Novel Text Steganography through Special Code Generation

Indradip Banerjee
Department of Computer Science and Engineering, University Institute of Technology, The University of Burdwan, Burdwan, India.
ibanerjee2001@yahoo.com

Souvik Bhattacharyya
Department of Computer Science and Engineering, University Institute of Technology, The University of Burdwan, Burdwan, India.
souvik.bha@gmail.com

Prof. Gautam Sanyal
Department of Computer Science and Engineering, National Institute of Technology, Durgapur, India.
gautam.sanyal@cse.nitdgp.ac.in

Abstract

Encrypted messages sending frequently draws the attention of third parties, perhaps causing attempts to break and reveal the original messages. Steganography is introduced to hide the existence of the communication by concealing a secret message in an appropriate carrier like text, image, audio or video. Steganography is of Greek origin and means "Covered or hidden writing". The carrier can be sent to a receiver without any one except the authenticated receiver only knows existence of the information. Steganography is often being used together with cryptography and offers an acceptable amount of privacy and security over the communication channel. It is an emerging area which is used for secured data transmission over any public media. In this paper, a novel text based steganography technique based on the use of indefinite articles 'a' or 'an' in conjunction with the non-specific or non-particular nouns in English language have been proposed. The authors also introduced a new code representation technique (SSCE - *Secret Steganography Code for Embedding*) at both ends in order to achieve high level of security. Before the embedding operation each character of the secret message has been converted to SSCE Value and then embeds to cover text. Finally stego text is formed and transmits to the receiver side. At the receiver side different reverse operation has been carried out to get back the original information.

Keywords: Steganography, SSCE (Secret Steganography Code for Embedding), Security, Cover Text, Stego Text

1. Introduction

Information hiding is a general term encompassing many sub disciplines. One of the most important sub disciplines is steganography [1], [8] as shown in Figure 1. Steganography, is derived from a work by Johannes Trithemus (1462-1516) entitled "Steganographia" and comes from the Greek (στεγανό-ς, γραφ-ειν) defined as "covered writing" [2]. It is an ancient art of hiding information in ways a message is hidden in an innocent-looking cover media so that will not arouse an eavesdropper's suspicion. Steganography differs from cryptography in the sense that where cryptography focuses on keeping the contents of a message secret, steganography focuses on keeping the existence of a message secret [3, 4].

Paper Identification Number: CI-3.4
This peer-reviewed paper has been published by the Pentagram Research Centre (P) Limited. Responsibility of contents of this paper rests upon the authors and not upon Pentagram Research Centre (P) Limited. Copies can be obtained from the company for a cost.

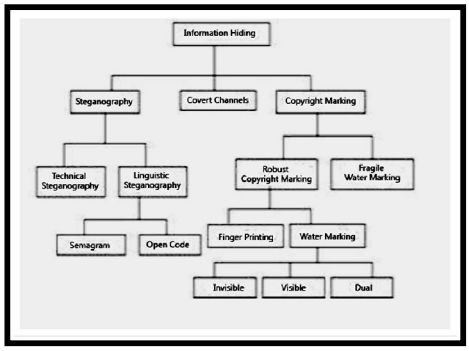


Figure 1: A Classification of Information Hiding techniques

Another form of information hiding is digital watermarking, which is the process that embeds data called a watermark, tag or label into a multimedia object such that watermark can be detected or extracted later to make an assertion about the object. The object may be an image, audio, video or text only [5],[6]. A covert channel could be defined as a communications channel that transfers some kind of information using a method originally not intended to transfer this kind of information. Observers are unaware that a covert message is being communicated. Only the sender and recipient of the message notice it. Steganography works have been carried out on different media like images, video clips, text, music and sound [13],[20].Among them image steganography is the most popular of the lot. In this method the secret message is embedded into an image as noise to it, which is nearly impossible to differentiate by human eyes [14], [15], [16]. In video steganography, same method may be used to embed a message [17], [18]. Audio steganography embeds the message into a cover audio file as noise at a frequency out of human hearing range [16]. One major category, perhaps the most difficult kind of steganography is text teganography or linguistic steganography because due to the lack of redundant information in a text compared to an image or audio. The text steganography is a method of using written natural language to conceal a secret message as defined by Chapman et al. [13]. Some Image steganographic algorithm with high security features has been presented in [19],[21],[22],[23],[24].

Figure 2 shows the basic text steganography mechanism. Firstly, a secret message (or an embedded data) will be concealed in a cover-text by applying an embedding algorithm to produce a stego-text. The stego-text will then be transmitted by a communication channel, e.g. Internet or mobile device to a receiver. For recovering the secret which sent by the sender, the receiver needs to use a recovering algorithm which is parameterised by a stego-key to extract the

secret message. A stego-key is used to control the hiding process so as to restrict detection and/or recovery of the embedded data to parties who know it [7],[8].

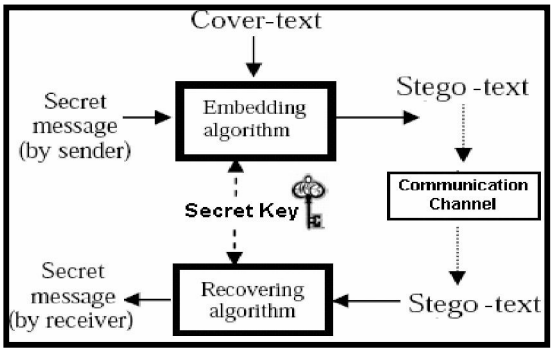


Figure 2: The Mechanism of Text Steganography

Text steganography can be classified in three basic categories [2] - format-based, random and statistical generation and linguistic method.

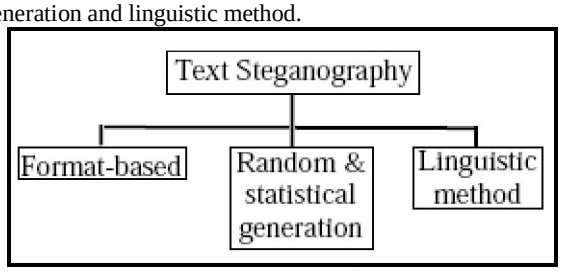


Figure 3: Three basic categories of text steganography

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. However, Bennett has stated that those format-based methods managed to trick most of the human eyes but it cannot trick once computer systems have been used.

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. A second approach to character generation is to take the statistical properties of word-length and letter frequency in order to create "words" (without lexical value) which will appear to have the same statistical properties as actual words in a given language. The hiding of information within word sequences, the actual dictionary items can be used to encode one or more bits of information per word using a codebook of mappings between lexical items and bit sequences, or words themselves can encode the hidden information.

The final category is linguistic method which specifically considers the linguistic properties of generated and modified text, frequently uses linguistic structure as a place for hidden messages. In fact, steganographic data can be hidden within the syntactic structure itself.

In this paper, a new approach of text steganography have been proposed based on the use of indefinite articles a or an in conjunction with the non-specific or non-particular nouns in English language .A new code representation method SSCE also have been proposed here to achieve high level of security.Before the embedding operation each character of the secret message has been encoded using SSCE Value and then embeds into cover text by the proposed text steganography method to form the stego text. This method is an integrated approach of new secret code generation along with a novel text based steganography method. Incorporating these two approaches in an embedding algorithm, a high embedding capacity of secret message can be achieved.

The proposed scheme has been inspired by the author's previous work [9] on a new approach of text steganography method by inserting extra blank space between the words of odd or even size of the cover according to the embedding sequence.

This paper is organized into the following sections. Section 2 describes the proposed model. Algorithms of various processes like embedding, extracting, encryption, decryption and GUI are discussed in Section 3. Analysis of the processes and results are discussed in Section 4. The last section draws the conclusion.

2. The Proposed Model

Figure 4 shows the block diagram of the proposed secret-key text steganographic model. The input messages can be in any digital form and are often treated as a bit stream. The input message is first encrypted using a new code generation technique **SSCE**. This encrypted message generates the secret key, (which may be called a message enabled key). Before embedding a checking has been done to find out whether the vowels and consonants are placed in the cover text as per the grammatical order, if not place it in proper order. For the improvement of security level, the SSCE code representation has been used to encrypt the message and then secret message has been embed to the cover text by inserting indefinite articles a or an in conjunction with the non-specific or non-particular nouns in English language based on the mapping information given in Fig 5 to form the stego text. At the receiver side other different reverse operation has been carried out to get back the original information.

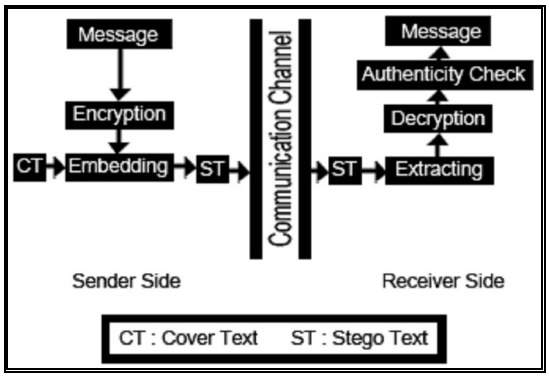


Figure 4: Proposed Text Steganography Model

Words		Bit Sequence
a	consonant	00
an	vowel	11
a	vowel	10
an	consonant	01

Figure 5: Mapping Technique

Solution Methodology

The proposed system consists of following two windows, one is the SENDER SIDE and the other is the RECEIVER SIDE. The user will be someone who is familiar with the process of information hiding and will have the knowledge of steganography systems. An encryption algorithm has been proposed prior to steganography for generation of encoded message. The user should be able to select a plain text message from a file, another text to be used as the carrier (cover text) and then use the proposed embedding method which will hide the encrypted message in the selected cover text and will form the stego text. The user at the receiver side should be able to extract the message from the stego text with the help of different reverse process in sequential manner to un-hide the message from the stego text. The GUI of the proposed solution has been shown in figure 6.

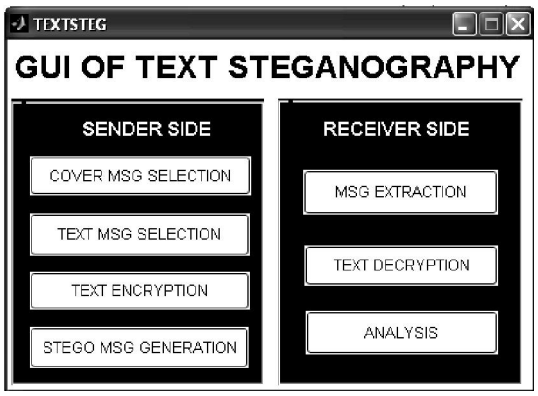


Figure 6: Solution Methodology

3. Algorithms

In this section, algorithms for different processes used both in the sender side and receiver side are described.

3.1 Algorithm for Message Encryption / Decryption

- Select the message and pick one by one character.
- Convert to its ASCII equivalent.
- Change ASCII code to our generated code from *SSCE* Table (Figure 7).
- Convert to its character equivalent.

3.2 Algorithm for Message Embedding

- Select the message and encrypt the message with *SSCE value.*
- Select the cover text to embed the message. Check whether the selected text is capable of embedding. If not possible repeat this step otherwise continue.
- Check the message sequence and pick first two bit sequence (MSG).
- Starting from the first word of the cover text (TX)
  - If MSG='11' then find out the word (an) from the TX and check whether the next word's first character is vowel.
  - Else If MSG='10' then find out the word (an) from the TX and check whether the next word's first character is vowel. Change (an) to (a).
  - Else If MSG='01' then find out the (a) from the TX and check whether the next word's first character is consonant. Change (a) to (an).
  - Else If MSG='00' then find out the word (a) from the TX and check whether the next word's first character is consonant.
- Repeat the above step for the remaining bit sequence of the message (two bit at a time).
- Save the embedding position in a separate file and encode it with *SSCE value* and send it to the receiver separately.

3.3 Algorithm for Message Extracting

- Select the newly generated text (stego text) after message embedding and their positions.
- Select the embedding position in TX
  - If there is word (an) and next word's first character is vowel, then MSG='11'
  - Else If there is word (a) and next word's first character is vowel, then MSG='10'
  - Else If there is word (an) and next word's first character is consonant, then MSG='01'
  - Else If there is word (a) and next word's first character is consonant, then MSG='00'

3.4 Algorithm for GUI

In this section the two algorithimic approach is discribed one for the function of the Sender Side and another for the Receiver Side.

3.3.1 Sender side

- Select the Cover Text from the set of Text files.
- Check whether the selected text is capable to do the embedding or not. If not possible then error.
- Select the message in text form.
- Encode the message through *SSCE value*.
- Embed the encrypted message in the cover text to form the stego text.
- End

3.3.2 Receiver side

- Receive the text with embedded message along with positions.
- Extract the encrypt form of message from the Stego Text.
- Decrypt the message with the help of the *SSCE value.*
- End

Secret Steganography Code for Embedding(SSCE) Table

ASCII	SSCE	ASCII	SSCE	ASCII	SSCE	ASCII	SSCE	ASCII	SSCE	ASCII	SSCE	ASCII	SSCE	ASCII	SSCE	ASCII	SSCE	ASCII	SSCE
10	1	1	26	2	52	3	78	4	104	5	130	6	156	7	181	8	206	9	231
20	2	11	27	12	53	13	79	14	105	15	131	16	157	17	182	18	207	19	232
30	3	21	28	22	54	23	80	24	106	25	132	26	158	27	183	28	208	29	233
40	4	31	29	32	55	33	81	34	107	35	133	36	159	37	184	38	209	39	234
50	5	41	30	42	56	43	82	44	108	45	134	46	160	47	185	48	210	49	235
60	6	51	31	52	57	53	83	54	109	55	135	56	161	57	186	58	211	59	236
70	7	61	32	62	58	63	84	64	110	65	136	66	162	67	187	68	212	69	237
80	8	71	33	72	59	73	85	74	111	75	137	76	163	77	188	78	213	79	238
90	9	81	34	82	60	83	86	84	112	85	138	86	164	87	189	88	214	89	239
100	10	91	35	92	61	93	87	94	113	95	139	96	165	97	190	98	215	99	240
110	11	101	36	102	62	103	88	104	114	105	140	106	166	107	191	108	216	109	241
120	12	111	37	112	63	113	89	114	115	115	141	116	167	117	192	118	217	119	242
130	13	121	38	122	64	123	90	124	116	125	142	126	168	127	193	128	218	129	243
140	14	131	39	132	65	133	91	134	117	135	143	136	169	137	194	138	219	139	244
150	15	141	40	142	66	143	92	144	118	145	144	146	170	147	195	148	220	149	245
160	16	151	41	152	67	153	93	154	119	155	145	156	171	157	196	158	221	159	246
170	17	161	42	162	68	163	94	164	120	165	146	166	172	167	197	168	222	169	247
180	18	171	43	172	69	173	95	174	121	175	147	176	173	177	198	178	223	179	248
190	19	181	44	182	70	183	96	184	122	185	148	186	174	187	199	188	224	189	249
200	20	191	45	192	71	193	97	194	123	195	149	196	175	197	200	198	225	199	250
210	21	201	46	202	72	203	98	204	124	205	150	206	176	207	201	208	226	209	251
220	22	211	47	212	73	213	99	214	125	215	151	216	177	217	202	218	227	219	252
230	23	221	48	222	74	223	100	224	126	225	152	226	178	227	203	228	228	229	253
240	24	231	49	232	75	233	101	234	127	235	153	236	179	237	204	238	229	239	254
250	25	241	50	242	76	243	102	244	128	245	154	246	180	247	205	248	230	249	255
		251	51	252	77	253	103	254	129	255	155								

Figure 7: SSCE Value Table

4. Analysis of the Results

There are mainly three aspects that should be taken into account when discussing the results of the proposed method of text steganography. They are security, capacity and robustness. The authors simulated the proposed system and the results are shown in the figures 8, 9, 10 and 11. This method satisfies both security aspects and hiding capacity requirements. It generates the stego text with minimum degradation which is not very revealing to people about the existence of any hidden data, maintaining its security to the eavesdroppers. Besides the security level has increased through the encoding of the secret message before embedding operation. This method hides two bit per word in the cover text which reflects the high embedding capacity of the system. Although the embedding capacity of the proposed method depends upon the embedding sequence along with the pattern of the cover text.

4. 1 Similarity Measure

For comparing the similarity between cover text and the stego text, the Jaro-Winkler distance for measuring similarity between two strings has been computed. The Jaro-Winkler distance [12] is a measure of similarity between two strings. It is a variant of the Jaro distance metric [10], [11] and mainly used in the area of record linkage [10] (duplicate detection). The higher the Jaro-Winkler distance for two strings is, the more similar the strings are. The score is normalized such that 0 equates to no similarity and 1 is an exact match. The Jaro distance metric states that given two strings s_1 and s_2 their distance d_j is $d_j = \frac{1}{3}\left[\frac{m}{s_1} + \frac{m}{s_2} + \frac{m-t}{m}\right]$, where m is the number of matching characters and t is the number of transpositions. Two characters from s_1 and s_2 respectively are considered matching only if they are not farther than $\left\lfloor \frac{\max(|s_1|,|s_2|)}{2} \right\rfloor - 1$. Each character of s_1 is compared with all its matching characters in s_2. The number of matching (but different sequence order) characters divided by two defines the number of transpositions. The Jaro score of comparing cover text and stego text is 0.9022, which means they are closely similar. Besides comparison through histogram technique has been done. It has been observed that the histogram of the cover text and the stego text is almost identical.

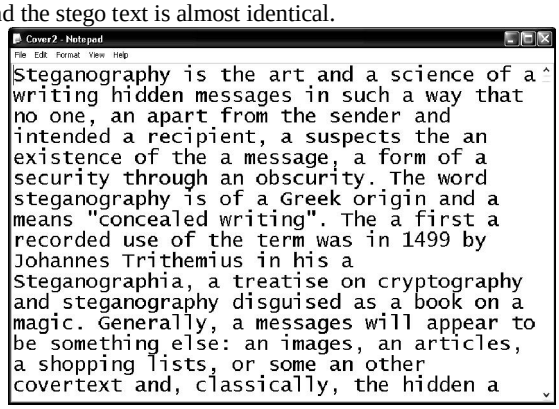


Figure 8: Cover Text

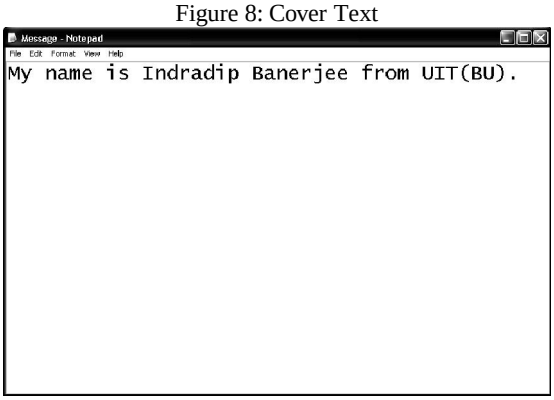


Figure 9: Secret Message

Figure 10: Encrypted Secret Message

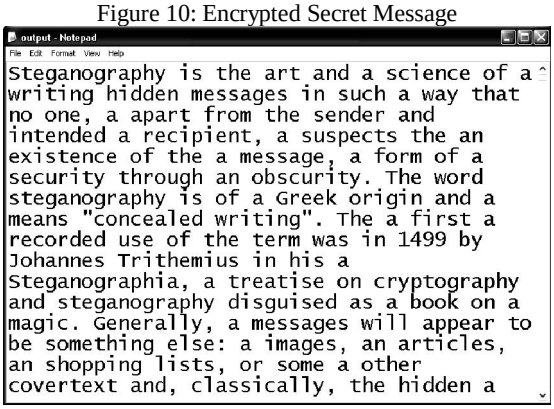


Figure 11: Stego Text

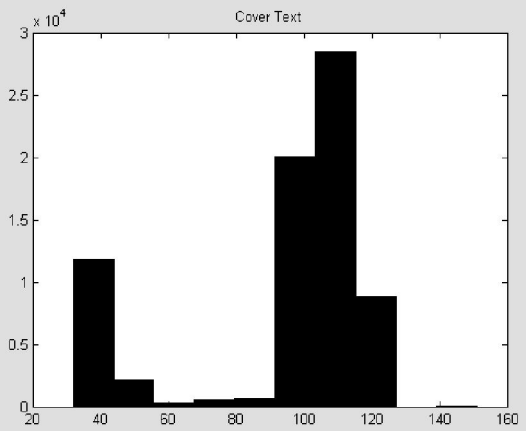


Figure 12: Histogram of Cover Text

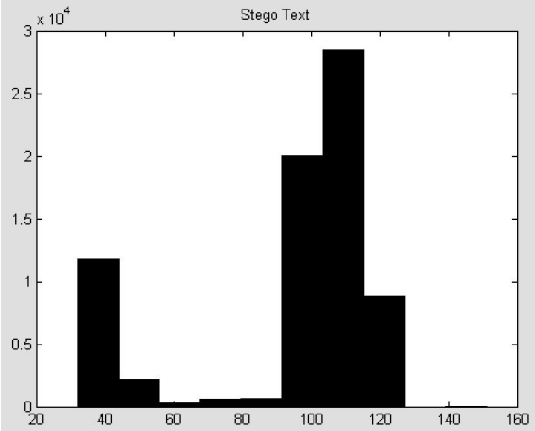


Figure 13: Histogram of Stego Text

5. Conclusion

In this paper the authors presented a new approach of text steganography method by changing the (a), (an), vowel & consonant rule of the cover text according to the embedding sequence and also in some cases only the (a)-consonant and (an)-vowel between the words of the original cover text may be used as mapping the embedding sequence. This property generates the stego text with minimum degradation. This property enables the method to avoid the steganalysis also. The new BASE Value (SSCE - Secret Steganography Code for Embedding) has been used to 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. The future work should be focused to improve the capacity of the embedding scheme by incorporating some compression technique on the secret message.

6. References

[1] Dr. Mohammed Al-Mualla and Prof. Hussain Al-Ahmad, "Information Hiding: Steganography and Watermarking". [Online]. Available: http://www.emirates.org/ieee /information_hiding.pdf.

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

[3] Ross J. Anderson and Fabien A.P. Petitcolas, "On the limits of steganography," IEEE Journal on Selected Areas in Communications (J-SAC), Special Issue on Copyright & Privacy Protection, vol. 16 no. 4, pp 474-481, May 1998.

[4] T Mrkel,JHP Eloff and MS Olivier ."An Overview of Image Steganography,"in proceedings of the fifth annual Information Security South Africa Conference, 2005.

[5] Digital Watermarking: A Tutorial Review S.P.Mohanty, 1999.

[6] N. F. Johnson, S. Jajodia, "Exploring Steganography: Seeing the Unseen," IEEE Computer, February 1998, pp.26−34.

[7] "Spy Gadgets in World War II: Microdots", 2007. [Online]. Available: http://www.mi5.gov.uk/output/Page303.html [Accessed Feb. 15, 2008].

[8] Fabien A. P. Petitcolas, Ross J. Anderson and Markus G. Kuhn. "Information Hiding – A Survey", Proceedings of the IEEE, special issue on protection of multimedia content, July 1999, pp. 1062 – 1078.

[9] Souvik Bhattacharyya, Indradip Banerjee and Gautam Sanyal. Design and implementation of a secure text based steganography model. In Proceedings of 9th annual Conference on Security and Management (SAM) under The 2010 World Congress in Computer Science,Computer Engineering and Applied Computing(WorldComp 2010), LasVegas,USA, July 12-15,2010.

[10] M. A. Jaro. Advances in record linking methodology as applied to the 1985 census of tampa florida. Journal of the American Statistical Society. 84:414–420, 1989.

[11] M. A. Jaro. Probabilistic linkage of large public health data file. Statistics in Medicine 14 (5-7)., pages 491–498, 1995.

[12] W. E. Winkler. The state of record linkage and current research problems. Statistics of Income Division, Internal Revenue Service Publication R99/04., 1999.

[13] Kran Bailey Kevin Curran. An evaluation of image based steganography methods. International Journal of Digital Evidence,Fall 2003, 2003.

[14] D. Kahn. The Codebreakers - the comprehensive history of secret communication from ancient times to the Internet. Scribner, 1996.

[15] Z. Duric N. F. Johnson and S. Jajodia. Information Hiding: Steganography and Digital Watermarking - Attacks and Countermeasures. Kluwer Academic, 2001.

[16] N.F. Maxemchuk J.T. Brassil, S. Low and L. O.Gorman. Electronic marking and identification techniques to discourage document copying. IEEE Journal on Selected Areas in Communications, 13:1495–1504, 1995.

[17] G. Doerr and J.L. Dugelay. A guide tour of video watermarking. Signal Processing: Image Communication, 18:263–282, 2003.

[18] G. Doerr and J.L. Dugelay. Security pitfalls of frameby-frame approaches to video watermarking. IEEE Transactions on Signal Processing, Supplement on Secure Media, 52:2955–2964, 2004.

[19] Souvik Bhattacharyya, Indradip Banerjee and Gautam Sanyal. "Implementation of a Novel Text Based Steganography Model" Proceeding of "National Conference on Computing and Systems (NACCS)", 29/01/2010, Dept. of Computer Science, The University of Burdwan.

[20] T Mrkel,JHP Eloff and MS Olivier ."An Overview of Image Steganography,"in proceedings of the fifth annual Information Security South Africa Conference ,2005.

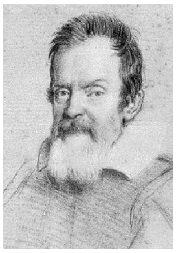
[21] "Study of Secure Steganography model" by Souvik Bhattacharyya and Gautam Sanyal at the proceedings of "International Conference on Advanced Computing & Communication Technologies (ICACCT-2008),Nov, 2008, Panipat, India"

[22] **"**An Image based Steganography model for promoting Global Cyber Security" by Souvik Bhattacharyya and Gautam Sanyal at the proceedings of "International Conference on Systemics,Cybernetics and Informatics (ICSCI-2009),Jan, 09,Hyderabad,India."

[23] **"**Implementation and Design of an Image based Steganographic model" by Souvik Bhattacharyya and Gautam Sanyal at the proceedings of " IEEE International Advance Computing Conference "(IACC-2009)"

[24] A Novel Approach to Develop a Secure Image based Steganographic Model using Integer Wavelet Transform" at the proceedings of International Conference on Recent Trends in Information, Telecommunication and Computing (ITC 2010)" by Souvik Bhattacharyya, Avinash Prasad Kshitij and Gautam Sanyal. (Indexed by IEEE Computer Society).

Galileo Galilei



Galileo Galilei (Italian pronunciation: [ɡaliˈlɛo ɡaliˈlɛi]; 15 February 1564 – 8 January 1642), commonly known as **Galileo**, was an Italian physicist, mathematician, astronomer and philosopher who played a major role in the Scientific Revolution. His achievements include improvements to the telescope and consequent astronomical observations, and support for Copernicanism. **Galileo has been called the "father of modern observational astronomy",[6] the "father of modern physics", the "father of science", and "the Father of Modern Science"**. Stephen Hawking says, "Galileo, perhaps more than any other single person, was responsible for the birth of modern science."

The motion of uniformly accelerated objects, taught in nearly all high school and introductory college physics courses, was studied by Galileo as the subject of kinematics. His contributions to observational astronomy include the telescopic confirmation of the phases of Venus, the discovery of the four largest satellites of Jupiter (named the Galilean moons in his honour), and the observation and analysis of sunspots. Galileo also worked in applied science and technology, inventing an improved military compass and other instruments.

Galileo's championing of Copernicanism was controversial within his lifetime, when a large majority of philosophers and astronomers still subscribed to the geocentric view that the Earth is at the centre of the universe. After 1610, when he began publicly supporting the heliocentric view, which placed the Sun at the centre of the universe, he met with bitter opposition from some philosophers and clerics, and two of the latter eventually denounced him to the Roman Inquisition early in 1615. In February 1616, although he had been cleared of any offence, the Catholic Church nevertheless condemned heliocentrism as "false and contrary to Scripture",[10] and Galileo was warned to abandon his support for it—which he promised to do. When he later defended his views in his most famous work, *Dialogue Concerning the Two Chief World Systems*, published in 1632, he was tried by the Inquisition, found "vehemently suspect of heresy", forced to recant, and spent the rest of his life under house arrest.

For more detail visit
http://en.wikipedia.org/wiki/Galileo_Galilei
Editorial team, ICSCI-2011