Volume 2, No. 1, January 2011
Journal of Global Research in Computer Science
ISSN-2229-371X
REVIEW ARTICAL
Available Online at www.jgrcs.info

Digital Steganalysis: Review on Recent Approaches

Indra Kanta Maitra
Research Fellow, Dept of Computer Science & Engg,
University of Calcutta, 92 A.P.C. Road,
Kolkata – 700009, India
ikm.1975@ieee.org
Sanjay Nag
Research Scholar, Dept of Computer Science & Engg,
University of Calcutta, 92 A.P.C. Road,
Kolkata – 700009, India
sanjaynag75@gmail.com
Biswajita Datta
Lecturer, Department of Computer Sc. & Engineering
St. Thomas College of Engineering and Technology
Kolkata, India
Prof. Samir Kumar Bandyopadhyay
Professor, Dept of Computer Science & Engineering,
University of Calcutta, 92 A.P.C. Road,
Kolkata – 700009, India
skb1@vsnl.com

Abstract: Steganography is the art and science of secret communication, aiming to conceal the existence of a communication, which has been used in military, and perhaps terrorists. Steganography in the modern day sense of the word usually refers to information or a file that has been concealed inside a digital Picture, Video or Audio file. In steganography, the actual information is not maintained in its original format and thereby it is converted into an alternative equivalent multimedia file like image, video or audio, which in turn is being hidden within another object. Information Security is becoming an inseparable part of Data Communication. In order to address this Information Security, Steganography plays an important role. The digital media steganalysis is divided into three domains, which are image steganalysis, audio steganalysis, and video steganalysis. DNA sequences possess some interesting properties, which can be utilized to hide data. This paper is a review of the recent steganography techniques and utilization of DNA sequence appeared in the literature.
Keywords: Steganalysis, Computational Intelligence, Image Steganalysis, Audio Steganalysis, Video Steganalysis, DNA , Data hiding, Complementary pair, and Data recovery

INTRODUCTION

Steganography is often confused with cryptology because the two are similar in the way that they both are used to protect important information [34]. The difference between the two is that Steganography involves hiding information so it appears that no information is hidden at all. If a person or persons views the object that the information is hidden inside of he or she will have no idea that there is any hidden information, therefore the person will not attempt to decrypt the information. There are two main purposes in information hiding: (1) to protect against the detection of secret messages by a passive adversary, and (2) to hide data so that even an active adversary will not be able to isolate the secret message from the cover data. Information hiding system can be divided into four areas which are Covert Channels, Steganography, Anonymity, and Copyright Marking. A survey of current information hiding has shown that steganography is one of the recent important sub disciplines. This is because most of the proposed information hiding system is designed based on steganography. Today, steganography is most often associated with the high-tech application where data are hidden with other information in an electronic file.

Today's steganographic systems uses multimedia objects like image, audio, video etc., as cover media because people often transmit digital pictures over email and other Internet communication [1-3]. In modern approach, depending on the nature of cover object, steganography can be divided into five types:

a. Text Steganography
b. Image Steganography
c. Audio Steganography
d. Video Steganography
e. Protocol Steganography

Three major computational intelligence methods have also been identified in the steganalysis domains which are Bayesian, neural network, and genetic algorithm. Each of these methods has its own pros and cons.

Many of the new attacks in steganography are derived by analysing steganography techniques. This process of analysing steganographic protocols is carried out in order to detect and extract secret messages. The process is called steganalysis which is generally starts with several suspected information streams but uncertain whether any of the

information stream contains hidden messages. The goal of steganography is to avoid suspicion on the existence of hidden messages whereas steganalysis aims to discover the hidden message from useless covert messages in a given text or data. Hence, steganalysis is the process of detecting steganography by analysing variances among bit patterns on unusually large file size [4-6].

One of the significant techniques used in steganalysis system is computational intelligence (CI). Thus, this study believed that CI can be implemented to solve steganalysis problems. Hence, this study suggests that to have a good steganalysis tool, the implementation of steganalysis system should involve some degree of CI [7-8].

In recent years, research work has been carried out on DNA-based data hiding schemes [9] . Most of them use the biological properties of DNA sequences. DNA sequence is composed of four nucleotides A, C, G, and T. Hence; we need to transform the representation format of the nucleotides such that the hiding techniques can be used to conceal the secret message in a DNA sequence.

The rest of the paper is organized as follows: In Section 2 we discuss review works on recent text hiding methods. In subsequent Section we conclude the paper with references.

REVIEW ON RECENT METHODS

Someone takes the first letter of each word of the previous sentence to see that it is possible and not very difficult. Hiding information in plain text can be done in many different ways [4]. Many techniques involve the modification of the layout of a text, rules like using every n-th character or the altering of the amount of white space after lines or between words [2-3]. The last technique was successfully used in practice and even after a text has been printed and copied on paper for ten times, the secret message could still be retrieved. Another possible way of storing a secret inside a text is using a publicly available cover source, a book or a newspaper, and using a code which consists for example of a combination of a page number, a line number and a character number. This way, no information stored inside the cover source will lead to the hidden message. Discovering it relies solely on gaining knowledge of the secret key.

To hide information, straight message insertion may encode every bit of information in the image or selectively embed the message in "noisy" areas that draw less attention—those areas where there is a great deal of natural colour variation. The message may also be scattered randomly throughout the image. A number of ways exist to hide information in digital media. Common approaches include

- Least significant bit insertion
- Masking and filtering
- Redundant Pattern Encoding
- Encrypt and Scatter
- Algorithms and transformations

Each of these techniques can be applied, with varying degrees of success.

Least significant bit (LSB) insertion is a common and simple approach to embed information in an image file. In this method the LSB of a byte is replaced with an M's bit. This technique works better for image, audio and video steganography. To the human eye, the resulting image will look identical to the cover object.

Masking and filtering techniques are mostly used on 24 bit and grey scale images. They hide info in a way similar to watermarks on actual paper and are sometimes used as digital watermarks. Masking images entails changing the luminance of the masked area. The smaller the luminance change, the less of a chance that it can be detected. Observe that the luminance in Figure 2 is at 15% in the mask region if it was decreased then it would be nearly invisible [1, 4-5]. Masking is more robust than LSB insertion with respect to compression, cropping, and some image processing. Masking techniques embed information in significant areas so that the hidden message is more integral to the cover image than just hiding it in the "noise" level. This makes it more suitable than LSB with, for instance, lossy JPEG images. JPEG images use the discrete cosine transform to achieve compression.

DCT is a lossy compression transform because the cosine values cannot be calculated exactly, and repeated calculations using limited precision numbers introduce rounding errors into the final result. Variances between original data values and restored data values depend on the method used to calculate DCT [6, 7- 8].

Embedding secret messages in digital sound is usually a more difficult process than embedding messages in other media, such as digital images. In order to conceal secret messages successfully, a variety of methods for embedding information in digital audio have been introduced. These methods range from rather simple algorithms that insert information in the form of signal noise to more powerful methods that exploit sophisticated signal processing techniques to hide information. The list of methods that are commonly used for audio steganography is given below.

- LSB coding
- Parity coding
- Phase coding
- Spread spectrum
- Echo hiding

Protocol steganography allows users who wish to communicate secretly to embed information within other messages and network control protocols used by common applications. This form of unobservable communication can be used as means to enhance privacy and anonymity as well as for many other purposes, ranging from entertainment to protected business communication or national defense. The term protocol steganography refers to the technique of embedding information within messages and network control protocols used in network transmission [4]. In the layers of the OSI network model there exist covert channels where steganography can be used [5]. An example of where information can be hidden is in the header of a TCP/ IP packet in some fields that are either optional or are never used. A paper by Ahsan and Kundur provides more information on this [4].

Video files are generally a collection of images and sounds, so most of the presented techniques on images and audio can be applied to video files too [3]. When information is hidden inside video the program or person hiding the information will usually use the DCT (Discrete Cosine Transform) method.

DCT works by slightly changing the each of the images in the video, only so much though so it's isn't noticeable by the human eye. To be more precise about how DCT works, DCT alters values of certain parts of the images, it usually rounds them up. The great advantages of video are the large amount of data that can be hidden inside and the fact that it is a moving stream of images and sounds. Therefore, any small but otherwise noticeable distortions might go by unobserved by humans because of the continuous flow of information [4, 6].

The statistical analysis method can be used against audio files too, since the LSB modification technique can be used on sounds too. Except for this, there are several other things that can be detected. High, inaudible frequencies can be scanned for information and odd distortions or patterns in the sounds might point out the existence of a secret message. Also, differences in pitch echo or background noise may raise suspicion. Like implementing steganography using video files as cover sources, the methods of detecting hidden information are also a combination of techniques used for images and audio files.

However, a different steganographic technique can be used that is especially effective when used in video films [2-7]. The usage of special code signs or gestures is very difficult to detect with a computer system. This method was used in the Vietnam War so prisoners of war could communicate messages secretly through the video films the enemy soldiers made to send to the home-front [4, 6].

Computational intelligence (CI) is the study of the design of intelligent agents which involves iterative development or learning. Computational intelligence includes neural networks, evolutionary computation (genetic algorithms and swarm intelligence) and other optimization algorithms. Techniques for handling uncertainty, such as bayesian, fuzzy logic, certainty theory fit into both categories. All these techniques use a mixture of rules and associated numerical values. Commonly, the implementation of computational intelligence, and their hybrids methods in steganalysis environment are collectively referred to as *intelligent steganalytic systems* (ISS). Nowadays, many researchers have applying CI on steganalysis environment. Most of their results have proven that the application of CI methods has given a great influence on steganalysis performance. They have also identified that the steganalysis environment can be divided into three (3) domains which are image steganalysis, audio steganalysis,

Currently, several methods for detecting image steganography with CI such as LSB embedding [8], spread spectrum steganography, and LSB matching , have been successfully steganalyzed [4].

*a) **Bayesian:***

On analyzing an image, one steganalysis approach [4] had proposed to estimate the hidden message based on a Bayesian framework. Message embedding in bit planes of an image is modelled as a binary symmetric channel. However, this method does not work for LSB embedding due to the lack of statistical structure in the bit plane.

*b) **Neural Network:***

A neural network [5] has been applied to analyse the possible occurrences of certain image pattern through histogram to detect the presence of data. They have used neural network approach to check for those discrepancy patterns and trains itself for better accuracy by automating the whole process from decomposition, signature searching, detection and elimination of the detection framework.

In another study, method based on neural network [6] has proposed to gather statistics features of images to identify the underlying hidden data. This study used neural network to analyze object digital image based on three different types of transformation, which are Domain Frequency Transform (DFT), Domain Coefficient Transform (DCT) and Domain Wavelet Transform (DWT). Meanwhile, the work of detection of wavelet domain information hiding techniques [7] has suggested statistical analysis on the texture of an image. Wavelet coefficients in each sub-band of wavelet transform are modelled as a Generalized Gaussian distribution (GGD) with two parameters. It appears that those parameters are a good measure of image features and can be used to discriminate stego-images from innocent images. Neural network is adopted to train these parameters to get the inherent characteristic of innocent and stego-images. Other study also claimed [8] that an artificial neural network capable of supervised learning results in the creation of a surprisingly reliable predictor of steganographic content, even with relatively small amounts of embedded data.

The interesting result is that clean colour images can be reliably distinguished from steganographically altered images based on texture alone, regardless of the embedding algorithm. Another study [7] that utilized an artificial neural network as the classifier in a blind steganalysis system. They found that an artificial neural network performs better in steganalysis than Bayes classifier due to its powerful learning capability. Thus, IEEE Computer Society [2] has suggested artificial neural network technology system (ANNTS). This technology is designed to recognize the digital files containing messages hidden by scanning an image or other file. ANNTS can accurately identify steganographic images between 85% and 100% of the time.

c) Genetic Algorithm:

Through a computational immune system (CIS) [2], a genetic algorithm approach has been used in blind steganography detection. They have developed CIS classifiers, which evolved through a genetic algorithm (GA), that is able to distinguish between clean and stego images by using statistics gathered from a wavelet decomposition. A further study [4] has investigated an artificial immune system (AIS) approach to novel steganography detection for digital images. AIS typically mimic portions of the

biological immune system (BIS) to provide a solution to a computational problem. Meanwhile, an application of genetic algorithm to optimal feature set selection in supervised learning using Support Vector Machine (SVM) for image steganalysis [3] has also presented. A genetic algorithm approach was used to optimize the feature set used by the classifier. Experimental results showed that the correct identification rates was as high as 98%, and as low as less than 2%.

d) Hybrid Method:

There are two studies have been done on hybrid technique of image steganalysis. These studies have proven that the effectiveness of the AI hybrid in the dynamic environment is as good as Dynamic Evolving Neural Fuzzy Inference System (DENFIS) which was presented by [7].

Recently, biological techniques become more and more popular, as they are applied to many kinds of applications, authentication protocols, biochemistry, and cryptography. One of the most interesting biology techniques is deoxyribo nucleic acid and using it in such domains. Hiding secret data in deoxyribo nucleic acid becomes an important and interesting research topic. Some researchers hide the secret data in transcribed deoxyribo nucleic acid, translated ribo nucleic acid regions, or active coding segments where it doesn't mention to modify the original sequence, but others hide data in non-transcribed deoxyribo nucleic acid, non-translated ribo nucleic acid regions, or active coding segments [9-11].

Unfortunately, these schemes either alter the functionalities or modify the original deoxyribo nucleic acid sequences. As a result, how to embed the secret data into the deoxyribo nucleic acid sequence without altering the functionalities and to have the original deoxyribo nucleic acid sequence be able to be retrieved is worthy of investigating.

Data Hiding Scheme adopts the reversible contrast mapping technique to hide the secret message in a DNA sequence, respectively. DNA sequence is composed of four nucleotides A, C, G, and T. Hence, we need to transform the representation format of the nucleotides such that the hiding techniques can be used to conceal the secret message in a DNA sequence.

To hide data, we need one of three things: the ability to insert a sequence containing
the data, to alter an existing innocuous sequence, or to find redundancy in an existing sequence and leveraging it to hide data.

First, each nucleotide symbol of the DNA sequence is converted into a binary string. A convenient strategy is to encode each nucleotide with two bits in alphabetical order. For example, the nucleotide A is encoded with ‘00’, C is encoded with ‘01’, G is encoded with ‘10’, and T is encoded with ‘11’. Next, several bits of the binary formatted DNA sequence are combined to form a bit string, and then the bit string is converted to a decimal integer. Each integer in the decimal formatted DNA sequence is called a word. Let w be the length of a bit string to form a word. Let us take a DNA sequence ‘AGTTCAGTA’ as an example. The binary format of the sequence is ‘001011110100101100’. Assume that w = 6, the first six bits ‘001011’ are converted to the decimal integer 11 because (001011)2= (11)10. Hence, the decimal format of the DNA sequence ‘AGTTCAGTA’ is ’11 52 44’. After that, the decimal formatted DNA sequence can be used to conceal the secret message. Reverse process will be done to find out the original message.

A different marking procedure is proposed in [5]. A map of transformed pairs and the sequence of LSBs for all non-transformed pairs are first collected. Then, the entire sequence LSB plane is overwritten by the payload and by the collected bit sequences. Thus, all the information needed to recover any original word pair is embedded into the pair itself or very close to it. In the case of cropping, except for the borders where some errors may appear, the original words of the cropped sequence are exactly recovered together with the embedded payload. For word pairing on row or column direction, there are no problems of synchronization. Some control codes should be inserted in the payload to validate watermark integrity.

CONCLUSIONS

Many different techniques exist and continue to be developed, while the ways of detecting hidden messages also advance quickly. Since detection can never give a guarantee of finding all hidden information, it can be used together with methods of defeating steganography, to minimize the chances of hidden communication taking place. Even then, perfect steganography, where the secret key will merely point out parts of a cover source which form the message, will pass undetected, because the cover source contains no information about the secret message at all. DNA as a storage medium is extremely effective. It is compact, biodegradable, and consumes very little energy. Today it is used to propagate species, encode protein synthesis, and solve complex computational problems. Who knows what it will do in the future? Recognizing this, techniques for hiding data to catalog, annotate, watermark, and/or encrypt information in this medium can have tremendous purpose. This paper proposes the original idea of hiding data in DNA.

REFERENCES

1. Johnson, N. F. and Jajodia, S. (1998). Exploring steganography: Seeing the unseen. Computer, 31(2): 26–34.
2. Saraju P. Mohant. Digital Watermarking: A Tutorial Review
3. Niels Provos, Peter Honeyman, Hide and Seek: Introduction to Steganography (2003).
4. F.A.P.Petitcolas, et al.,”Information Hiding – A Survey”, Proceedings of the IEEE, Vol.87, No.7, July 1999, pp.1062-1078.
5. B.Pfitzmann, ”Information Hiding Terminology”, Proc. of First Int. Workshop on Information Hiding, Cambridge, UK, May30-June1, 1996, Lecture notes in Computer Science, Vol.1174, Ross Anderson(Ed.), pp.347-350.
6. David Kahn, ”The History of Steganography”, Proc. of First Int. Workshop on Information Hiding,

Cambridge,UK, May30-June1 1996, Lecture notes in Computer Science, Vol.1174, Ross Anderson (Ed.), pp.1-7.

7. Roshidi Din, and Hanizan Shaker Hussain, and Sallehuddin Shuib, "Hiding secret messages in images: suitability of different image file types," *WSEAS TRANSACTIONS on COMPUTERS*, vol. 6(1), January 2006, pp. 127 -132.

http://www.worldses.org/journals/computers/computers-anuary2007.doc

8. G. Luo, X. Sun, L. Xiang, and J. Huang, "An evaluation scheme for steganalysis-proof ability of steganalysis algorithms", *International Conference on Intelligent Information Hiding and Multimedia Si Processing* (IIHMSP), vol. 2, 26-28 Nov 2007, pp. 126 - 129.

9. H.J. Shiu ,d, K.L. Ng , J.F. Fang , R.C.T. Lee , C.H. Huang, "Data hiding methods based upon DNA sequences", Information Sciences 180 (2010) 2196–2208.

10. A.L. Lehninger, D.L. Nelson, M.M. Cox, Principles of Biochemistry, Worth, New York, 2000.

11. A. Leier, C. Richter, W. Banzhaf, H. Rauhe, Cryptography with DNA binary strands, BioSystems 57 (2000) 13–22.

