Special Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com

Linguistic Steganography and Their Applications in Protecting Data - A Review

K.Bharathi SenthilKumar, M.C.A.,M.Phil.,
Assistant Professor, PG & Research Department of Computer Science, Joseph Arts and Science College,
Thirunavalur, Villupuram, Tamil nadu, India

***Abstract*--** Steganography is the science of hiding the information into the other information so that the hidden information appears to be nothing to the human eyes. Steganography schemes are discussed for different file formats. There are many ways to hide information inside an image, audio/video, document etc. 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. The type of steganography which is defined as a collection of techniques and methods that allows the hiding of any digital information within texts based on some linguistic knowledge is known as linguistic steganography. Linguistic steganography is of various forms and have various applications. This paper makes a study on various forms of linguistic steganography and their applications in protecting data.

***Keywords*--** *Steganography, Linguistic Steganography, Data Protection.*

I. INTRODUCTION TO STEGANOGRAPHY

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 addition with encryption for additional security of sensitive data. This method improves one of the largest issues of encrypting data 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. Steganography hides a message inside another message and views like a normal sound, graphic or other file. 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. 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.

To hide the sensitive or private information steganography is the best process. In other words the process of hiding a file inside another file i.e. videos, pictures or audio files is also known as Steganography. The data is usually encrypted when file or information is hidden inside a carrier file. Steganography is always confused with cryptography because both Steganography and Cryptography are common in the way that they are used for securing the information. The information is not modified in Steganography and it is just enclosed into the cover file.

II. APPLICATIONS OF STEGANOGRAPHY

EC-Council 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:

A. Medical Records

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

B. 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.

C. Digital Music

Steganography acts as a digital signature which is used to secure music from being copied by introducing subtle alterations into a music file. Huge number of information are carried by modifying digital audio files. Some files denotes that the content is under copyright. Most of the sophisticated versions of steganography consists of information about the artist.

D. Workplace Communication

Steganography can be also used as an effective method for employees who miss privacy in the workplace to bypass normal communication channels.

E. 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. The movie industry requires copyright DVDs to secure itself against piracy.

III. LINGUSTIC STEGANOGRAPHY

According to author linguistic steganography is defined as the collection of techniques and methods that allows the hiding of any digital information within texts based on some linguistic knowledge is referred to as linguistic steganography. To 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.

IV. FORMS OF LINGUISTIC STEGANOGRAPHY

The steganography is basically categorized into 3 types:

1. Technical steganography
2. Linguistic Steganography,
3. Digital Steganography.

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

A. Semagrams:

The process of hiding the information by the use of signs or symbols is called as semagrams. Everyday physical objects or innocent viewing to convey a message such as doodles or the components positioning on a website or desk is used by visual semagram. This technique involves painting, music, drawing, letter or any other symbol to hide the information. Hiding a message by modifying the carrier text appearance such as font type or size, adding extra spaces into it or varied flourishes in handwritten text or letters is called as text semagrams.

B. Open codes:

Hiding a message with a legitimate carrier message that are obvious to an unsuspecting observer is called as open codes. Sometimes these types of carrier message is referred as the overt communication whereas the hidden message is referred as the covert communication. These Open codes makes use of openly readable text. And the sentences or words present in this text are hidden in a vertical or reversed order. The letter must be in chosen selected place of the text. This classification is further subdivided into covered ciphers and jargon codes.

C. Jargon codes:

The language that is understood by a group of people but is meaningless to other people is referred as jargon codes. It 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 the words themselves are altered rather than exchanging individual letters . A subset of jargon codes is cue codes where specific prearranged phrases convey meaning. A brief carrier message is used by the cue code to signal the existence of an event whose semantics have been prearranged.

D. Covered Ciphers:

Ciphers that 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 is known as covered or concealment. Theis a covered cipher is subdivided into grill and null ciphers.

E. Grill ciphers

A template is employed by grill cipher which is used to cover the carrier message and the words that exist in the template openings are the hidden message. Encrypting a plaintext by writing it onto a sheet of paper through separate pierced cardboard or paper sheet is possible in grill ciphers. The actual text can be read only when a recognized pierced sheet is placed on the message. It is very critical to decipher and crack a grill cipher as only the person with the grill can decipher the hidden message.

F. Null Ciphers

According to some prearranged set of norms, hiding the message such as reading every 5th word or view at the 3rd character in each word is known as null cipher. Hiding the message within a huge number of useless data is null cipher. The original data may be mixed with the unused data in any order. Only the person who knows the order to understand it are permitted to decipher it.

V. PROTECTING DATA USING LINGUISTIC STEGANOGRAPHY

The computation power 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. The linguistic properties of modified and produced text of linguistic steganography are particularly assumed, and in several cases, these linguistic structure uses the space in which the messages are hidden.

One solution to predict ungrammatical sequence of lexical items is syntactic analysis, which can be used and one solution prediction by syntactic steganalysis is used to assure that structures are syntactically proper from the initiation. 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. 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 natural tool that is used for encoding the bit is created by context free grammars from tree framework . The tree structures are used as optimizing data structures in several computer science areas from compilers to sort algorithms. In the simplest scheme the right branch is “1” and the left branch is “0” at any point where there is a branch in syntax.

The simplest way to produce correct texts syntactically is described by Mann and Thompson to produce them from syntax itself. It seems obvious, but predicts a way to hide information within text which may not be given more importance. A custom-made context free grammar is proposed by Wayner in which any optional branch in a production denotes a sequence of bits.

The grammar is in Greibach Normal Form, in order to avoid left recursion in parsing i.e. non-terminals come only at productions end. 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, extract the bit sequences. If this grammar is represented as a tree according to pre order traversal the bits are encoded. When a determination is made at non terminal node 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. Finally 0 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. Developing grammars of this size may be considered as a difficult activity. Even if the sentences produced by these grammars are readable, 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.

There is one step towards such an approach comes from Davida and Chapman who developed the NICETEXT system is to do 2 things: the first thing targets to create "interesting parts of speech sequences", and second thing targets 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. 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.

Dictionaries are capable with style sources if they consist of all kinds required by the source templates of style. These code dictionaries consist of words classified by type and also consists of bit values which denotes every word for the steganographic data encoding.

The style source is initiated with a sentence model, which is a syntactic template with extra information for sentence formatting. The syntactic framework is selected when producing a text from a sentence model table. This table 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, there is a greater chance that the produced text will be human readers because the style varies greatly from different syntactic frameworks.

VI. STEGANOGRAPHY WITH AUDIO AND IMAGE

The 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 other than the sender and recipient knows is also known as steganography. Most of the system support either one form by name image or audio, but 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 hiding the data in an audio format is called as Steganography with audio and hiding the data in the form of image is called as Steganography with image. The message together with an image and audio is encrypted by this system and it is send to recipient's end such that the data is much secured. The following figure explains the working of the system:

The following are the steps of working of this system:

1. The system uses a symmetric key (K1) to encrypt the secret message.
2. In order to increase the system's overall robustness, the encrypted data is encoded with an error-correcting code that is of low rate.
3. With the help of the second key (K2) a pseudorandom signal is then generated, and it is used to modulate the encoded information.

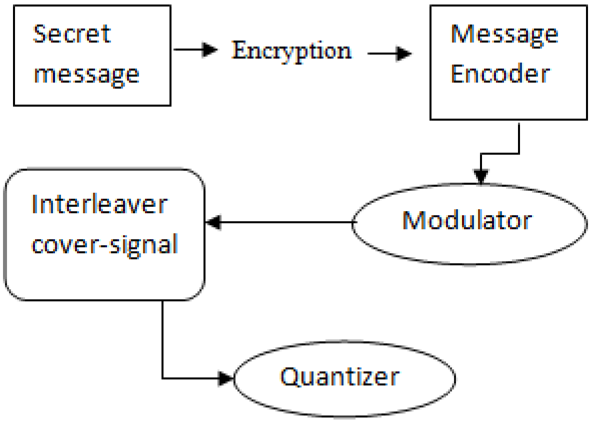

Figure1: Steganography with Image and Audio.

4. With the help of a cover signal the resultant signal is then interleaved.
5. In order to create a brand new digital audio file this signal is additionally quantized that has in itself the original data or message and send to recipient.
6. This process is reversed to extract the original message by the recipient.

The data is prevented from regular hacking and is more secured when compared with conventional steganographic systems.

A. Image Steganography:

Image Steganography is one of the steganographic modules of the system. The facility provided to hide data in an image. It has two sub modules as follows,

a. Embed

The module which is used by the user to select an image file and provide data to hide is known as embed. The data provided in this module may be a simple text or a text file. If the file is to be hidden in this method, first the file must be selected and then select an image file. Analyzing the image is first comprised by the embedding process if it is sufficient for hiding the data. Then the data is made to hide in the binary format of the image.

b. Extract

The module in which user can select an image file from which the hidden data is extracted is known as extract. The reverse process of embedding it is the extraction of data from the image.

B. Audio Steganography:

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

a. Embed

The module in which user can select an audio file and provide a facility to hide the data is known as embed. Even here the data may be in a simpler text form or a text file. If the file is to be hidden in this method, first the file must be selected and then select an audio file. Analyzing the audio is first comprised by the embedding process if it is sufficient for hiding the data. Then the data is made to hide in the binary format of the audio.

b. Extract

The module in which user can select an audio file from which the hidden data is extracted is known as extract. The reverse process of embedding it is the extraction of data from the audio.

CONCLUSION

The context free grammar is used to mimic the normal text syntactic statistical profile making a steganographic text conform to rhetorical and semantic standards which enhances the stego text into the rhetorically and semantically statistical profile of cohesive texts. And these do not manage the contents of steganography. Thus the linguistic encloses production targets to defend against both linguistic steganalysis. This paper concludes with the protection of data or file using various forms of linguistic steganography. And also the different applications to protect these data. Steganography with audio and video module is used for protecting the data.

Acknowledgment

"This paper is a revised version of a paper entitled [An Overview Of Various Forms Of Linguistic Steganography And Their Applications In Protecting Data] presented at [JGRCS 2012, volume 3]."

References

[1] M. K. Kaleem (2012), " An Overview Of Various Forms Of Linguistic Steganography And Their Applications In Protecting Data."Wollega University, Nekemte, Ethiopia.

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

[13] 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.

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

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

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

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

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