Eng. & Technology, Vol.24, No.6, 2005

# Analysis of Information Hiding Techniques in the Texts

Dr. Abdul Monem S. Rahma* Abdalla Mohamed T. Al-Dhao**

Received on: 5/7/2004
Accepted on: 14/3/2005

**Abstract**

Nowadays; information hiding represents one of important sciences, which was found more importance by the researchers. The goal of steganography is to transmit messages to destination in full secrecy. From that much media types used to hide information like (images, sound, disks, and networks).

The philosophy of steganography depends on three elements in general: The cover message, hidden message and stego file, in addition to the key which is sometimes used in hiding and extracting processes.

Some information hiding techniques in the texts have discussed the advantages and disadvantages of these methods and their compressions, in addition to the discussion of sub-techniques that are related with the two basic operations: hiding and extracting processes. The paper outlines that information hiding in texts demands hard potential in both hiding process and steganalysis of the hidden message. In addition, data-hiding techniques in texts bring some vulnerabilities like other information hiding techniques in other media types.

## تحليل طرق إخفاء المعلومات في النصوص

الخلاصة

إخفاء المعلومات يمثل أحد العلوم المهمة في هذا العصر ، والذي لاقى اهتماما متزايداً من قبل الباحثين. إن هدف إخفاء المعلومات هو ايصال الرسائل إلى الطرف الآخر بسرية تامة. لذلك أوساط عديدة استخدمت لاخفاء المعلومات كالصورة ، الصوت ، الشبكات والأقراص.

إن فلسفة الاخفاء بصورة عامة ترتكز على ثلاثة عناصر أساسية هي : الغطاء (Cover) والرسالة المخفية (Hidden Message) والرسالة بعد الاخفاء (Stegofile) ، بالاضافة الى المفتاح (Key) الذي يستخدم أحيانا في عمليتي الاخفاء (Hiding Process) والاسترجاع (Extracting Process) ، وأن هاتين العمليتين ترتبطان بعدد من العمليات الفرعية الأخرى مثل : التشفير ، وفك الشفرة ، وضغط البيانات.

ناقش البحث بالتحليل عدد من طرق إخفاء المعلومات في النصوص ، وقد ركز على إبراز سلبيات وإيجابيات هذه الطرق مع مقارنتها. فضلاً عن مناقشة العمليات المساعدة الأخرى المرتبطة بالعمليتين الأساسيتين (الاخفاء والاسترجاع). تم التوصل إلى أن بعض أساليب الاخفاء في النصوص يتطلب جهداً عالياً لاخفاء الرسالة ، مثلما تتطلب جهداً عالياً أيضاً لمعرفة الرسالة المخفية. كذلك تم التأكيد على أن طرق الاخفاء في النصوص تحمل بعض نقاط الضعف مثل نظيراتها في طرق الاخفاء في الأوساط الأخرى.



* Dept.of Computer Science, University of Technology
** iraq commission for computers and information

https://doi.org/10.30684/etj.24.6.2
University of Technology-Iraq, Baghdad, Iraq/2412-0758

## 1. Introduction

Steganography literally means: "covered writing" and encompasses methods of transmitting secret messages through innocuous cover carriers in such a manner that the existence of the embedded messages is undetectable [10]. Perhaps the biggest advantage that steganography has over cryptography is the fact that transmission of secret information is non-observable. However, unlike cryptography, steganography requires a magnitude of overhead to hide small amount of information [13].

There are many different media types that information can be hidden in, like (Images, sound, disk space, and network packets), with different techniques. Most of techniques focus on hiding data in images or sound, and there are a few researches on data hiding in texts.

This paper focuses on analyzing of existent methods of information hiding in the text.

The analysis will discuss the text media and the advantages and disadvantages of text hiding techniques. In addition, the paper makes comparison of text techniques and the vulnerability of these techniques. No doubt text hiding techniques bring some weaknesses like other media of information hiding (image, sound, disk, networks), which also bring some similar vulnerability.

## 2. Text Media:

A text can be defined in a closed frames or line borders, and constructed from words written in a determinate language, for example Arabic or English language. Figure (1) expresses the idea of text media[14].

Text hiding means " Documents may be modified to hide information by manipulating positions of lines and words" [16]. In addition most of text hiding techniques aim to hide the data in text files by encoding the text with some other processes like compression and ciphering. Figure (2) shows the general model of information hiding in the texts [8].

The most popular techniques of hiding data in texts are [12]:

a. Old SPY used by German in WWII.
b. Line – Shift Coding.
c. Word – Shift Coding.
d. Feature Coding.
e. White Space Techniques.
f. Syntactic Techniques.
g. Symantic Techniques.
h. Typography Technique.
i. Context Free Grammar Techniques.
j. Dictionary Techniques.
k. Linked List Technique.
l. Hiding Text in HTML Files.
m. Hiding Text in XML Files.

## 3. Steganography Terminolo-gy:

The philosophy of steganography in the text and other media depends on three or four basic elements, related each with another: The **cover message**, which can be seen from any one, second: the **hidden message** which must be seen only from the receiver and the key which is used some times in hiding process, in addition to the stego file, which contains both: the cover message and the hidden message.

According to the terminology proposed by Pfitzmann to the first International Workshop on Information Hiding, the noisy message known as: **cover** and the bits carrying the noise is the **cover bits**. The bits embedded as pseudo-random noise are secret bits. The cover bits

substituted with the secret bits are called **hiding bits** [3].

The cover messages can be divided in two classes. The cover may be a stream cover or a random access cover. The former is a contiguous data stream, like a telephone connection and the latter may be a file, for example a wave file. In a stream cover it may not be possible to tell in advance where the cover message begins, where it ends or how long it will be [3]. In addition, the schemes for extracting data can be as follows:

a. **Cover escrow schemes**: where the original cover signal is needed to reveal the hidden information.
b. **Blind schemes:**These schemes allow direct extraction of the embedded data from the modified signal without knowledge of the original cover. [11]:

According to above definitions the stego message can be defined by the following formula:

I = f ( **I,m,K**) Such that:

**I** represents the Original Object.

**m** represents the hidden message.Let:

**K** represents the key that the two parties share.

In addition to the above formula the steganalysis algorithm of the hidden message (active steganalysis) works according to the following points [6]:

a. Estimate the embedded message length.
b. Estimate location(s) of the hidden message.
c. Estimate the secret key used in embedding.
d. Estimate some parameters of the stego-embedding algorithm.
e. Extract the hidden message (grand goal!).

## 4. Analysis of Techniques of Text Hiding:

As described before there are many ways to hide data in texts. These techniques differ from one to another. In the following sections the analysis of some of them is given.

### 4.1. Old German Spy Technique used in WWII:

During World War II, unencrypted transmissions were used to send embedded data. A German spy sent the following message [9]:

"*Apparently neutral's protest is thoroughly discounted and ignored. Isman hard hit. Blockade issue affects pretext for embargo on by-products, ejecting suets and vegetable oils.*"

After taken the seconed letter of each word is taken the following message can be extracted.

"Pershing sails from NY June 1"

**The advantages of this technique are:**

a. It is very simple in comparison with other stegano techniques and does not need experience to accomplish the hiding and extracting processes.
b. The cover is variable and can be changed with any sent message.
c. It does not need other processes related with information hiding like: encoding, ciphering and compression.
d. The key used (which is a second letter) is variable and can be replaced by another position.
e. There is no constant relation between positions to get specified rule for steganalysis of the message because of the

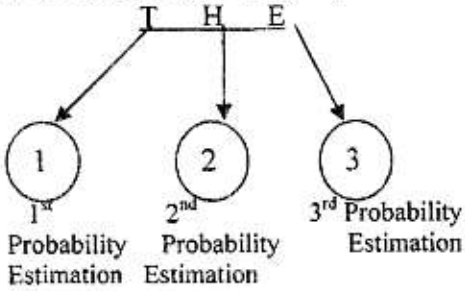
variabilities in positions and the words.

**The Vulnerable aspects of the above technique are:**

Using of the key may lead to successful steganalysis of the message. For example, if the string found in carrier includes the word the, then the key is limited to the values 1, 2, or 3. According to the previous formula the substitution of the key used which is (2), can be as in follows:

$I = f(I,m,2)$

If we analyze the word (the):




It can be seen that, the probability estimation (2) will show the weakness of this technique. In addition such a contrived message would never pass a big RAM analysis that analyzes the text for the probability that it is Standard English by comparing adjacent words.

**4.2. Line-shift Coding Method:**

In this method, text lines are vertically shifted to encode the document uniquely. Encoding and decoding can generally be applied either to the file format of the document, or the bitmap of a page image. However, if a document is marked with line-shift coding, it is particularly difficult to remove the encoding if the document is in paper format. Each page will need to be rescanned, altered, and reprinted. This can be complicated even further if the printed document is a photocopied, as it will then suffer from effects such as blurring, and salt-and-pepper noise. Figure (3) represents an example of this technique [11]:

**The advantages of this technique are:**

**a.** The far away of observation in the spaces between lines. Because of in the normal documents (specially in word documents) the spaces between lines appear in an ordinary form. In addition, this characteristic demands: some neighbor's lines to be unmodified.

**b.** Hard potential is required to detect the hidden message, because the hiding process is done in multi – complex steps, like: scanning and encoding of the shifted lines.

**c.** Rescanning in the recovered copy if its digital representation is not available.

**The Vulnerable aspects of the this technique are:**

**a.** The horizontal and the vertical profile of the copy can be compiled.

**b.** Recomposing the mark process can be available.

**4.3. White Space Techniques:**

There are three techniques of using white space to encode data. These methods are: inter-sentence spacing, end-of-line spaces, and inter-word spacing in justified text. In the method in inter sentence spacing the binary message can be encoded into the text by placing either one or two spaces after each terminating character, e.g., a period for English prose, a semicolon for C-code, etc. A single

space encodes a "0," while two spaces encode a "1." [4].

In the method of end of line spaces the encoding of data can be done by inserting spaces at the end of lines. The data are encoded allowing for a predetermined number of spaces at the end of each line. Two spaces encode one bit per line, four encode two, eight encode three, etc., dramatically increasing the amount of information we can encode over the previous method in the text which has been selectively justified, and has then had spaces added to the end of lines to encode more data. Rules have been added to reveal the white space at the end of lines [4].

The most effective sub-technique of white space is enterring word spaces, in which data are encoded by controlling where the extra spaces are placed. One space between words is interpreted as a "0." Two spaces are interpreted as a "1." This method results in several bits encoded on each line [4].

One of the provided steganographic software is software called SNOW, which uses the white space to hide secret messages. SNOW hides the message in ASCII text without affecting the visual representation of the text [15]. **The SNOW program runs in two main processes:**

**Hiding process:** which can be done according to the following steps:

Message-> optional compression -> optional encryption->concealment in text

**Extracting Process:** which can be done according to the following steps:

Extract data from text -> optional decryption -> optional uncompression -> message

[1] provides software uses of white space technique as combined method with Context Free Grammar (CFG) approach to hide message in text files. The provided approach is similar to SNOW software, which uses compression and encoding in both hiding and extracting stages. The difference between SNOW and this system is the construction of (CFG) rules and productions, which is the basic and first step in the (CFG) system.

Another new software provided, is a graphical editor [12], which is designed according to the Object Oriented Programming (OOP) and software engineering concepts as a similar software to word processor. The graphical editor succeeds in manipulating the problem of white space technique (**opening the file in the word processor**). In addition: the newest technique increases the amount of hidden message, that's by benefiting from the spaces found in the header file format of the editor. Similarly, as in two previous systems the graphical editor uses encoding and ciphering sub processes. After applying the graphical editor to some examples it is seen that the size of the hidden message is approximately equal to (27%) relative to the cover message.

**The advantages of these techniques:**

a. Changing the number of trailing spaces has little chance of changing the meaning of phrases or sentences.
b. They are unlike line shift coding or word shift coding which need scanning the document in the hiding stage.
c. Causal reader is unlikely to take notice of slight modifications to white space (far away from observations).
d. They can hide more bits when applying (inter word spaces) sub technique.

e. They can be used as a combined method with most techniques of hiding data in the texts. For example they can be combined with dictionary technique, (CFG) technique or Linked list technique.
f. they can be done with any text.

**The Vulnerable aspects of these techniques:**

The most popular vulnerable aspect of white space techniques is the shifting of spaces when opening the file with a common word processor.

**4.4. Syntactic Techniques:**

Syntactic method means: changing the diction and structure of text without significantly altering meaning or tone. There are many circumstances where punctuation is ambiguous or when mispunctuation has low impact on the meaning of the text. For example, the phrases "bread, butter, and milk" and "bread, butter and milk" are both considered correct usage of commas in a list [4].

There are some techniques which use this approach; we can discuss [1] technique, which uses (CFG) approach. The provided system works according to the following algorithm:

a. **Production Rules:** Construct production rules by using CFG.
b. **Encoding stage:** in which the sender derives one specific string out of the CFG, which will act as the stego-object.
c. **Compression stage:** That is by using Haffman codes.
d. **Hiding Stage:** Hiding the compression paths from the secret message by using a combination of white space technique (inter-word spaces) and syntactic technique that utilizes punctuation
e. **Decoding Stage:** This step includes some sub steps to extract the final message.

**The Advantages of this technique [1]:**

a. The hiding process of the binary paths gives good result, so it offeres more spaces between words to be used in hiding process.
b. Using of (CFG) increases the security of the hidden message, because of using determined rules to generate the word.
c. No need for the knowledge of original message in the decoding process (receiving process).

**The Vulnerable aspects of this technique are:**

a. The limitation of the system by using production rules. For example if the system uses Finite State Automata (FSA) then it will become more flexible and effective.
b. It is difficult to select meaningful type categories without considering the eventual grammatical requirements of a natural-language style-source.

**4.5. Dictionary Techniques:**

There are some techniques which use dictionaries in information hiding in texts. The famous two techniques dealing with this approach are:

**4.5.1. Raghad System [2]:**

This technique depends on constructing a dictionary, which contains words sorted in alphabetical order. These words are written in lower case (small letter) so that each letter in the word is converted to its code (ASCII). Each word in the dictionary can represent a single (letter or symbol) of the secret message according to their codes. This System is characterized by the following things [2]:

a. The system can encode the message directly in the text by

exploiting the natural form of the lower case of the letter without altering the form of the letter.

b. The system allows the user to deal with a dictionary and pick up the suitable words from it to build the stego message for any secret message.

c. The system is flexible with some options such that the user can build the stego message manually or automatically by two basic sub techniques: the stationary features form of a letter or by using the dictionary.

**The advantages of this method [2]:**

a. The system does not need the original message because it exploits the natural form of the lower case without altering the form of the letter. This will increase the security of the system. This method has more fitness than feature coding method because the latter method will alter the form of the letter to encode the message, then the receiver needs the original message to decode the secret message.

b. The system does not consume a large amount of space for encoding the message. To encode a letter in a word, this will need only a selected word from a dictionary to encode the entire letter.

c. The system will be protected from the word processor because it does not use the characteristics of open-space method.

**The Vulnerable aspects of this technique are:**

a. The limitation of the system because of using small letters.

b. The using of unique key in hiding process, which is one letter in the dictionary words. This can lead to successful steganalysis because encoding techniques are known.

c. Using of binary codes to encode the secret message can lead to successful steganalysis because: there are some steganalysis techniques which deal with analyzing of binary codes. For example, suppose the analyst uses the following formula:
Let: s(k) denote a cover message.
Let: w(k)} denote the hidden message.
Then the stego message be obtained as:
$z(k) = s(k) + w(k), k = 1, 2, \ldots, N$
According to the characteristics of the system the steganalysis, is blind scheme which does not need the cover; that means:

a. $s(k) = 0$ in the above formula.

b. K takes the values: 0,1,2,3,…..,N which correspond to: 000000, 0000001,0000010, 000011,….,N in the system. These binary codes can be estimated.

**4.5.2. NICETEXT System:**

It is another software that uses a dictionary technique combined with syntactic method (CFG) technique. This technique is provided by [7] in software called (NICETEXT) program. The basic idea of the Nicetext is to hide ciphertext or plain text according to the following algorithm:

a. Independent of the input ciphertext it chooses a contextual template, if current one is exhausted.

b. It reads the next word type from the template.

c. It reads enough bits from input ciphertext to select a particular word of the proper type from the dictionary.

d. Output the word.

e. Repeat until end of input ciphertext.

**The advantages of this technique:**

a. The ability of hiding either plain text or ciphertext.
b. High security by using both steganography and cryptography in hiding and extracting processes.
c. Capacity of hiding message by using big and flexible dictionaries.
d. Difficult steganalysis of hidden message because the system has the property to generate more than one hidden message.

**The Vulnerable aspects of this technique:**

a. The most obvious problem with the manual method is that it takes too long to enter large lists.
b. Nicetext focuses on creating large, sophisticated dictionaries with thousands of words, these words demand soft categories for forward and reverse transformation.

### 4.6.Linked Lists Technique:

Most techniques of data hiding in texts aim to hide information in **text files**. [12] Providing a new approach to hide data in word document files. The provided technique uses a graphical editor, which is designed according to V.C++ and software engineering concepts to be similar as word processor. Linked lists are used as a searching technique between two objects: the cover which contains all characters of the hidden message and the hidden object which must be less than or equal to the cover object.

Unlike previous techniques this method does not use encoding and decoding, compression and decompression or ciphering and deciphering because it hides full character randomly in the linked list data field. For example: if the cover is constructed from (200) characters, then the system can hide maximally (200) characters. The hiding operation can be done by storing the positions of appearances of the hidden message in independent object like array and then extract the content of the array to get the hidden message [12].

**The advantages of linked list technique:**

a. It need encoding, compression or other sub operations, just hiding and extracting processes.
b. The amount of hidden message is equal to the cover message.
c. It hides data in Word Document Files.
d. It can be combined with other techniques like white space.
e. No need for the cover message after accomplishing hiding process and saving the stego file (the cover can be changed to other object like image, sound or video).
f. After applying the system to some examples it is obvious that the size of the hidden message is (100%) relative to the cover size. In addition if linked list is applied in combination with white space technique then the size of hidden message will increase to (110%).

**The Vulnerable aspect of linked lists:**

Steganalysis can be done for this technique by trying to access memory locations.

## 5. Conclusion:

Most of existing techniques of data hiding in texts are similar in hiding and extracting processes and sub processes. These similar operations

can be seen in table (1), which gives a comparison between most of text hiding techniques.

From the above table, it is obvious that most of hiding techniques use multiple sophisticated operations like: ciphering, deciphering, encoding, decoding, compression and decompression. The advantages of these characteristics (multiple processes) are the increasing of security of hidden message. In spite of using these sophisticated operations,that does not prevent steganalysis of hidden message in the texts, and still there are some weaknesses as with other information hiding techniques, which bring similar vulnerabilities. However most of steganalysis techniques focus on steganalysis of images, this property decreases the detection of hidden message in texts. In addition , the popularity of the text file is due to the characteristics of good steganography system, which are: large capacity, high-level security and high imperceptibility, data hiding in texts can be characterized by the following properties:

a. Some of techniques of data hiding in texts are characterized by hiding more information (Capacity) like Dictionary, CFG and linked list. In addition most of techniques are characterized by the ability to be combined with other techniques like Dictionary, CFG and linked list. These properties will increase the capacity of the hidden message. Figure (4) shows the estimated capacity of most of data hiding in texts.
b. Some techniques use ciphering (and/or) key processes, in hiding process like Nicetext techniques and white space techniques. This property will increase the security of hidden message, because of using both steganography and cryptography.

**References:**

[1] Abdul Wahab, H., B., **Information Hiding in Written Text Using Context Free Grammar (CFG)**, MSc. Thesis, University of Technology, Department of Computer Science and Information System, Baghdad, 2001.

[2] Al-Shamkhy, R., A., **Hiding Text in Text Using Dictionary Method,** MSc, Thesis, Department of Computer Science and Information System, University of Technology, Baghdad, 2001.

[3] Anderson, R., **Information Hiding,** First International Workshop, Cambridge, 1996.

[4] Bender, W., and Gruhl, D., **Information Hiding & Data Hiding,** IBM System Journal Vol. 35. http://www. Research.com

[5] Chandramouli, S., T., **A Mathematical approach to Steganography**, Electrical and Computer Engineering, Stevens Institute of Technology, California, 2002.

[6] Chandramouli, S., T., **Active Steganalysis of Sequential Steganography**, Electrical and Computer Engineering, Stevens Institute of Technology, 2000.

[7] Chapman, M., T., **Hiding the Hidden, A Software System for Concealing Ciphertext As Innocuous Text**, MSc. Thesis, University of Wisconsin-Milwaukee, Department of Computer Science, 1998.

[8] Inque, S., and Makino, K., **A Proposal on information hiding methods using XML,** University of Tokyo, 2001.

[9] johnson and jajodia , S., **information Hiding Steganography and Water Marking and Countermeasures , Kluwer** , Academic Publisher Group, 2001.

[10] **Johnson** and jajodia , S., **information Hiding** Second International Workshop, Oregon ,1998.

[11] Popa, R., **An Analysis of Steganographic Techniques**, the Politehnica" University of Timisoara, Faculty of Automatics and Computers, Department of Computer Science and Software Engineering, 1998.

[12]Tome, A., M., **Design and Implementation of Text's Graphical Editor for Informational Hiding Purposes,** MSc, Thesis, Informatics Institute for Post Graduate, Baghdad, 2004.

**Internet References:**

[13] **How SNOW works,** 2002.http://www.madchat.org/crypto/stegano/java/snow.desc.html

[14] **Hides Arbitrary Data in any Text,** German, 2001. http://www.Texthide.com

[15]**Steganography and Steganalysis ,** 2001. http://www.Giac.oro/Practical/Qureshi-Waheed.Doc

[16] **Rich Text Media,** 2001. http://www.Markitingterms.Dictionary/richmedia.html

Johnson a

Johnson a

A text can be defined in a closed frame or line border and constructed from words written in a determinate language.

يكتب النص عادة في مساحة مغلقة يحدها اطار ويتكون النص من مجموعة من الكلمات بلغة محددة .

Figure (1) A model for Arabic and English texts media

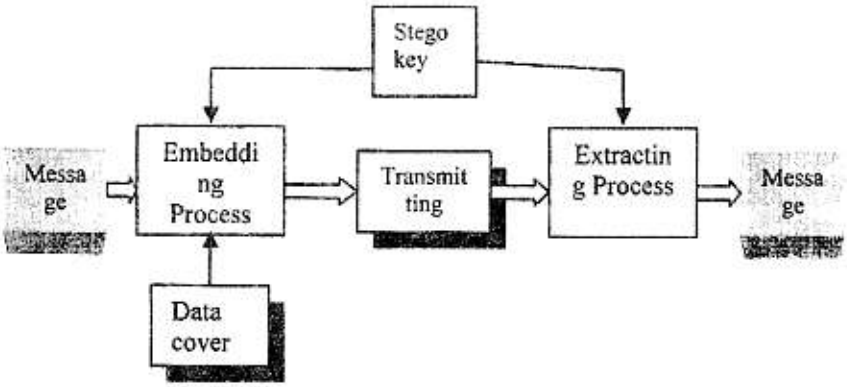



Figure (2) General model of hiding data in texts

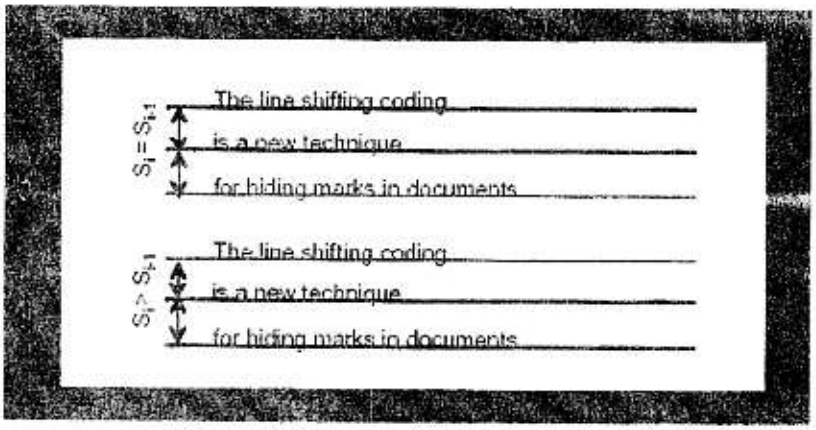
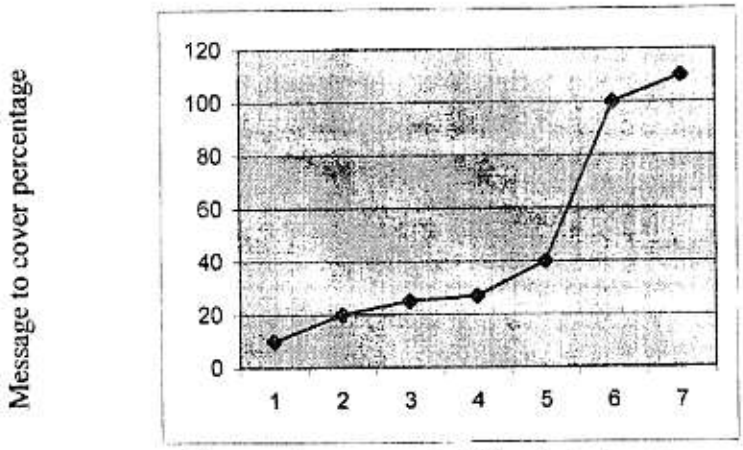



Figure (3) Example of Line Shift – Coding.

Table (1) Comparison between techniques of data hiding in texts

| Index | Processes / Hiding Technique | Encoding Decoding | Compression Decompression | Ciphering Deciphering | Key | Type of File | Size of Hidden message Relative to the Cover |
|---|---|---|---|---|---|---|---|
| 1 | SPY | No | No | No | Yes | Any Type | 10% |
| 2 | Line Shift Coding | Yes | No | No | No | Text File | 10% |
| 3 | White Space | Yes | Yes | Yes | Yes | Text File | 20-30% |
| 4 | Syntactic | Yes | Yes | Yes | Yes | Text File | 25-30% |
| 5 | Dictionary | Yes | Yes | Yes | Yes | Text File | 40-50% |
| 6 | Linked Lists | No | No | No | No | Word Document | 100% |
| 7 | Combination of Linked lists& White Space | Yes | No | Yes | No | Word Document | 130% |




The number represents hiding techniques

Figure (4) The Estimated Capacity of data Hiding in Texts