Available online at: www.basra-science-journal.org

ISSN –1817 –2695

Steganography in MS Excel Document Using Unicode System Characteristics

Aliea Salman Saber and Wid Akeel Awadh

Computer Science Department ,Basrah University

Basrah . Iraq

Email: alieea@yahoo.com

Email: alishashim2009@yahoo.com

Received 20-3-2012, Accepted 23-1-2013

Abstract

The massive explosion in computer technology and communication has encouraged the development of steganography and watermarking, breathed the spirit of steganography, and made it the top security technique to protect and preserve the rights and privacy. So, steganography is one of the important types of information hiding techniques that has evolved a lot in the recent years.

Most text steganographic methods take the formatted text documents, such as, MS Word, PDF, PPT …etc., as a cover to hide secret information. This study concerns the steganography in MS Excel document and proposes a new steganographic method for hiding information efficiently by Unicode system characteristics technique. In the Unicode standard, there are two different codes for seven numbers (9 ,8 ,7 ,3 ,2 ,1 ,0) in Arabic and Persian Language. The seven numbers of (9 ,8 ,7 ,3 ,2 ,1 ,0) have the same shape but different codes in Unicode table. As a result, we can hide information in MS Excel Document, by using one of these two codes.

This proposed method has a high capacity. It can hide one bit in each number in the cover_file, and it is not make any apparent changes in the original text. So it satisfies perceptual transparency. In our method, we have increase the level of security by encrypting the secret message before embedding it by using Advanced Encryption Standard (AES-128) algorithm.

Keywords MS Excel document, Secret message, Cryptography, Unicode standard, Text steganography method, Human visual system.

1. Introduction

The application of computer in real life is increasing every day. So, the need to secure data is becoming more and more essential part of message or data transfer. Information security became a part of our daily life. Among the different techniques, hidden exchange of information is one of the concerns in the area of information security. Various methods like cryptography, steganography, coding have been used for this purpose. However, during recent years, steganography has attracted more attention in recent years [1].

Steganography is the art and science of writing hidden messages in such a way that no-one, apart from the sender and intended recipient, suspects the existence of the message, a form of security through obscurity [2]. The word "*steganography*" is of Greek origin and means "concealed writing" from the Greek words "*steganos*" meaning "covered or protected", and "*graphein*" meaning "to write".

In steganography, the information is hidden in a cover media so nobody notice the existence of the secret information. Steganography works have been carried out on different medium such as images [3], video clips [4], music and sounds [5].Steganography scheme can be view as [6] **Cover_medium + hidden_data + steg_key = steg_medium**

There are three important parameters in designing steganography methods: perceptual transparency, robustness and hiding capacity. These requirements are known as "the magic triangle" [7]. Steganography may have different applications. For example, it can be used by medical doctors to combine explanatory information with X-ray images. It can be useful in communication for codes self error correction. It can embed corrective audio or image data in case corruption occurs due to poor connection or transmission, copyright protection, preventing e-document forging, and other application [8].

There is the major distinction between steganography and other methods of exchange hidden information. For example, in cryptography method, people become aware of the existence of information by observing coded information although they are unable to comprehend the information. However, in steganography, nobody will understand the existence of information in the resources [9].

Text steganography is the most difficult kind of steganograhy which is due largely to the relative lack of redundant information in a text file as compared with a picture or a sound file [10]. The structure of text document is normally very similar to what is seen, while in all other cover media types (audio, picture, video), the structure is different than what we observe, making the hiding of information in other than text easy without a notable alteration. The advantage of prefer text steganography over other media is its smaller memory occupation, simpler communication, sends more information and needs less cost for printing, as well as some other advantages.

Today, the computer systems have facilitated hiding information in texts. The rang of using hiding information in text has also developed. From among the most important of these technologies, one can name of hiding information in electronic texts, web pages and documents.

2. Unicode Standard

The Unicode standard is the international character-encoding standard used for presenting the texts for computer processing. This standard is compatible to the second version of ISO/IEC 10646-1:2000 and have the same characters and codes as ISO/IEC 10646.

Unicode enables us to encode all the characters used in writing the languages of the world. This standard uses the 16-bit

encoding, which provides enough space for 65536 characters; that is to say, it is possible to specify and define 65536 characters in different moulds such as numbers, letters, symbols, and a great number of current characters in all different languages of the world. This standard covers a mathematical and technical symbols, punctuation marks, arrows, and miscellaneous marks. Moreover, because of the wideness of the space dedicated to the characters, this standard also includes most of the symbols necessary for high-quality typesetting. The languages whose writing systems can be supported by this standard are Latin (covering most of the European languages), Cyrillic (Russian and Serbian), Greek, Arabic (including Arabic, Persian, Urdu, Kurdish), Hebrew, Indian, Armenian, Assyrian, Chinese, Katakana, Hiragana (Japanese), and Hangeul (Korean). An Arabic Unicode table (takes the range 0600-06FF, form_A) represents standard forms of all characters used in Arabic language, and another Unicode table (takes the range FE70-FEFF) represents Arabic presentation forms _B that has all Arabic characters with isolated form.

Unicode table has been developed to cover the characters of the languages which use Arabic writing system. Among these languages we can mention Persian, Urdu, Pashto, Sindhi, and Kurdish. This standard has detailed and careful explanations about the implementation methods including letters-connection method, the exhibition of the right-to-left and bi-direction texts [11].

3. Advanced Encryption Standard (AES)

This standard specifies the Rijndael algorithm, a symmetric block cipher that can process data blocks of 128 bits, using cipher keys with lengths of 128, 192, and 256 bits. Rijndael was designed to handle additional block sizes and key lengths; however they are not adopted in this standard.

Throughout the remainder of this standard, the algorithm specified herein will be referred to as “The AES algorithm” The algorithm may be used with the three different key lengths indicated above, and therefore these different “flavors” may be referred to as “AES-128”, “AES-192”, and “AES-256”. For the AES algorithm, the length of the input block, the output block and the State is 128 bits. This is represented by Nb = 4, which reflects the number of 32-bit words (number of columns) in the State. For the AES algorithm, the length of the Cipher Key, K, is 128, 192, or 256 bits. The key length is represented by Nk = 4, 6, or 8, which reflects the number of 32-bit words (number of columns) in the Cipher Key.

For the AES algorithm, the number of rounds to be performed during the execution of the algorithm is dependent on the key size. The number of rounds is represented by Nr, where Nr = 10 when Nk = 4, Nr = 12 when Nk = 6, and Nr = 14 when Nk = 8[12]. The only Key-Block-Round combinations that conform to this standard are given in table1.

Table1: Key-block round combinations

	Key Length (Nk words)	Block Size (Nb words)	Number of Rounds (Nr)
AES-128	4	4	10
AES-192	6	4	12
AES-256	8	4	14

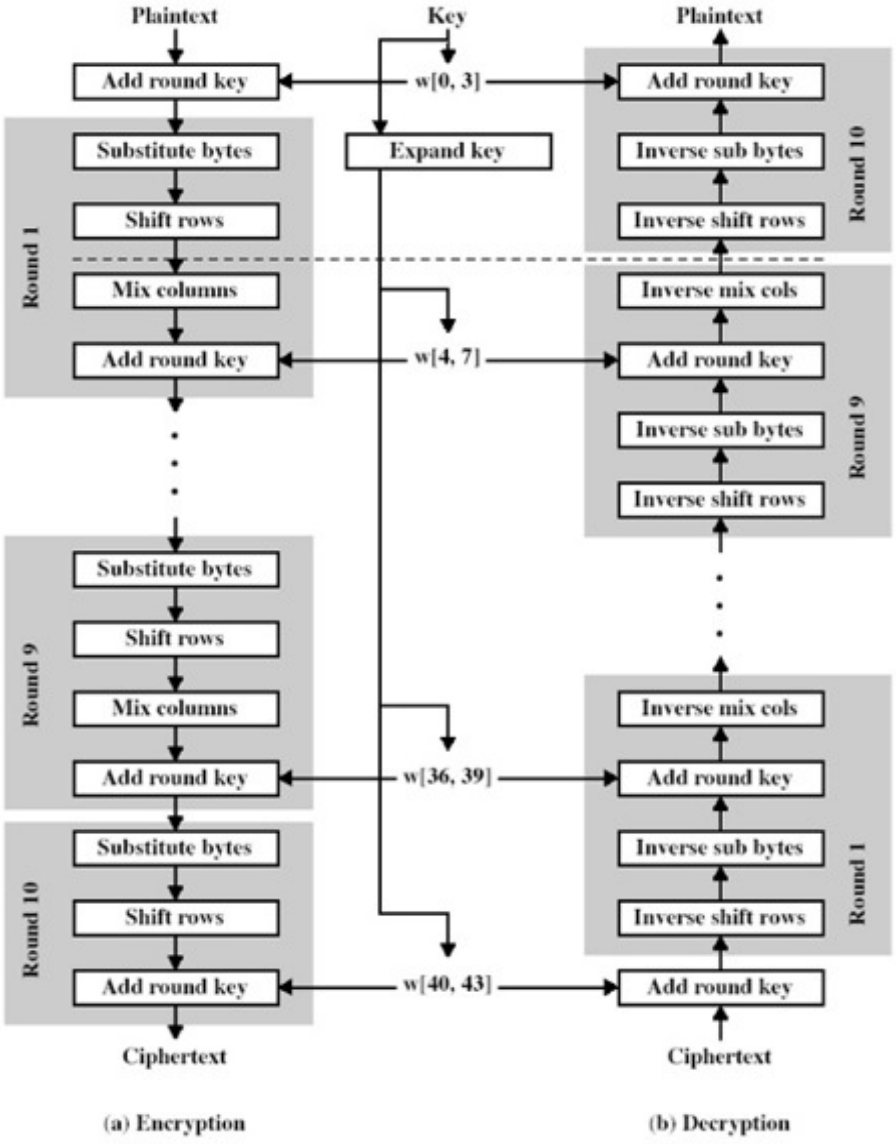


Figure (1): AES Encryption and Decryption

4. Previous Works

Open space method, hiding information is done through embedding information by utilizing the space characters in a plain text document. In this method these space characters are placed at the end of paragraph lines or between the words. Open space method could be used more widely for the Excel document. Most cells in a MS Excel document are empty so that these can be used for embedding space without changing the document's appearance [13].

Line and word shifting methods, shifting text lines vertically and shifting words horizontally could embed information in these methods. Security of this method depends on the availability of varying the distances between words and lines to puzzle intruders. 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. But these methods can not be directly applied in Excel document. However, change the height of rows or the width of columns slightly to embed information [14][15].

Character features methods, some features of characters in a text are changed to embed information in these methods. For example, it can change the font color slightly to embed the information in the text. Steganography based on character features can hold a large quantity of secret information without making normal readers aware of the existence of such information in the text. Most steganography methods based on character feature are suitable for Excel document, such as changing the text size, color, alignment…etc. However, modifying a single character without affecting the whole string in the cell is not possible in an Excel document. Therefore, the steganographic effect in an Excel document is not as good as that in the Word document [15].

Text abbreviation or acronym**,** another method for hiding information is the use of abbreviations. In this method, very little information can be hidden in the text. For example, only a few bits of information can be hidden in a file of several kilobytes, as in the case of Excel document [16].

File structure methods**,** many documents contain readily available spaces that can be used inside their file structures. In these methods, some unused space is used to embed information. Meta-data is an example; it is ingrained in file structures but not visible to the user without special tools. Some files also have unused space. In these spaces, bits can be overwritten without any adverse or obvious effect on the file. These spaces create an opportunity to hide information. In Excel file, there are 420 bytes continuous block below the header where can embed data easily [17].

Text rotates in MS Excel documents, this method is implemented by slightly rotating the angle of the text inside the cell to reduce the visible detection of the embedded information. Measuring the text angle of the cell retrieves the secret information. Experiments for different threshold in the algorithm are presented and the result show the proposed method not only has a good imperceptibility but also achieve high embedding rate while most of cell in Excel document are short in length [18].

5. Proposed Method

This study proposes a novel steganographic method by Unicode system characteristic, to embed secret bits in Excel worksheet. Because of structure characteristic of Excel file that contain a large amount of numeric data, we propose a new and innovative way to exploit the use of Arabic numeric data to embed secret bits.

This study proposes a novel method by using Unicode system characteristic technique (similar numbers with different codes in Unicode table) to embed the secret message in the Excel worksheet. There is on the Unicode table seven numbers in Arabic and Persian language (0, 1, 2, 3, 7, 8, 9) have the same shape but different code as shown in the table2. So can be hidden secret message in Excel worksheet through using one of the two codes for these numbers in the Unicode table.

Table2: codes of Arabic and Persian numbers in Unicode table

Arabic Numbers	Code of Arabic Numbers in Unicode Table	Persian Numbers	Code of Persian Numbers in Unicode Table
0	0660	0	06F0
1	0661	1	06F1
2	0662	2	06F2
3	0663	3	06F3
4	0664	۴	06F4
5	0665	۵	06F5
6	0666	۶	06F6
7	0667	7	06F7
8	0668	8	06F8
9	0669	9	06F9

Our method work in two stages, **Embedding Stage** and **Extraction Stage**.

❖ **Proposed algorithm method For Embedding:-**

Input: - Cover MS Excel document, Plain-text.

Output: - Stego MS Excel document.

Step1: Enter plain-text, and then, we encode the plain-text into secret message by using (AES-128) algorithm.

Step2: Open cover–text (MS Excel document).

Step3: Convert the secret message into binary bits (0, 1).

Step4: Check if the bits of the secret message less than the bits of the cover–text, if condition is true continue to step 5, otherwise, go to step 7.

Step5: The size of the secret data is hidden at the beginning of sheet by the process describe in step 6, in order to prevent the extraction of the additional data at the time of the extraction from the stego-text.

Step6: To embed the size and secret data, Take each non empty cell in the sheet, and check each symbol in a cell:

- if the symbol is one of the Arabic numbers (،0 9 ،8 ،7 ،3 ،2 ،1), and if the bit "1" is to be hidden, we change the code of Arabic numbers to code of Persian number, and do not change the code of Arabic number for hiding bit "0".

Step7: end embedding process.

❖ **Proposed algorithm method For Extraction :-**

Input: - Stego MS Excel document.

Output: - Extracted plain text.

Step1: Stego MS Excel document.

Step2: Extract the size of hidden data that embedded at the beginning of the text and extract the hidden data by the process describe in step 3.

Step3: Take each cell in the sheet, and check each symbol in the cell:

- If the code of symbol is code of (،8 ،7 ،3 ،2 ،1 ،0 9) in Persian language, that mean the bit "1" has been hidden in the text.
- If the code of symbol is code of (،8 ،7 ،3 ،2 ،1 ،0 9) in Arabic language, that mean the bit "0" has been hidden in the text.

Step4: Collocate the sequence of extracted bits.

Step4: Decoded the secret bits into plain text by (AES-128) algorithm.

Step5: End extraction process.

6. Experiment Results

The proposed method was implemented by using Microsoft Excel 2003 and Microsoft Visual Basic6.0 software and running on the Intel core(TM) 2 Due 2.4 GHz CPU, and 1 GB RAM hardware platform. We tested the implemented system by taking different messages of different length and hiding them in some Excel sheets of different size. The results that are got from experiment are recorded and can be summarized in Table3.

Table3: Example of text steganography method proposed

Secret Massage	Cover File		
Number of bits(bits)	**File type**	**Number of numeric data**	**Capacity(%)=** Number of bits/number of numeric data
128	Excel	477	26.8
512	Excel	797	64.2
1152	Excel	1180	97.6
1792	Excel	2996	59.8
2816	Excel	3159	89.1

The use of Unicode system characteristics does not make any apparent changes in the original text by hiding data, while in most text steganography methods, such as line shifting, word shifting, and especially the open spaces method, it is evident that the text has been changed. Therefore, even if the reader has the original text, it is impossible for him to realize the hiding of the data by merely observing the appearance of the text. However, the original texts are usually not available to the observer with text steganography methods. Therefore, the main goal of text steganography that is, the impossibility of detection of the presence of data-has been achieved. In case of printing the stego-text containing the hidden data, the hidden data will be lost, because, as mentioned earlier, due to the hiding of the data in the text, the text appearance remains unchanged and only the internal structure of

the saved file is changed. Consequently, the hidden data will be lost because of losing the internal data of the file and we cannot extract the hidden data from the printed copy of the text. As a result, this method is limited to hide data in electronic documents (e-document).

7. Conclusion:-

This paper presents a steganography technique which useful for embedding data in Excel document. It benefits from Unicode system characteristics to hold secret information**.** Although various data hiding methods which based on the text documents are focus on the formats of TXT, MS Word, PDF, PPT…etc, a few studied methods of embedding data in MS Excel document Compared with other text formats, the expression form of Excel document data are quite different. Therefore, steganography which is based on Excel document deserves to be further investigations.

Our method has an excellent perceptual transparency because the stego text which the user sees is exactly similar to the original text. Therefore, the hiding capacity of our method is very high, we hide "1" bit in each number in Excel sheet. However, it is robust to digital copy-past operation, which means that copying and pasting the text between computer programs preserve hidden information.

REFERENCES

[1] E. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Information Hiding-A Survey", Proceedings of the IEEE, pp. 1062–1078, 1999.

[2] Steganography, http://en.wikipedia.org.

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

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

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

[6] Kessler, c. Gary, "An Overview of Steganography", the Computer Forensics Examiner issue of Forensic Science Communications, July 2004**.**

[7] N. Cvejic, Algorithms for Audio Watermarking and Steganography. Finland: Oulu University Press, 2004.

[8] N. F. Maxemchuk and S. Low, "Marking Text Documents", in Proceedings of the IEEE International Conference on Image Processing, Santa Barbara, CA, USA, pp. 13–16, 1997.

[9] J.C. Judge, "Steganography: Past, Present, Future", SANS white paper, November 30,2001, http://www.sans.org/rr/papers/index.php?id=552, last visited: 1 May 2006.

[10] Memon,jibran a. and khowaja, k. and K. Hameedullah, "EVALUATION OF STEGANOGRAPHY FOR URDU /ARABIC", Journal of Theoretical and Applied Information Technology, 2008**.**

[11] The Unicode Standard, URL: http://www.unicode.org, last visited: 31 September 2011.

[12] Specification for the ADVANCED ENRYPTION STANDARD (AES), Federal Information Processing Standard Publication 197, November 26, 2001.
[13] M. H. Shirali-Shahreza and M. Shirali-Shahreza, "Persian/Arabic CAPTCHA", IADIS International Journal on Computer Science and Information Systems (IJCSIS), pp. 63–75, 2006.
[14] K. Bennett, "Linguistic Steganography: Survey, Analysis, and Robustness Concerns for Hiding Information in Text", Purdue University, CERIAS Tech, 2004.
[15] K. Rabah, "Steganography–The Art of Hiding Data", Information Technology Journal, pp. 245–269, 2004.
[16] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for Data Hiding", IBM Systems Journal, pp. 313–336, 1995.
[17] Castiglionea, A., A.D. Santisa and C. Sorienteb, Taking advantages of a disadvantage: Digital forensics and steganography using document metadata. J. Syst. Software, 80: 750-764, 2007.
[18] Bin Yang, Xingming Sum, ligyun Xiang, Zhiqiang Ruan and Ruizhen, "Steganography in MS Excel Document using Text-rotation Technology", Information Technology Journal, 10(4):889-893, China, 2011.

التضمين في ورقة العمل أكسل بأستخدام مواصفات نظام اليونيكود

علياء سلمان صابر و ود عقيل جواد
قسم الحاسبات/ كلية العلوم / جامعة البصرة

الخلاصة:

أن ألأنفجار الهائل في تقنية الحاسوب وألأتصالات شجع على أحياء وتطوير الكتابة المخفية والعلامة المائية, حيث نفخت هذه التقنيات بروح الكتابة المخفية وجعلتها تتصدر تقنيات ألأمنية من حماية وضغط وحقوق وخصوصية. لذلك تعتبر الكتابة المخفية من أهم أنواع التضمين الذي تطور كثيرا في السنوات الماضية.

أغلب طرق التضمين النصي تستخدم نصوص من نوع Word, PDF, PPT كملف غطاء لأخفاء المعلومات السرية. أما هذا المقال فأنه يتناول طريقة جديده لأخفاء المعلومات في ملفات من نوع أكسل بألأعتماد على تقنية مواصفات نظام اليونيكود (الرموز المختلفة للأرقام المتشابهة في جدول اليونيكود).

ويوجد في مواصفة اليونيكود رمزان مختلفان للأرقام (0, 1, 2, 3, 7, 8, 9) في اللغتين العربية والفارسية. هذه ألأرقام لها نفس الشكل وكودين مختلفين في جدول اليونيكود لذلك يمكن أخفاء البتات السرية من خلال أستخدام أحد الكودين المختلفين لهذه ألأرقام.

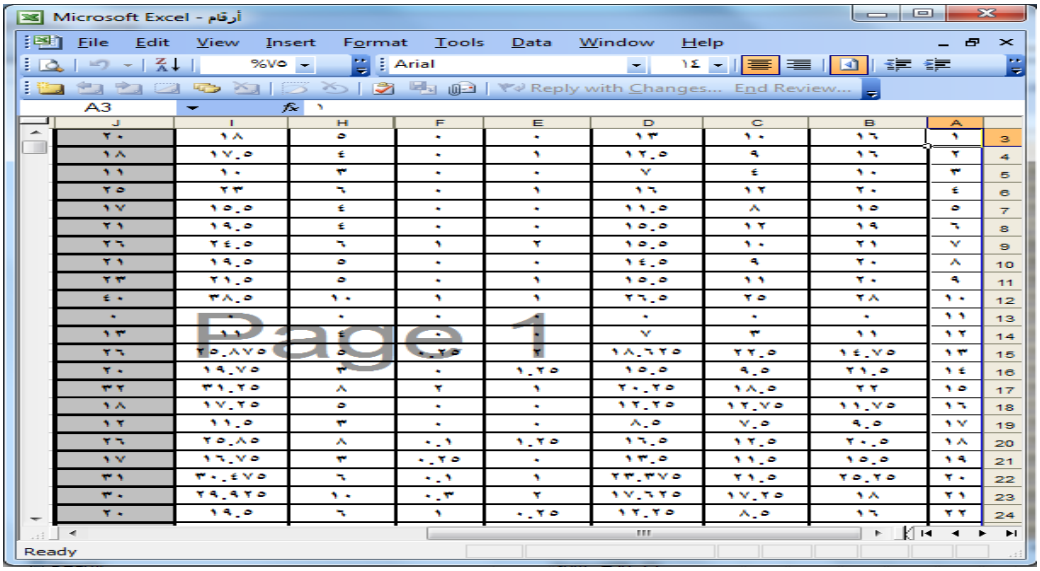
هذه الطريقة تمتلك سعة أخفاء عالية في ملفات ألأكسل نظرا لما تتمتع به هذه الملفات من كثرة البيانات الرقمية, كما أنها لا تسبب أي تغيير في شكل النص ألأصلي أي أنها تحقق شفافية عالية. في طريقتنا حققنا مستوى عالي من ألأمنية من خلال تشفير الرسالة قبل تضمينها بأستخدام طريقة التشفير القياسي المتقدم (AES).

EXAMPLE :-

Plain-Text: -

لمحة تاريخية عن كلية العلوم
التأسيس : 1964
أهدافها : رفد البلد بخريجيها من حملة البكلوريوس والماجستير والدكتوراه في ألأختصاصات الكيمياء وعلم الحياة والفيزياء والرياضيات والحاسبات وعلم الارض

Cover-Text :-



Steg-Object :-