A Novel Approach of Bulk Data Hiding using Text Steganography

Shivani*, Virendra Kumar Yadav*, Saumya Batham*

*Department of Computer Science and Engineering, ABES Engineering College, Ghaziabad, 201009, India

Abstract

Steganography enables individual to attain privacy by providing efficient techniques to hide data. The term steganography derives its existence from 'Johannes Trithemus' literary work titled "Steganographia". It can be applied to several digital medium such as text, image, audio and video. Research work is going on to improve the performance metrics so that a large amount of data can be hidden by consuming less time and memory overhead. Text steganography can be achieved by using text file as a cover. Text file consumes less storage space and can be accessed with low bandwidth. Zero Distortion Technique [14] is applied on text along with abbreviation method for achieving optimal results. For encryption Indexed Based Chaotic Sequence approach has been used [11]. The results obtained by the trio effect are evaluated on the basis of performance metrics like hiding capacity, time consumption and results compared with other text steganography techniques.

1. Introduction

Sending confidential data over internet is risky task. Primary concern is to protect data from intruders. As data is transmitted via digital medium, it has certain disadvantages like tampering, easy to access, illegal use, copyright violation etc. Information hiding is a branch of computer science which deals with hiding data, object or function details [4]. Encapsulation and information hiding can be used interchangeably [3]. For this purpose several techniques can be used out of which steganography and cryptography are preferred, steganography is combined with cryptography, combination is termed as metamorphic cryptography [1]. In 1983 cryptographer ‘Gustavus Simmons’ published first work on digital steganography, which is known as 'prisoner's problem'.

Cryptography provides security by using encryption techniques but it creates suspicion in the mind of third party who is intercepting data [10]. To avoid this suspicion, steganography is used (which is an art of “concealed
writing”). It is used to hide confidential and sensitive data inside a digital medium. Steganography is broadly classified into four categories i.e. text, image, audio and video steganography. Among these, text steganography has some limitations such as lack of redundancy, sensitive to changes etc. In other digital mediums like image, sound and video redundancy is higher [3]. A steganalyst is a person who is skilled in detecting hidden messages from a digital media. Steganalysis is the study of extracting hidden messages from the digital medium which is done with the help of steganography [3].The text file which is used for hiding data is known as cover text and the result obtained after embedding the data in cover text is known as stego text. There are three inputs to steganography algorithm i.e. cover text, algorithm and text to hide. It is performed in two phases: First, data is embedded in cover text and then extracted from the stego text.

The work is categorized as follow: Section 2 contains related work contributed in the field of text steganography. Section 3 contains proposed work i.e. zero distortion technique on text. Section 4 contains experimental results and comparison tables.

2. Related Work

Zero distortion technique is a novel approach for performing steganography [14][15], in which digital medium is used as a reference and secret data is hidden in such a manner so that there is no distortion in the cover medium. Zero distortion technique has been applied to gray as well as on color images[14][15]. In color images large amount of data can be hidden as compared to gray images because color image contains RGB bands. Advantages of using images is that changes are almost imperceptible to human vision and disadvantage is stego image is send to receiver’s end so chances of attack are more. This technique is applied on text i.e. to hide secret text within cover text. Cryptography is combined with steganography to increase security. For encryption purpose Indexed Based Chaotic Sequence [11][12] has been used.

Text steganography is more complicated than image steganography. It co-relates with linguistic steganography (data is hidden in cover text). There are several methods for performing text steganography like:

- Syntactic Method: It deals with syntax or format of text to hide data. In this method punctuation is inserted in cover text to hide data such as full stop (.), comma (,) etc at correct place [2].
- Semantic Method: It uses synonym of the word for hiding data and substitute target word with its synonym. But sometimes it may alter the actual meaning of the text file [5].

| Word | Synonym |
|------|---------|
| Big  | Large   |
| Small| Little  |
| Smart| Intelligent |

- Abbreviation or acronym Method: Mohammad Sirali-Shahreza and M.Hassan Shirali-Shahreza from Iran proposed a method by substituting word with its acronym [7]. A very less amount of data can be hidden in this method.

| Acronym | Translation |
|---------|-------------|
| 2day    | Today       |
| Rslt    | Result      |
| ASAP    | As Soon As Possible |

- Word Spelling: Different countries have different vocabulary. Mohammad Shirali-Shahreza in his research paper [8] has elaborated that same words spelled differently in British English and American English can be used to hide data.
Table 3. Word Spelling method

| American English | British English |
|------------------|-----------------|
| Center           | Centre          |
| Tire             | Tyre            |
| Defense          | Defence         |

- Line shifting: In this method lines of text are vertically shifted up to an extent (for example 1/400 inch up or down). It is used widely for printed text [6].
- Word Shifting: This method varies horizontal distance between words by inserting white spaces [8]. This method is time consuming as well as hidden text is perceptible to human vision.
- Feature Coding: As discussed in [17] [18] syntax is altered to introduce redundancy. Letters are stretched or shortened with respect to their dimension. It also deals with different attributes of text like color etc to hide data.
- Random Character Sequence: In this method a random string with single letters are generated containing same letters as the cover [19]. Two groups are formed for hiding purpose. If 0 is to hidden then group A is used and if 1 is to be hidden then group B is used.
- Inter sentence space: This method [19] uses space between sentences to hide message. The primary disadvantage is that very less amount of data can be hidden.
- Method based on curves: This method [19] deals with the shape of letters i.e. two groups are on the basis of their structure. Group A contains letters having curves and group B contains letters without curves.

Table 4. Method based on curves

| Group | Group Name       | Bit | Letters                  |
|-------|------------------|-----|--------------------------|
| A     | With curves      | 0   | B, C, D, G, J, O, P, Q, R, S, U |
| A     | Without curves   | 1   | A, E, F, H, I, K, L, M, N, T, V, W, X, Y, Z |

Approach Based on Vertical Straight Line: In this method [19] groups are formed on the basis of vertical straight line appearing. Group A contains words having vertical line and Group B contains words not having vertical straight line.

Table 5. Approach based on vertical straight line

| Group | Group Name       | Bit | Letters                  |
|-------|------------------|-----|--------------------------|
| A     | With curves      | 0   | B, C, D, G, J, O, P, Q, R, S, U |
| A     | Without curves   | 1   | A, E, F, H, I, K, L, M, N, T, V, W, X, Y, Z |

- Quadruple categorization: In this method [19] four groups are formed on the basis on if there is a curve, middle horizontal straight line, single straight vertical line or multiple straight vertical lines.
Table 6. Quadruple Categorization

| Group | Group Name                                      | Bit | Letters                        |
|-------|-------------------------------------------------|-----|--------------------------------|
| A     | Curved Letters                                  | 00  | C, D, G, O, Q, S, U           |
| B     | Letters with middle Horizontal straight line     | 01  | A, B, E, F, H, P, R           |
| C     | Letters with one vertical straight line          | 10  | I, J, K, L, T, Y             |
| D     | Letters with diagonal line                      | 11  | M, N, V, W, X, Z             |

3. Proposed Work

3.1. Zero Distortion Technique on Text

Zero Distortion Technique [14][15] is a novel technique which results zero distortion in the cover image i.e. no changes occurs as the digital medium is taken as a reference only. Other techniques tend to change the actual content of the cover and compromise its originality and changes occurred in digital medium are perceptible to human vision. Zero Distortion Technique mainly works on locations. If bit values matched then location of the cover binary bit is saved in a matrix known as matrix of locations. This matrix contains the actual information from which secret text can be constructed. To increase the security, Indexed Based Chaotic Sequence is used which encrypts the matrix of locations.

3.2. Text Steganography Method

For text steganography abbreviation method is used [7]. This method is used for hiding large amount of data. It includes a database of words which contains acronyms and their translation. Such types of words are very popular and used in our daily life in social networking medium like facebook, whatsapp, messages, hike, yahoo messenger etc. Such words are also known as slangs. For example See is replaced with C, You is written as U etc. This technique enables to hide large amount of data and also time consuming.

3.3. Encryption Techniques

Encryption is a mechanism to encode data so that only intended users can read data. Its origin is derived from a Greek word meaning secret writing [10].

It enhances security of the system. For encryption Indexed Based Chaotic Sequence [11][12] is used because it provides non linearity and randomness. Its main characteristic is, it generates a matrix with value ranging between 0-1. For initial values it shows constant behaviour but in the interval of 3.54 - 4 its values shows chaotic behaviour. Chaos word is used to define a situation of confusion [12].

![Bifurcation diagram for logistic map of μ](image-url)
From bifurcation diagram it can be concluded that in the range of 0-1 there is a constant line (no change occurs) and graph shows sudden increase in the range of 1-3. But randomness is observed in the range of 3.54 – 4. Indexed Based Chaotic Sequence possess a unique property i.e. if the value of $X_0$ and $\mu$ differs from the value fixed at sender’s end then secret data can’t be extracted. So only intended users can extract data from the stego text.

It is known as Indexed Based Chaotic Sequence as it does encryption based on the index values. This technique generates values in the range of 0-1 for a given matrix by using formula.

\[
X_{n+1} = \mu * X_n * (1-X_n)
\]  

(1)

The values of $X_0$ and $\mu$ are set at sender’s end and pass along with the Chaotic Sequence matrix to the other end. The range of $X_0$ is 0-1 and $\mu$ is 3.54 – 4. These values are constant. Then values and indexes are sorted in ascending order.

\[
[value \ index] = \text{sort(matrix_containg_values_generated)}
\]  

(2)

After sorting, indexes of the sorted values are retrieved. These indexes are the encrypted values which are passed to the other end along with stego text.

3.4. Techniques for embedding and extracting text using Zero Distortion Technique on text

In this paper we are extending our previous work [14],[15]. Performance metrics are computed generated which shows better results. For applying steganography on text, abbreviation method is used[19]. The major limitation of abbreviation method is less amount of data can be hidden but when combined with Zero Distortion Technique large amount of secret text can be hidden. Abbreviation method is used to reduce the size of the secret text. So if a large input secret text is taken, first size of the secret text will be reduced then zero distortion technique is being applied. So this technique overcomes the limitation of abbreviation method.

3.4.1. Algorithm for embedding text data into cover text

Secret text file will be embedded in the cover text file using this algorithm.

**Input:** A cover text file ($T_C$), a secret text ($T_S$).

**Output:** Chaotic Sequence Matrix ($C_{OS}$)

**Step 1)** Read the cover text file ($T_C$) and secret text file ($T_S$).

**Step 2)** Calculate the length of the cover text file ($T_C$).

// For calculation purpose

**Step 3)** Conversion of cover text file ($T_C$) into ASCII and then into binary format.

**Step 4)** Reduce the size of the secret text file ($T_S$) using abbreviation method, obtained reduced file is ($T_{RS}$)

// Reduction is performed so that an large amount of data can be hidden

// Database of abbreviated words has already been created

**Step 5)** Conversion of reduced secret text file ($T_{RS}$) into ASCII and then into binary format.

**Step 6)** For all $i=1$ to 8 repeat steps 7 to 10

**Step 7)** For $j=1$ to rows_of_cover_text_file

**Step 8)** Matching of bits of cover text file ($T_C$) and reduced secret text file ($T_{RS}$) is performed

**Step 9)** If matched then save the location of cover text file ($T_C$) in matrix of locations ($L_{OS}$) of dimensionality n rows and 8 column. Increase the value of location and count variable by 1

// n is the length of the characters in reduced secret text ($T_{RS}$)

// Count variable is used to check whether complete data has been hidden or not

**Step 10)** Else increase location only.

// resulting is the matrix of locations ($L_{OS}$). Now perform encryption using Indexed Based Chaotic Sequence

**Step 11)** Save the dimensions of matrix of locations ($L_{OS}$) in variable m x n
Step 12) Initialize the value of $\mu$ in the range 3.5-4 and $X_0$ in 0-1
Step 13) For $k = 1$ to $m \times n$
Step 14) Apply Indexed Based Chaotic Sequence using formula
\[ X_{n+1} = \mu \times X_n \times (1-X_n) \] (1)
Step 15) Sort the values generated by the above formula and extract the indexes of the sequence.
Step 16) Reshape the indexes into $m \times n$ matrix same as the dimension of matrix of locations ($L_{0S}$).
//This is the Chaotic Sequence i.e. $C_{0S}$ which will be passed to the decryption end.
Step 17) If count variable is equal to the length of the reduced secret text. Then message displays “Secret data has been embedded successfully”.
Step 18) Else message displays “Text has not been embedded .Size of the cover text file is small”.

3.4.2. Algorithm for extracting data from the cover text file

INPUT: Chaotic Sequence ($C_{0S}$), value of $X_0$ and $\mu$, cover text file ($T_C$).
OUTPUT: Secret text file ($T_S$)

Step 1) Extract the matrix of locations ($L_{0S}$) from the Chaotic Sequence by using the same formula.
//Decryption has been performed
\[ X_{n+1} = \mu \times X_n \times (1-X_n) \] (1)
Step 2) Conversion of cover text file ($T_C$) into ASCII and then into binary format.
Step 3) Calculate the length of the cover text file
Step 4) Match the values of matrix of locations ($L_{0S}$) and the matrix of cover text. If matched then put the values of the location from the matrix of cover text.
Step 5) Else increase the location of the cover text.
//matrix containing secret text has been created in binary format
Step 6) Conversion of matrix from binary to ASCII format.
Step 7) Conversion of ASCII format to character format.
Step 8) Display the secret text ($T_S$).

3.5. Execution of Algorithms

3.5.1. Execution of algorithm for embedding secret text

Execution involves two phases: Firstly secret data is embedded and matrix of locations is formed ($L_{0S}$). Secondly encryption is performed using Indexed Based Chaotic Sequence.

- **Embedding data into cover text file**

We are taking a cover_data.txt and a secret_data.txt as cover and secret input file respectively. Algorithm proceeds as follows:

Step 1) Read the cover text file ($T_C$) and secret text file ($T_S$).

![Fig. 2. Cover text file(cover_data.txt)](image)

![Fig. 3. Secret text file(secret_data.txt)](image)

Step 2) Calculate the length of the cover text file ($T_C$).
\[ \text{Length}=3564 \text{ bytes} \]

//size of sample text
Step 3) Conversion of cover text file \((T_C)\) into ASCII and then into binary format.

![ASCII format of cover text file](image1)

![Binary format of cover text file](image2)

Step 4) Reduce the size of the secret text file \((T_S)\) using abbreviation method. The size of secret text file is 1000 bytes which are reduced to 900 bytes by using abbreviation method. File can be further reduced if large database is used.

![Reduced secret text file](image3)

![Reduced secret text file](image4)

STEP 5) Conversion of reduced secret text file \((T_{RS})\) into ASCII and then into binary format.

![ASCII value of reduced secret text file](image5)

![Binary format of reduced secret text file](image6)

Step 6) Matching of bits of cover text file \((T_C)\) and reduced secret text file \((T_{RS})\) is performed. Matrix of locations \((L_{0S})\) is generated.
Encryption is performed using Indexed Based Chaotic Sequence

Step 7) Apply encryption by using Indexed Based Chaotic Sequence using formula

\[ X_{n+1} = \mu \times X_n \times (1 - X_n) \]  

Random values generated for matrix of locations (L0S) by the above formula are in the range of 0-1.
Sort the values of the above matrix in ascending order.

Extract the indexes associated with sorted values
This is the resultant Chaotic Sequence matrix which will be passed to the other end along with value of $X_0$ and $\mu$ that are constant values fixed at sender’s end for performing encryption.

Step 8) If message has been hidden successfully a message will be displayed which is calculated on the basis of count variable. If count variable is less than secret text than complete message has not been hidden i.e. a large cover text is needed. Display the stego cover text (if needed), no changes will be visible in cover text as it has been taken only as a reference i.e. zero distortion the cover text file.

4. EXPERIMENTAL RESULTS

| Text Steganography Approach | Message Text size(Bytes) | Cover Text size(Bytes) | No. of Bytes can hide(Bytes) | Time Overhead (ms) |
|-----------------------------|--------------------------|------------------------|-------------------------------|--------------------|
| Zero Distortion Technique + Abbreviation Method | 500 | 2640 | 800 | 20006 |
| Method based on curve | 800 | 2640 | 172 | 37,996 |
| Method based on vertical straight line | 800 | 2640 | 161 | 27,593 |
| Quadruple Categorization | 800 | 2640 | 145 | 26,562 |
| Inter Word space | 800 | 2640 | 58 | 20,825 |
| Feature Coding | 800 | 2640 | 65 | 19,160 |
| Random Character | 800 | 2640 | 48 | 31,702 |

| Text Steganography Approach | Message Text size(Bytes) | Cover Text size(Bytes) | No. of Bytes can hide(Bytes) | Time Overhead (ms) |
|-----------------------------|--------------------------|------------------------|-------------------------------|--------------------|
| Zero Distortion Technique + Abbreviation Method | 1000 | 3564 | 1000 | 27,802 |
| Method based on curve | 1000 | 3564 | 252 | 32,999 |
| Method based on vertical straight line | 1000 | 3564 | 220 | 30,017 |
| Quadruple Categorization | 1000 | 3564 | 205 | 32,269 |
| Inter Word space | 1000 | 3564 | 79 | 21,926 |
| Feature Coding | 1000 | 3564 | 70 | 17,950 |
| Random Character Sequence | 1000 | 3564 | 78 | 32,814 |
Conclusion

In simple words, steganography is an art of hiding message in such a fashion, presence of hidden message cannot be conceived. Several researchers are working in this area to improve the efficiency of steganographic algorithms. Proposed algorithm in this paper improves the data hiding capacity. With the help of proposed algorithm users can hide more amount of data without producing any distortion in the cover image. It means changes reflected are almost negligible.

Acknowledgement

I would like to thanks my parents, grandparents whose constant support motivates me in my hard time when I was developing this technique.

References

1. Thomas Leonitin Philjon. J, Venkateshvara Rao. N , “Metamorphic Cryptography - A Paradox between Cryptography and steganography Using Dynamic Encryption”, IEEE -International Conference on Recent Trends in Information Technology, ICRTIT 2011
2. W.Bender, D.Gruhl,N.Morimoto,A.Lu, “Technique For Data Hiding”Ibm Systems Journal, Vol.35, Issue 3&4, 1996, Pp.316-336.
3. Moerland, T. “Steganography and Steganalysis”.Leiden Institute of Advanced Computing Science, www.liacs.nl/home/ moerl/privtech.pdf
4. 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, pp. 1062-1078, July 1999.
5. M. H. Shirali-Shahreza, M. Shirali-Shahreza, “A new approach to persian/arabic text steganography,” Proc. 5th Int. Conf. Computer and Information Science, Washington, 2006, pp.310-315.
6. Y. Kim,K.Moon,And LOh,”A Text Watermarking Algorithm Based On Word Classification And Inter-Word Space Statistiocs",Proceeding Of The Seventh International Conference On Document Analysis And Recognition,2003,Pp775-779.
7. Mohammad Sirali-Shahreza, M. Hassan Shirali-Shahreza, Text Steganography in Chat, 1-4244-1007/07 © 2007 IEEE
8. Mohammad Shirali-Shahreza, Text Steganography by Changing Words Spelling, ISBN 978-89-5519-136-3, Feb. 17- 20, 2008, ICACT 2008
9. L. Robert, T.Shannugapriya, “A study on digital watermarking techniques,” Int. Journal of Recent Trends in Engineering, vol.1, no.2, May 2009.
10. William Stalling,Cryptography and network security:Principles and Practices (4th edition), Prentice 2006, ISBN – 978-81-775-8774-6
11. S.Batham, A. Acharya, V.K. Yadav, R.Paul.”A New Video Encryption Algorithm Based on Indexed Based Chaotic Sequence”.CONFLUENCE-2013, IET digital library.http://digital-library.theiet.org/content/conferences/10.1049/cp.2013.2307
12. Aradhana Soni and Anuja Kumar Acharya , ”A Novel Image Encryption Approach Using An Index Based Chaos And DNA Encoding and Its Performance Analysis” , IJCA (0975-8887) Volume 47-No. 23,June 2012.
13. Anuja Kumar Acharya , “Image Encryption Using A Chaos Base Encryption Algorithm”, International Conference On Communication,Computing and Security( ICCC 2011)
14. Shivani , Yadav.V , Batham.S , “An Approach to Image Steganography using Strength of Indexed Based Chaotic Sequence”, SSCC- 2014 (Springer)
15. Shivani , Yadav.V , Batham.S , ”Zero Distortion Technique: An approach to image steganography on color images using strength of Chaotic Sequence”,accepted in ACM ICTCS 2014.
16. Beenish Mehboob,Rashid Aziz Faruqui , “A Steganography Implementation”, IEEE 2008.
17. M. Shirali-Shahreza, M. H. Shirali-Shahreza, “Text steganography in SMS,” Proc. Int. Conf. Convergence Information Technology, Washington, 2007, pp. 2260-2265.
18. K.F. Rafat, “Enhanced text steganography in SMS,” Proc. of the 2nd Int. Conf. Computer, Control and Communication, Karachi, 2009, pp.1-6.
19. Shraddha Dulera, Devesh Jinwala and Aroop Dasgupta, “EXPERIMENTING WITH THE NOVEL APPROACHES IN TEXT STEGANOGRAPHY”, International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.6, November 2011