system. It is for security reason to prevent unauthorized users accessing the tool so as to prevent the data to be extracted from the steged media.

Image Steganography:

Image Steganography is one of the steganographic modules of the system. It provides facility of hiding data in an image. It has two sub modules as follows,

a. Embed: - Embed is the module in which user can select an image file and provide data to hide. The data may be a simple text or a text file. If the file is to hidden the file must be selected first and an image file has to be selected. The embedding process first comprises of analyzing the image if it is sufficient for hiding the data. Then the data is made to hide in the binary format of the image.

b. Extract: - Extract is the module in which user can select an image file from which the hidden data is extracted. The extraction of data from the image is the reverse process of embedding it in them.

Audio Steganography:

Audio Steganography is one of the steganographic modules of the system. It provides facility of hiding data in an audio file. It has two sub modules as follows,

a. Embed: - Embed is the module in which user can select an audio file and provide data to hide. The data may be a simple text or a text file. If the file is to hidden the file must be selected first and an audio file has to be selected. The embedding process first comprises of analyzing the audio if it is sufficient for hiding the data. Then the data is made to hide in the binary format of the audio.

b. Extract: - Extract is the module in which user can select an audio file from which the hidden data is extracted. The extraction of data from the audio is the reverse process of embedding it in them.

CONCLUSION

The use of context free grammar to mimic the normal text syntactic statistical profile making a steganographic text conform to rhetorical and semantic standards enhances the stego text into the rhetorically and semantically statistical profile of cohesive texts which do not manage the contents of steganography. Thus the linguistic encloses production targets to defend against both linguistic and statistical steganalysis. The context free grammar based syntactic mimicking of Wayner is, when there is a possibility of predicting statistical inconsistencies in a cover text because of structural factor and that structural factor must be used as a means for producing the solution, guaranteeing that the profile of statistical of the produced approximates cover that of normal text in terms of specific structural factor. Thus the future work in rhetorical and semantic mimicking will be required to consist rhetorical and semantic structural analyses. Further tree based structure exploitation of the ontology must be used as a source not only for encoding bits through relations and the concepts between them, but also for assuring that the texts stayed "on topic" in the production process by paying attention to the classes and concepts and lexical components which they are associated with, composed of and retrieved from. The future work will establish these choices in the course of configuring coherent cover generation mechanisms semantically.

REFERENCES

[1]. Cross M and Shinder D L (2008), Scene of the Cybercrime, Syngress Publishing Inc., USA, p 525.

[2]. Last M and Kandel A (2010), Web Intelligence and Security, IOS Press, USA, p 95.

[3]. Gupta M and Sharman R (2008), Handbook of Research on Social and Organizational Liabilities in Information Security, Information Science, USA.

[4]. Kizza J M (2010), Ethical and Social Issues in the Information Age, Springer, New York, p 273.

[5]. Alam M A, Siddiqui T and Seeja K R (2009), Recent Developments In Computing And Its Applications, I K International publishing House, New Delhi, p 528.

[6]. Martin K M (2012), Everyday Cryptography, Oxford University Press, New Delhi.

[7]. EC-Council (2010), Computer Forensics: Investigating Data and Image Files, Cengage Learning, USA, p 1-2.

[8]. Fridrich J (2004), Information Hiding: 6th International Workshop, IH 2004, Toronto, Canada, May 23-25 2004, Revised Selected Papers, Springer, New York, p 180.

[9]. Goje A C, Gornale S S and Yannawar P L (2007), Proceedings Of The 2Nd National Conference On Emerging Trends In Information Technology (Eit-2007), I K International Publishing, New Delhi, p 201.

[10]. Shih F Y (2010), Image Processing and Pattern Recognition: fundamentals and Techniques, John Wiley & Sons, New Jersey, p 477.

[11]. Zelkowitz M (2011), Security on the Web, Academic Press, UK, p 54.

[12]. Arnold M K, Schmuker M and Wolthusen S D (2003), Techniques and Applications of Digital Watermarking and Content Protection, Artech House, USA, p 17.

[13]. Chapman, Mark T (1998), Hiding the Hidden: A Software System for Concealing Ciphertext as Innocuous Text. Milwaukee: University of Wisconsin-Milwaukee.

[14]. Chapman, Mark and George Davida (1997), "Hiding the Hidden: A Software System for Concealing Ciphertext as Innocuous Text" Proceedings of Information Security, First International Conference, Lecture Notes in Computer Sciences 1334. Berlin: Springer, 333-345.

[15]. Chapman, Mark, George I. Davida, and Marc Rennhard (2001), "A Practical and Effective Approach to LargeScale Automated Linguistic Steganography." Proceedings of the Information Security Conference (ISC '01), Malaga, Spain, p 156-165.

[16]. Wayner, Peter (1992), "Mimic Functions." Cryptologia XVI: 3, p 192-213.

[17]. Wayner, Peter (2002), Disappearing Cryptography: Information Hiding: Steganography & Watermarking (second edition), Morgan Kaufmann, San Francisco.

[18]. Johnson, Neil F (2000), "Steganalysis" In Information Hiding: Techniques for Steganography and Digital Watermarking, Artech House, Boston, p 79-93.

[19]. Mann and Thompson (1988), "Rhetorical Structure Theory: Toward a Functional Theory of Text Organization." Text, 8:3, p 243-281.

[20]. Provos, Niels (2001), "Defending Against Statistical Steganalysis.", CITI Technical Report 01-4, University of Michigan.

