A New Approach of Text Steganography Based on Mathematical Model of Number System

Kunal Kumar Mandal
Dept. of Computer Applications,
National Institute of Technology, Durgapur
West Bengal, India
kunal1984@gmail.com

Angshuman Jana
Dept. of Computer Applications,
National Institute of Technology, Durgapur
West Bengal, India
janaangshuman@gmail.com

Vineet Agarwal
Vice President
FX Derivatives Trading
Deutsche Bank AG, Singapore
agar.vineet@gmail.com

Abstract — In today's world the art of sending & displaying hidden information especially in public places, has received much attention and faced many challenges. Therefore, different methods have been proposed so far for hiding information in various cover media. Steganography is the method to hide a message inside another message without drawing any suspicion to others so that the message can only be detected by its intended recipient. With other steganography methods such as Image, Audio, Video, a number of text steganography algorithms have been introduced. This paper presents a new approach for steganography where introduced a new approach of using a new kind of number system for Steganography through a directed weighted graph cover media.

Keywords— Steganography, Number System, Directed Graph, Information hiding.

I. INTRODUCTION

The growing possibilities of modern communication need special means of security especially on computer network. The network security is becoming more important as the number of data being exchanged on the Internet increases. Therefore, the confidentiality and data integrity are required to protect against unauthorized access. This has resulted in an explosive growth of the field of information hiding. Various methods like Cryptography, steganography, coding and soon have been used for this purpose. However, during recent years, steganography perhaps has attracted more attention than others. Steganography is the art and science of hiding information such that its presence cannot be detected [1]. A secret message is encoded in such a manner that the existence of the information is hidden. Paired with existing communication methods, steganography can be used to carry out hidden exchanges. In implementing steganography,the main objective is to hide the information under cover media so that outsiders may not discover the information contained in the said frame. This is the major distinction between steganography and other methods of hidden exchange of information. For example, in cryptography method, people become aware of the existence of information by observing coded information although they will be unable to comprehend the information. However, in steganography, nobody will understand the existence of information in the resources [1]. Most of steganography works have been carried out on pictures [2, 3], video clips [4, 5], music and sounds [6]. Text steganography is the most difficult kind of steganography this is due largely to the relative lack of redundant information in a text file as compared with a picture or a sound file [8]. This paper provides new method for hiding information using weighted directed graph based on a new number system. The work done is basically representing any number in a different form. The primary difference between this representation and a normal decimal representation is that in this system we can represent a very large no. by more than 1 smaller numbers. Say if a no. is of 6 digits then it can be represented by two nos. of 3or 4 digits each. Thus we do not get any advantage in terms of space that is the no. of digits used is not less than what is used in normal decimal but the advantage is that we are using 2 nos. which are of smaller magnitude instead of a single no. with a large magnitude.

The remaining part of this paper is structured as follows: In section II, discus related work on steganography. Our proposed model with technical approach is being described section III. The section IV focuses on implementation details on weighted directed graph cover media. Finally conclusion and feature work are considered as well in section V.

II. PREVIOUS WORK

This section describes some of the related works on information hiding along with their advantages and disadvantages.

A. Text Steganography in Markup Languages.

In this method, one of the features of markup languages is used to hide information [10]. For instance feature of HTML document is their tags case insensitivity. For example, the tag <BR> can be also used as <Br> and <br>. As a result one can do text steganography in HTML documents by changing the small or information in first method of text steganography by

978-1-4799-2397-7/14/$31.00 ©2014 IEEE

comparing these words with words in normal case and in second case by information in first method of text steganography by comparing these words with words in normal case and in second case by comparing the tags positions. Then by using appropriate function in both hidden information is extracted. However these methods are not for all markup languages like in the WML., all tags are case sensitive and as a result first text steganography method cannot be employed on it but second text steganography method can be employed.

B. Text Steganography in Specific Characters in Words.

In this method, some specific characters from certain words are selected [11]. For example the first words of each paragraph are selected in a manner that by placing the first characters of the selected words side by side, it's forms secret or hidden information is extracted. This method requires strong mental power along with it takes a lot of time and it also requires special text because not all type of texts can be used in this method.

C. Line Shifting Method.

In this method, the lines of the text are vertically shifted to some degrees [12, 13]. For example, some lines are shifted 1/300 inch up or down in the text and information are hidden by creating a hidden unique shape of the text. This method is feasible for printed texts. However, in this method, the distances can be observed by using special instruments of distance assessment and necessary changes can be introduced to destroy the hidden information. Also if the text is retyped or if character recognition programs (OCR) are used, the hidden information would get destroyed.

D. Word Shifting.

In this method, by shifting words horizontally and by changing distance between words, information are hidden in the text [12, 14]. This method is acceptable for texts where the distance between words is varying. This method can be identified less, because change of distance between words to fill a line is quite common. But if somebody was aware of the algorithm of distances, he can compare the present text with the algorithm and extract the hidden information by using the difference. The text image can be also closely studied to identify the changed distances. Although this method is very time consuming, there is a high probability of finding information hidden in the text. The same as in the method described under 2-3, retyping of the text or using OCR programs destroys the hidden information.

E. Syntactic Methods.

By placing some punctuation signs such as full stop (.) and comma (,) in proper places, one can hide information in a text file [11]. This method requires identifying proper places for putting punctuation signs. The amount of information to hide in this method is trivial.

F. *Semantic Methods.*

In this method, we use the synonym of words for certain words thereby hiding information in the text [10, 15]. A major advantage of this method is the protection of information in case of retyping or using OCR programs (contrary to methods listed under 2-3 and 2-4). However, this method may alter the meaning of the text.

G. Feature Coding.

In this method, some of the features of the text are altered[16]. For example, the end part of some characters such as h, d, b or so on, are elongated or hortened a little thereby hiding information in the text. In this method, a large volume of information can be hidden in the text without making the reader aware of the existence of such information in the text. By placing characters in a fixed shape, the information is lost. Retyping the text or using OCR program destroys the hidden information.

H. Abbreviation.

Another method for hiding information is the use of abbreviations. In this method, very little information can be hidden in the text [8]. For example, only a few bits can be hidden in a file of several kilobytes.

I. Open Spaces.

In this method, hiding information is done through adding extra white-spaces in the text [8, 17]. These whitespaces can be placed at the end of each line, at the end of each paragraph or between the words. This method can be implemented on any arbitrary text and does not raise attention of the reader. However, the volume of information hidden under this method is very little. Also, some text editor programs automatically delete extra white-spaces and thus destroy the hidden information.

J. Vertical Displacement Of The Points In Persian Letters.

In this method, text steganography is applied on Persian text [18, 19]. One of the characteristics of Persian language is abundance of points in its letter. In Persian 18 letters out of 32 letters have points. From these 18, 3 letters have 2 points each, 5 letters have 3 points each and the other 10 letters have 1 point each. These 1 point letters are used to hide the information by shifting position of point a little bit vertically high with respect to the standard point position in the text.

III. PROPOSED TECHNIQUE

There are various approach for text stagonography. In this paper representation a new approach, that is secret Message will pass through directed weighted graph cover media apply new model of number system. For our proposed scheme, on Indian Languages. This paper proposes a new scheme on how to efficiently apply new model of number system for embedding secret messages. Our proposed model deals with a new kind of number system. This does not mean that a new set of symbols are used as the digits of the system. It uses the same 10 digits i.e. 0-9 as is used in decimal system but this system is somewhat different in its representation. In this representation any no is represented by an ordered pair of 2 numbers. These two numbers are then put in a particular formula to get the final value of the number the concept will be more clear by looking at the following example:- Let us consider a number say 748 represented in decimal number system. Now the value of the number 748 is computed as (7*100 + 4*10 + 8). Similarly if the number 748 was written in some other number system with base 'b' then the values is given as (7*b*b + 4*b + 8). So each number system need to operate on the digits to get the final value. In a similar manner our system also operates on the digits to get the final value but the computation is somewhat different. In our system the number 748 of decimal is written as (38, 7) and the computation is done by the formula ((x*(x+1)/2) + y) for an ordered pair (x, y). So the value of (38, 7) is given by: 38*39/2 + 7 = 19*39 + 7 = 741 + 7 = 748. Here each of the components i.e. x and y are treated as decimal. This concept is a different way of counting. That means ancient times counted a commodity by replacing an equal no of different commodity in place of the commodity to be counted. Thus they could not represent the quantity in terms of some standard. Then the decimal system evolved and we got a standard to represent any quantity. In the decimal system the counting is done by dividing the quantity in a group of ten. Thus for any number of 2 digits can be visualized as divided into a no of groups containing 10 units of the commodity each. (1,2,3,4,5,6,7,8,9,10),(11,12,13,14,15,16,17,18,19,20),(21,22,23,24,25………………

This number of group is equal to the tens digit of the number and the extra quantity left (less than 10) is the units digit of the number. But in my system the division is done differently:- Here we take any quantity of a commodity and start our counting as

(1),(1,2),(1,2,3),(1,2,3,4),(1,2,3,4,5),(1,2,3,4,5,6),(1,2,3,4,***5***,6,7),(1,2,3,4,5,6,7,8)………

i.e. we divide the group in such a manner that the 1st group can contain 1 no. 2nd group can contain 2 numbers. and the nth group can contain 'n' numbers. thus if we want to represent the number given by the 5 in the 7th group then it is represented as (6,5) so the formula gives us the value as 6*7/2 + 5 = 21 + 5 = 26. This can be verified by counting the position of the italic ***5*** given above. It is the 26th number form left. So the ordered (x, y) pair is defined as:

X= No. of groups that have been fully completed including the number being considered.

Y= number of extra elements left.

Thus if we consider the number 7 of the 7th group then we see that including that '7' we can have 7 completed groups so x =7, but we do not have any extra element left so y = 0. So the number is given as (7, 0).

We consider directed weight graph as cover media and in this graph G(V,E) contain number of vertexes and edges , each vertex will be represented by the coordinator in X,Y plane and each edge assigned by different cost. Among several vertex one is the start vertex it will be indicated by (0,0) coordinator. Graph traverse always start from start vertex that mean (0,0) coordinator to it's all neighbour vertex and so on according to the direction of the graph. During traverse only one path has taken which is minimum cost in between two or set of vertexes and the end vertex coordinator of this path is a valid coordinator. These coordinator values represent the X and Y value in new number system approach. From our number system using X and Y values extract a decimal value which is the ascii value of a character in any secrete message. Like this way valid coordinator has been taken and using it extract the ascii value of a character in the secret message.

IV. THEORETICAL AND TECHNICAL BACKGROUND

In this paper our cover media is weighted directed graph. In this graph all vertices have a co-ordinate in x-y plane. And among all vertices one is start vertex represented by (0,0) co-ordinate among the several vertex the different cost has been assigned. And our secret message is "DYNAMITE". Given below the fig.

A.. Encoding The Message

- First character of the above secrete message is "D" and it's ascii value is 68, and 68 make a co-ordinates that is (11, 2) where x = 11, and y = 2.

- Second character of the above secrete message is "Y" and it's ascii value is 89, and 89 make a co-ordinates that is (12, 11) where x = 12, and y = 11.

- Third character of the above secrete message is "N" and it's ascii value is 78, and 78 make a co-ordinates that is (11, 12) where x = 11, and y = 12.

- Fourth character of the above secrete message is "A" and it's ascii value is 65, and 65 make a co-ordinates or that is (10, 10) where x = 10, and y = 10

- Fifth character of the above secrete message is "M" and it's ascii value is 77, and 77 make a co-ordinates that is (11, 11) where x = 11, and y = 11.

- Sixth character of the above secrete message is “I” and it’s ascii value is 73, and 73 make a co-ordinates that is (11, 7) where x = 11, and y = 7.

- Seventh character of the above secrete message is “T” and it’s ascii value is 84, and 84 make a co-ordinate that is (12, 6) where x = 12, and y = 6.

- Eight character of the above secrete message is “E” and it’s ascii value is 69, and 69 make a co-ordinates that is (11, 3) where x = 11, and y = 3.

Above the all co-ordinates are valid co-ordinates and contain the secret message. That is shown in the bellow fig.

B. Decoding the message:-

- Step 1 : vertex scan start from (0,0) co-ordinates to all it’s adjacent coordinator that is (1,2), (7,8), (6,2), (11,2) according to direction below graph.

- Step 2: comparison the different type of cost in the directed graph. In below graph comparisons have board among cost 5,3,4,2.

- Step 3: minimum cost will be consider for actual movement from one node to another node. Here 2 is the minimum weight, movement will happen from (0,0) to (11,2).

- Step 4: after movement current vertex coordinators chosen is valid co-ordinate that represent value of x and y. here current vertex is (11,2) after move from (0,0). So 11,2 is the valid co-ordinate, where x = 11, and y = 2.

- Step 5: using x, y value on a new number system extract the decimal value. Formula is x*(x+1)/ 2 + y.

So, in the above example x = 11, x+1 = 12, and y = 2. (11 * 12) / 2 + 2 = 68.

- Step 6: the decimal value convert the character. In our case decimal value 68 is represent the character D (Ascii value of D is 68).

- Step 7: step 1 to 6 will be repeated decrypt the all character in the secret message.

V. CONCLUSION AND FUTURE RESEARCH SCOPE

In this paper we have introduced a new approach of using a new kind of number system for Steganography through directed weighted graph cover media. This system is taking the advantage of the existence of many different types of Steganography approach like feature code able characters in Indian Languages, line shifting, space adding etc . It is a totally new kind of approach where hide the secrete message is very easy on the basis of a new number system approach but without using any type of specific algorithm. Decrypt the message from a cover media based on a new number system approach and the general formula of the new number system for this no need any specific algorithm.

Future research can be made to develop new kind of cover media based on proposed number system or modify the above cover media to sparse the secret message.

REFERENCES

[1] C Cachin, "An Information-Theoretic Modelfor Steganography", in proceeding 2nd Information Hiding Workshop, vol. 1525, pp. 306-318, 1998

[2] R Chandramouli, N. Memon, "Analysis of LSB Based Image Steganography Techniques", IEEE pp. 1019-1022,2001.

[3] NT Johnson, S. Jajodia, "Staganalysis: The Investigation of Hiding Information", IEEE, pp. 113-116, 1998.

[4] D. Artz, "Digital Steganography: Hiding Data within Data", IEEE Internet Compufing, pp. 75-80, May-Jun 2001.

[5] Herodotus. The Histories. Penguin Books, London, 1996. Translated by Aubrey de Selincourt.

[6] G. Simmons, "The prisoners problem and the subliminal channel," CRYPTO, pp.51-67, 1983.

[7] Chen, T. S. Chen, M. W. Cheng, "A New Data Hiding Scheme in Binary Image," in Proc. Fifth Int. Symp. on Multimedia Software Engineering. Proceedings, pp. 88-93 (2003).

[8] J.C. Judge, "Steganography: Past, Present, Future", SANS white paper, November 30, 2001, http://www.sans.org/rr/papers/index.php?id=552,

[9] R. Chandramouli, and N. Memon, "Analysis of LSB based image steganography techniques", Proceedings of the International Conference on Image Processing, vol. 3, 7-10 Oct. 2001, pp. 1019 - 1022.

[10] M. Shirali Shahreza, “An Improved Method for Steganography on Mobile Phone”, WSEAS Transactions on Systems, vol. 4, Issue 7, July 2005, pp.955-957.

[11] G. Doërr and J.L. Dugelay, "A Guide Tour of Video Watermarking", Signal Processing: Image Communication, vol. 18, Issue 4, 2003, pp. 263-282.

[12] G. Doërr and J.L. Dugelay, "Security Pitfalls of Frame by Frame Approaches to Video Watermarking", IEEE Transactions on Signal Processing, Supplement on Secure Media, vol. 52, Issue 10, 2004, pp. 2955-2964.

[13] K. Gopalan, "Audio steganography using bit modification", Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,(ICASSP '03), vol. 2, 6-10 April 2003, pp.421-424.

[14] J.T. Brassil, S. Low, N.F. Maxemchuk, and L.O’Gorman, "Electronic Marking and Identification Techniques to Discourage Document Copying", IEEE Journal on Selected Areas in Communications, vol. 13, Issue. 8, October 1995, pp. 1495-1504.

[15] W. Bender, D. Gruhl, N. Morimoto, and A. Lu,"Techniques for data hiding", IBM Systems Journal, vol. 35,Issues 3&4, 1996, pp. 313-336.

[16] N. Provos and P. Honeyman”Hide and Seek ” An introduction to steganography, IEEE Security and Privacy, p.p 32-44, May/June2003.

[17] K. Bennett, "Linguistic Steganography: Survey, Analysis, and Robustness Concerns for Hiding Information in Text", Purdue University, CERIAS Tech. Report 2004-13

[18] T. Moerland, "Steganography and Steganalysis", May15, 2003,www.liacs.nl/home/tmoerlan/privtech.pdf,

[119] S.H. Low, N.F. Maxemchuk, J.T. Brassil, and L.O'Gorman, "Document marking and identification using both line and word shifting", Proceedings of

the Fourteenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'95), 2-6 April 1995,vol.2, pp. 853 - 860.

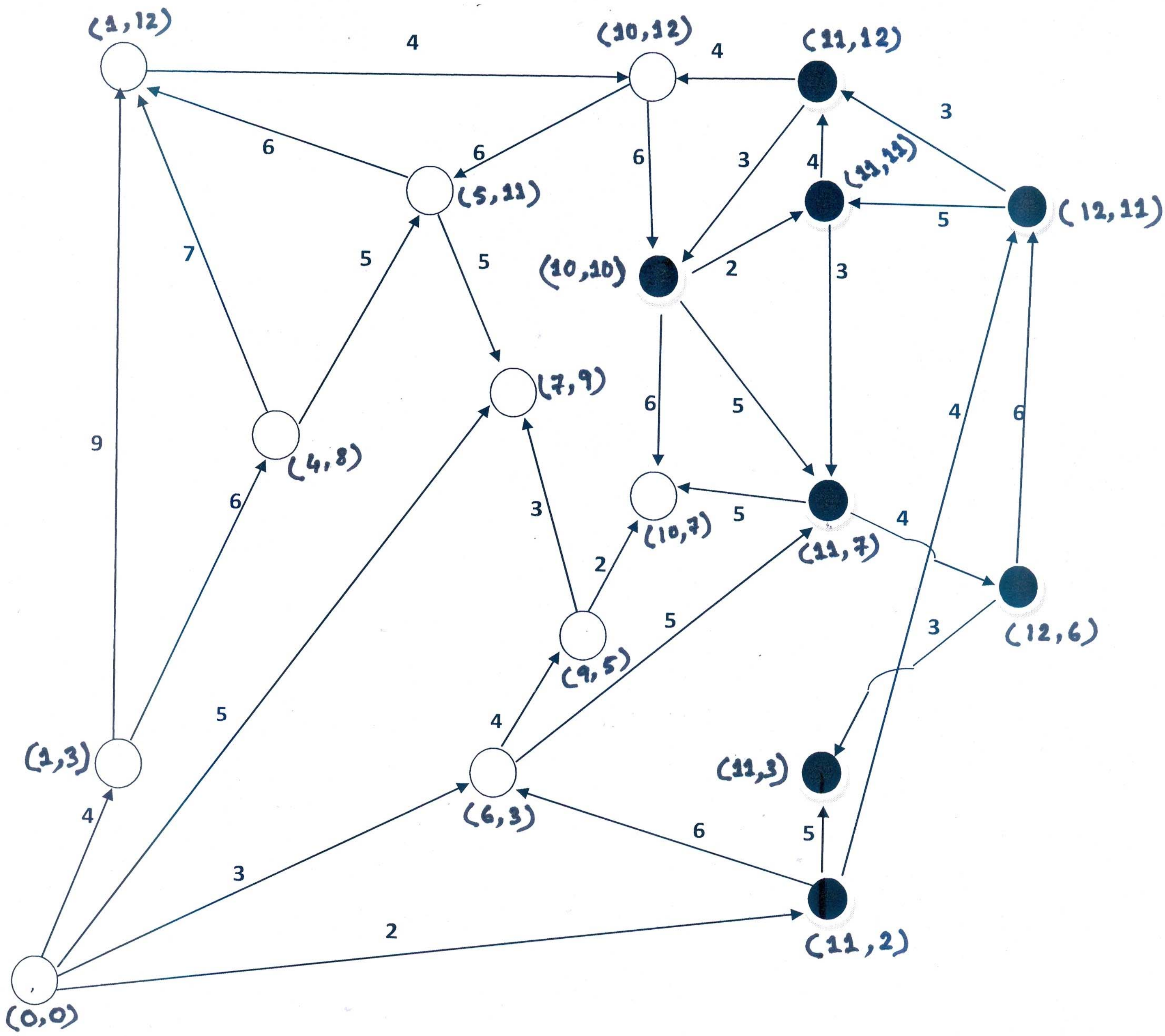
[20] A.M. Alattar and O.M. Alattar, "Watermarking electronic text documents containing justified paragraphs and irregular line spacing ", Proceedings of SPIE -- Volume5306, Security, Steganography, and Watermarking of Multimedia Contents VI, June 2004, pp. 685-695.

[21] Y. Kim, K. Moon, and I. Oh, "A Text Watermarking Algorithm based on Word Classification and Inter word Space Statistics", Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03), 2003, pp. 775–779

[22] M. Niimi, S. Minewaki, H. Noda, and E.Kawaguchi, "A Framework of Text-based Steganography Using SD Form Semantics Model", Pacific Rim Workshop on Digital Steganography 2003, Kyushu Institute of Technology, Kitakyushu, Japan, July 3-4, 2003.

[23] K. Rabah, "Steganography-The Art of Hiding Data", Information Technology Journal, vol. 3, Issue 3,pp.245-269,

Fig.


(1,12)
4
(10,12)
4
(11,12)
3
6
6
6
3
4
(11,11)
(5,11)
5
(12,11)
7
5
5
(10,10)
2
3
(7,9)
6
5
4
6
9
(4,8)
6
3
(10,7)
5
4
(11,7)
2
(12,6)
5
3
(9,5)
5
4
(1,3)
(11,3)
4
(6,3)
6
5
3
2
(11,2)
(0,0)