Principle of Graph Theoretic Approach to Digital Steganography

Vinay Kumar*, Scientist ‘D’, NIC, DIT, GOI, New Delhi, India, e-mail: vinay5861@gmail.com and S. K. Muttoo, Department of Computer Science, University of Delhi, India, e-mail: skmuttoo@cs.du.ac.in

* Corresponding Author

ABSTRACT

Steganography literally means secret writing. The technique has been used in various forms for 2500 years or long. It has found its application in various fields including military, diplomatic, personal and intellectual property applications. Briefly stated, steganography is the term applied to any number of processes that will hide a message within an object, where the hidden message will not be apparent to an observer. The paper describes the concept of finding natural relationship between a digital cover and a message. The relationship can be used to hide the information in cover without actually replacing or distorting any useful bits of the cover. It introduces a concept called sustainable embedding of message in a cover using natural relationship and representing it using graph theoretic approach.

KEYWORDS

Extra bytes, graph theoretic approach, pure steganography, secret key steganography, sustainable, natural embedding, spatial resource, bipartite graph, maximum bipartite matching, steganalysis.

1. INTRODUCTION

Steganography is the art of invisible communication. Its purpose is to hide the very presence of communication by embedding messages into innocuous-looking cover objects. Each steganographic communication system consists of an embedding algorithm and an extraction algorithm. To accommodate a secret message in a digital cover, the original cover is slightly modified by the embedding algorithm. The result is modified cover object that contains the secret message and it is called stego object. The important requirement for a steganographic system is its detectability by an attacker with probability not better than random guessing, given the full knowledge of the embedding algorithm, including the statistical properties of the source of cover object. Of course the stego key is not revealed. The most commonly used steganographic method is the Least Significant Bit (LSB) replacement. It works by embedding message bits as the LSBs of randomly selected pixels. A secret stego key shared by the communicating parties usually determines pixel selection.

The popularity of LSB embedding is due to its simplicity as well as the false belief that modifications of LSBs in randomly selected pixels are undetectable because of the noise commonly present in digital images of natural scenes. However, flipping the bits of the LSB plane does not occur naturally. The even pixel values are either unmodified or increased by one, while odd values are either decreased by one or left unchanged. This imbalance in the embedding distortion is utilized to mount successful attacks. The image in figure 1(a) is modified to embed “Graph Theoretic Approach” to get stego in figure 1(b). Through the naked eye, it is nearly impossible to detect any difference but any steganalysis tool can very easily find the presence of some message in the cover, given the cover 1(a).

![Figure 1](image)

The image is of size 100x100 pixels and 24 bit color, so there are total of

\[100 \times 100 \times 3 + 54 \text{ bytes} = 30,054 \text{ bytes}\]

are available to hide a message of 3756 characters. There are a few extra bytes also available in any image. Can it be utilized for the purpose? Can a message be hidden in a cover without replacing any useful bits? These are few questions that we try to answer through this paper by introducing a graph theoretic approach to steganography. Before explaining the concept let us have a brief description of types of steganography in use today.

2. TYPES OF STEGANOGRAPHIC METHODS

The steganographic technique used today can be broadly classified into three categories: Pure Steganography, Secret Key Steganography and Public Key Steganography.

PURE STEGANOGRAPHY

It is defined as a steganographic system that does not require the exchange of a cipher such as a stego-key. This method of secret communication is treated as least secure means of communication. The sender and receiver rely upon the presumption that no other parties are aware of this secret
message. But the amateur hackers who are working 24x7, can scan the cover (stego) using various tools available today to guess the presence of a possible secret message and can either alter or destroy it on the way.

SECRET KEY STEGANOGRAPHY
It is a method that involves exchange of stego-key pre or post communication of secret message steganographically. Secret Key Steganography takes a cover message and embeds the secret message inside of it by randomized technique. The random key that forms the secret key is also called stego-key. Only the communicating partners know the stego key and can reverse the process and read the secret message. Secret key steganography is more secure than pure steganography because only those parties who know the secret key can extract the secret message.

PUBLIC KEY STEGANOGRAPHY
It uses the concepts of public key cryptography [13]. Public key steganography is defined as a steganographic method that uses a public key and a private key to secure the communication between the parties wanting to communicate secretly. The sender will use the public key during the encoding process and the private key, which has a direct mathematical relationship with the public key, can decipher the secret message. Public key steganography provides a more robust way of implementing a steganographic system because it can utilize a much more robust and researched technology in Public Key Cryptography. Before any unwanted parties could intercept the secret message they have to first suspect the use of steganography and then they have to find a way to crack the algorithm used by the public key system. In this way it provides multiple layers of security.

The schematic diagram shown in the figure 2 explains the secret key steganography. Despite the security provided to steganographic method by stego key, bits replacement incorporated in cover file provides enough statistics to the hacker to guess the presence of message and classifying the media as stego.

![Figure 2](image.png)

To overcome this limitation of steganography, a graph theoretic approach is suggested where message is treated as naturally embedded without either replacing or exchanging the useful color bits of a cover file.

3. GRAPH THEORETIC APPROACH
A graph theoretic approach can be described combining meanings of three different words: graph, theoretic and approach. ‘Graph’ is a branch of mathematics dealing with the properties of diagrams to study the arrangements of objects and relationships between objects. ‘Theoretic’ is concerned primarily with theories or hypothesis rather than practical considerations; called “Theoretical science”. Setting aside this classical definition of theoretic, this term may be taken in this context as “concept of data structures used to store a graph in Computer science”. An ‘approach’ is described as a method used or steps taken in setting about a task, problem, etc. When the three words are combined, we may interpret “graph theoretic approach” as “a method that uses graph data structure in principle to solve a problem.” When it is used to solve a steganographic problem, it is called “Graph theoretic approach to Steganography”. There are two ways of using graph theoretic concept in digital steganography.

(i) Find relationship (if required) between smallest data unit of message and a group of such smallest unit of cover object and represent the relationship using a graph. If required, hide the relationships in the zero bytes of cover.

(ii) Use a graph as cover object and find redundancy in its feature like node or segment or its attributes and embed secret message in it.[9]

In the case of (ii) a data structure for storing graph has to be used so that features like node, segment and points constituting a segment should be stored [9]. In case of (i) it is explored to find association between bits string in cover and bit string in secret message. In case such association is found then following issues need to be addressed to implement the graph theoretic approach to digital steganography.

(i) What data structure should be used to represent such association?
(ii) And how this data can be embedded or assumed to be embedded in the cover?
(iii) What stego key can be used between the communicating partners?

In the following sections of the paper, possible approach is described. Authors of the paper are working in this area.

4. FINDING RELATIONSHIP
In digital world, every cover media like image, video, audio, speech etc can be treated as collection of data units. Each data unit is nothing but string of bits. Say the size of data units be k bits. Thus a cover can be treated as an array of such data units. Similarly a secret message to be embedded in the cover, may be treated as an array of data units each of size k bits. Let size
of message array is n. The size k can be adjusted in a way to avoid any padding. An embedding factor can be defined as ratio

\[ m = \frac{\text{size of cover array}}{\text{size of message array}} \]

It means that potentially m data units are available in cover for one data unit of secret message. Thus cover data units may be arranged in two-dimensional array of size n \times m. A maximum bipartite matching can be found using the natural presence of secret message in cover. Detail is given in the next section while finding a way to embed the relationship in cover.

For example, if m = 4 and k = 2, then one of the data unit “01” of message is said to be naturally present in the corresponding m data units of cover “10 00 01 10” at 3\text{rd} place. Also, if we take additional modulo 4 of “10 00 01 10” then it is also equal to 01 and hence it can be concluded that the part of secret message is naturally present in the corresponding part of the cover. There can be many other ways to find such relationship. If such correspondence is achieved for every data unit of the secret message then the message is embedded without any replacement of bits in the cover. But such ideal situation is rarely available.

In case some mismatch is found then it may be explored that i\text{th} data unit may corresponds to some combinations of j\text{th} block of m data units of cover file. If it is found, then the association can be represented as an arc (i, j). There may be one to many associations because one data unit may correspond to multiple m data units of cover file. At the end there may be some data units in message that are not associated to any of the block in cover. Thus, there are three possibilities:

(i) Direct relationship is found between i\text{th} data unit of message and j\text{th} block of cover.

(ii) Cross relationship is found between i\text{th} data unit of message and j\text{th} block of cover.

(iii) Some mismatched data unit of message say (some k\text{th}) data units.

After formulating the concept of finding relationships between bit patterns in message and that in cover, it is time to formulate a way to use the relationship to hide message in cover in such a way that no or very little modification of color bits of cover is required.

5. EMBEDDING THE RELATIONSHIP

The method is computation intensive. We first define a bipartite graph G = (V, E) taking every data unit of message as a node in the set L (left). Similarly every block of m data units from cover is taken as node in the set R (right). A \textit{bipartite} graph is a graph G whose vertex set \(V\) can be partitioned into two non empty disjoint sets \(L\) and \(R\) in such a way that every edge of \(G\) joins a vertex in \(L\) to a vertex in \(R\). Using the direct relationship and cross relationship, as described in the previous section, we find a maximum bipartite matching.

A matching of G is a subset M of the set E of edges of G, with the property that no two edges in M have a common vertex. Thus no left vertex is incident to more than one edge of the matching, and no right vertex is incident to more than one edge of the matching. The matching associates, or matches, some of the left vertices to some of the right vertices in a one-to-one way. Matchings are an important area of study in graph theory because many practical problems can be seen as requiring the discovery of an optimal matching in a bipartite graph.

![Figure 3](image.png)

A matching is a collection of edges. Each edge in the matching is a confirmation that data unit of secret message is naturally embedded in the block of data units of cover image and there is no need to flip any bits of the cover to hide that part of message. Clearly the best matching we can obtain would have three edges. \(M = \{(1, 7), (2, 8), (4, 5)\}\) is one such matching. There are a few others in the graph. A matching having a maximal number of edges is called a max-matching. How to use this concept to embed secret message in cover?

Let us take an example to illustrate it. Suppose four data units, denoted as node 1, 2, 3, and 4 in the bipartite graph of figure 3, of secret message are to be embedded in the four blocks of data units of cover image, denoted as node 5, 6, 7, and 8 in the graph. Node 1 cross matches with only one node 7. Node 2 directly match to node 6 and cross matches to node 8. Similarly node 4 directly matches to node 8 and cross matches to node 5. After scanning it found that data units 2 and 4 have direct relationship with corresponding blocks in cover and that part of secret message is naturally embedded. Nothing needs to be done to hide these parts of secret message. Node 1 cross matches with 7 so hiding (1,7) in the cover in the extra byte shall achieve the following:

(i) No color bits of cover is flipped to accommodate the message, and

(ii) In stead of writing message part, its indexes and cover block index is stored only.

Next the unmatched part i.e. node 3 is embedded as \((M, 3, <\text{message part}>)\) in extra byte. Extra bytes are those parts in a image that are padded and kept to maintain a fixed predefined file format of the image. It needs to be explored in a cover.
before applying this approach. We have explored a few file formats including BMP to get presence of extra bytes. At the end, a secret key is generated that contains <Size of the message, Data unit size>. The size of the message and data unit size together determines array size of message. The size of the image received at the recipient side and this information is used to find embedding factor. Once the information from the extra bytes, if any, is retrieved, it conveys how many are naturally embedded in the image and then the secret message can be reconstructed from image data itself.

6. STEGANALYSIS
Discussion of steganography necessarily leads to discussion of steganalysis. Steganalysis is the art of discovering the presence or transmission of stego content in the communication channel. From information security point of view it is crucial to thwart any attempt of stego communication. It is a probabilistic science. A steganalysis algorithm generally output a number within a given range to indicate the likelihood that stego was actually used on the input file. Steganalysis is a rapidly advancing science, and will continue to develop as long as steganographic algorithms are being created and used.

There are two approaches to the problem of steganalysis, one is specific to a particular steganographic algorithm. The other is developing techniques completely independent of the steganographic algorithm to be analyzed. Each of the two approaches has it’s own advantages and disadvantages. A steganalysis technique specific to an embedding method would give very good results when tested only on that embedding method and in all possibility it fails on other steganographic algorithm. However an independent steganalysis method may perform less accurately but it is capable to provide acceptable result on any new embedding algorithm. When the embedding algorithm is based on a graph theoretic approach that either does not change/replace any color bits of cover or changes very few colors bit, a general steganalysis algorithm can not work. In fact associations in the maximum bipartite matching takes care of the embedding process.

In general, the number of unmatched vertices is an upper bound on the number of changes to first-order statistics. An experimental result has shown that sufficiently good matching (< 3% unmatched) can be reached for natural cover data. This makes a graph theoretic approach practically undetectable by tests that look only at first-order statistics. It would be interesting to run a blind steganalysis scheme against this implementation to compare its detectability to the other tested algorithms. Adding a restriction to the set of edges could easily extend this graph theoretic approach and make it more secure against any general steganographic algorithm. However in the approach presented the steganography totally depends on the key pairs consisting of size of message and data unit size. To the third party none of the information is known and hence given the arbitrariness of message size and data unit size for any message, it is simply Herculean for any body to guess and extract the message.

7. CONCLUSION
A graph theoretic approach to steganography in an image as cover object helps in retaining all bits that participate in the color palette of image. The method is based on exploring maximum natural embedding and then finding relationship that conveys the presence of message in cover without either replacing or exchanging any bits of cover. This way the technique achieves sustainability. Sustainable steganography can be described as a method of hiding in such a way that no color bit is altered. Today various digital data formats are used in steganography. Most popular among them are bmp, doc, gif, jpeg, mp3, txt and wav because of the relative ease by which redundant or noisy data can be removed from them and replaced with a hidden message. BMP image is found to be more suitable because of presence of some redundant bytes at the quad word boundary. Since every digital cover file is simply stream of bits, an algorithm that treats cover as stream of bits, can be applied to any image/audio format with a little modification in finding zero bytes and header information. The concept presented in this paper for natural embedding can be further improved by using variable embedding factor k for adjusting its value whenever maximum natural embedding is achieved.

8. FUTURE SCOPE
Steganographic research is primarily driven by the lack of strength in the cryptographic systems on its own and the desire to have complete secrecy in an open-systems environment. Redundancy is not always useless. A lot of research is required to evolve techniques to naturally embed message in digital cover media. It can be used for the benefit of the society as well as for better administrative management by keeping any secret information secret and beyond the reach of spoiler by maintaining its utmost privacy. The rich resources of spatial data available under national spatial database project by many governments around the world may also be used for the purpose of steganography using graph-theoretic approach to steganography. “A successful steganography is one that neither disturbs nor replaces any useful bits of cover. A successful steganalysis is one that retrieves message from stego without any clue about it.”
9. ACKNOWLEDGEMENT
Inspiration comes from many sources. In fact it is always there. One has to look around to know the presence of something worth noticing. The paper is result of the present potential threat that the wide spread information exchange through network is facing from amateur hackers. The open world today wishes to exchange information using a public network infrastructure but in secured manner. Thankfully, the situation provides an opportunity to think about. We are also very grateful to all those who have been constantly encouraging us to go for such scientific research work besides the regular work which we are doing in our respective departments.

REFERENCES
[9] Kumar, V. and Muttoo, S. K. “A data structure for graph to facilitate hiding information in a graph’s segments -A Graph Theoretic Approach to Steganography”, Int. J. Communication Networks and Distributed Systems

BIOGRAPHICAL NOTES
1. Vinay Kumar* is a postgraduate in Mathematics and did his MCA from School of Computer and System Sciences, Jawaharlal Nehru University, New Delhi. He is working as a Scientist in National Informatics Center, DIT, Government of India. He is working in the area of Graph Theoretic Approach to Steganography. Numbers of papers are published in his name in Journals and conferences.
2. Dr. Sunil K. Muttoo is Reader in Department of Computer Science, University of Delhi, India. He completed his M. Tech from IIT Kharagpur and Ph.D. from University of Delhi, India. His specialized area is coding theory, information hiding, encryption, data security and privacy. He has been Research Scholar of ISRO sponsored Project “Gross Migration Scheme”.