EasyChair Preprint

№ 4214

An Efficient Framework for Text Document Security and Privacy

Umair Khadam, Muhammad Munwar Iqbal, Leonardo Mostarda and Farhan Ullah

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

September 19, 2020

An Efficient Framework for Text Document Security and Privacy

Umair Khadam [1], Muhammad Munwar Iqbal[2], Leonardo Mostarda[3], Farhan Ullah[4]

[1,2] Department of Computer Science, University of Engineering and Technology,
Taxila

[3]Computer Science Department, Camerino University, 62032 Camerino, Italy

[4]College of Computer Science, Sichuan University, Chengdu 610065, China

umair_khadim@live.com, munwariq@gmail.com, leonardo.mostarda@unicam.it, farhankhan.cs@yahoo.com

Corresponding Author: Farhan Ullah (farhankhan.cs@yahoo.com)

Abstract: Nowadays, with the help of advanced technologies, an illegal copy of digital contents can be shared easily. Which rise copyright and authentication problems. Digital text documents are generated and shared daily through different internet technologies such as the cloud, etc. The protection of these documents is a challenging task for researchers. In the past, steganography, cryptography, and watermarking techniques have been applied to resolve the copyright problem. However, most of the existing techniques are applicable for only plain text or protecting the document on the local paradigm. In the said perspective, we proposed a new technique to solve the problem of copyright and authentication on local and cloud paradigms. In this paper, we utilize some custom components of MS Word document for concealing the watermark into a text document. These components are not referred to as the main document and will not modify the content and format. The experimental analysis and results prove that the proposed method improves the watermark capacity, imperceptible and robust against formatting attacks.

Keywords: Steganography, Cryptography, Document Security, Copyright Protection, Digital Watermarking.

1. Introduction

In today's digital world, secure communications are required with rapidly evolving internet technology. Information security has gain importance in many areas like government applications, data storage, e-commerce, e-signature, banking, personal and corporate communication. The purpose of information security is to prevent third parties from accessing information for any purpose [1]. Data breaches are a significant challenge in the modern digital world because the critical data of the organization must be protected against unauthorized access. In the last five years, almost 10 billion records have been lost, exposed or stolen, with an average of five million records per day affected. Advanced digital technologies such as the cloud brought unlimited benefits to users, but they also cause problems for the original owner of the data against illegal copies. In the past, steganography, cryptography, and watermarking techniques have been used to provide ownership verification. Digital watermarking plays an important in this field of research. Where, a secret message also called watermark is embedded into the host content without compromising the data integrity [2].

When an illegal act occurs the same watermark, is used for ownership verification. Digital watermarking is classified into audio, video, text, and image, whereas, most of the watermarking research focuses on audio, video, and image [3]. The text watermark has now become very popular and become a hot area of research because text documents are almost part of all private and public sector organizations and need copyrights protection.

In recent years, cloud computing has been the most significant development in the field of information technology. It provides services to all organizations such as educational institutes, healthcare, banking, etc. via the internet by the pay-as-you-go model [4]. The security of data in cloud computing is the main issue for users. It is essential to ensure data security in many positions in data rest [3]. Digital text watermarking is not considered yet in the context of cloud computing. None of the existing digital watermarking technique provides secrecy to a text document in the cloud computing paradigm. We proposed an efficient digital watermarking framework based on MS Word document custom components that protect the text document authentication and verification. , many researchers worked in the field of digital text watermarking, and numerous techniques have been proposed for text document security and privacy. Three major categories of digital text watermarking are statistical, linguistic, and format-based techniques, as shown in Fig 1. Linguistic based techniques are divided into two significant types semantic and syntactic. In general, the semantic-based technique uses the synonym substitution method, where words synonym is used for embedding the watermark information. In the syntactic techniques the punctuation marks like full stop (.), comma (,), colon (:) and semicolon (;) etc. are placed to conceal the watermark in cover file [5, 6]. In the word and line spacing techniques, the words or lines are shifted up down to some degree to hide secret data.

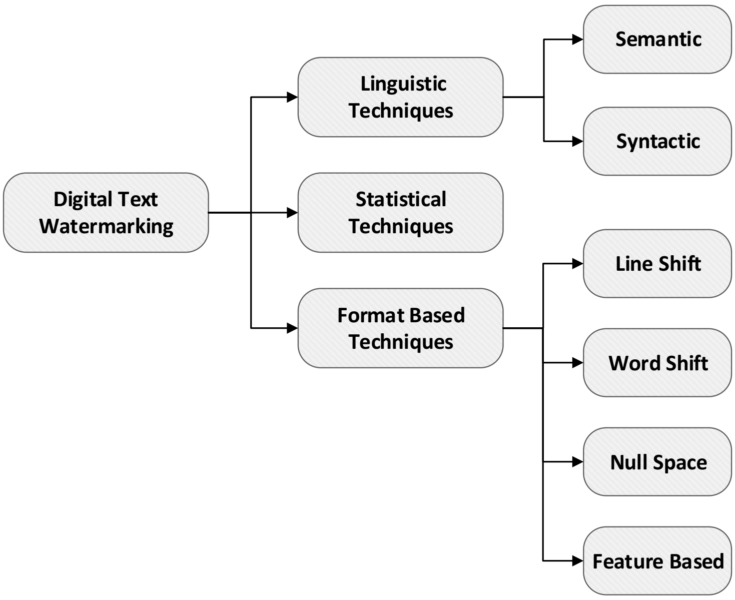

Fig 1. Digital text watermarking techniques classification

In the past, several techniques are designed for text documents. Khairullah et al. [7] presented a method based on invisible characters. The proposed technique sets the invisible characters to foreground colors such as the tab, space, or the carriage return characters, which can be obtained 24 bit per character. Similar English Font Types (SEFT) technique is proposed in [8], which utilizes the same English font for text watermarking. First of all, three different fonts are chosen which are identical, and then 26 characters and spaces are represented by a triple of capital letters. The proposed scheme is not considered as robust, because if the spaces between the text are removed then the watermark is ruined. Nuzhat et al. [5] introduced a zero text steganography approach that is based on multilayer partially homomorphic. The proposed technique implements multilayer security on a secret message. Rajeev et al. [9] suggested a technique that is based on Huffman compression. The proposed technique uses email forwarding data to conceal watermark and not consider robust.

Khosravi et al. [10] proposed an information hiding technique for PDF (Portable Document Format) based on justified text. First, the secret message is compressed by Huffman coding, then some unique lines of PDF files are chosen to conceal the information. The embedding operation takes place by replacing the added spaces with the regular spaces of the host rules. Alghamdi et al. [11] introduced a text steganography technique for the Arabic language. Markov Chain (MC) is implemented for encoder and decoder combined with Huffman Coding. The upper and lower bound are also computed for the stego-text. The proposed technique is format independent and less robust against attacks. Long et al. [12] suggest a coverless method based on web text, where a large number of web pages are used to conceal the secret message. The mature search engines are applied to obtain the secret information that is associated with web pages. Rizzo et al. [13] introduced a structural approach that protects digital content small portions. This approach is suitable for Latin symbols and white spaces which is based on homoglyph character substitution. Hence, the proposed system increases the hiding capacity of watermark but, not robust.

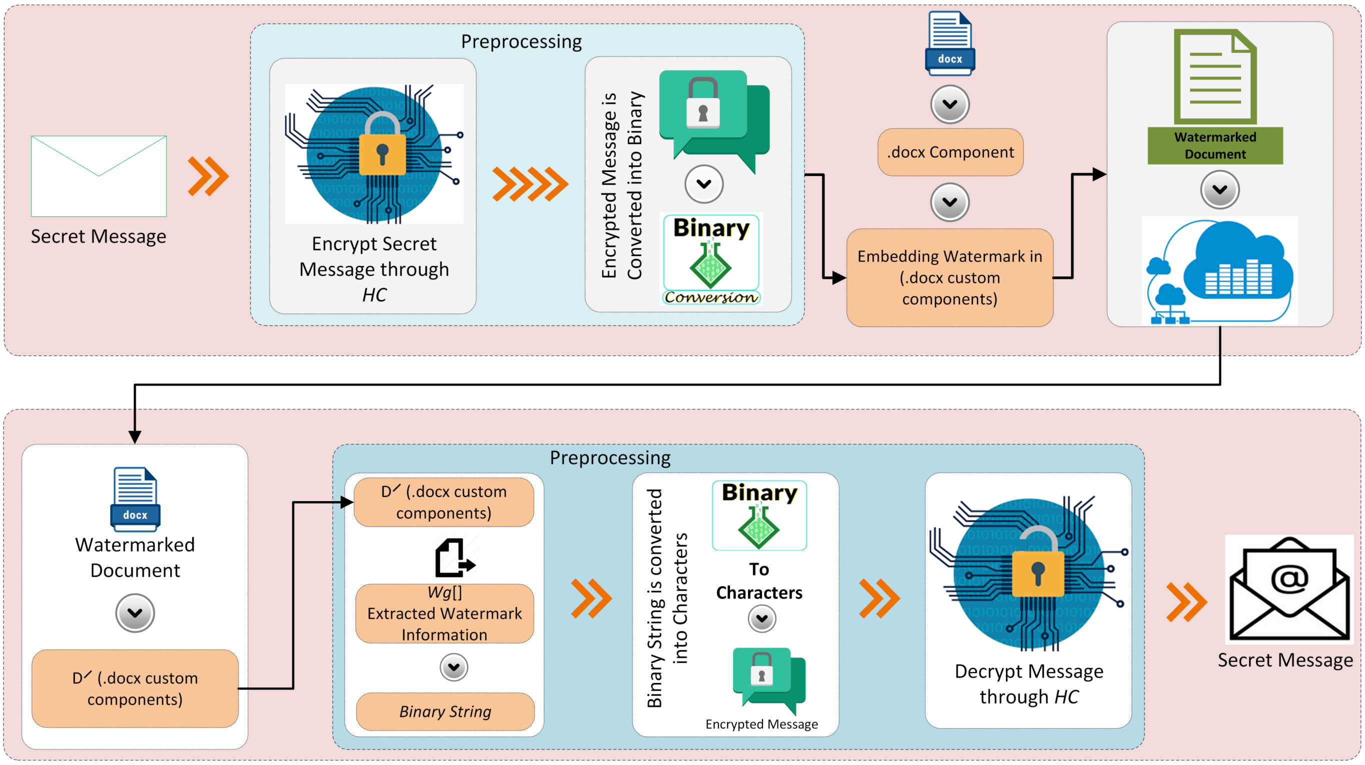
Preprocessing
Secret Message
Encrypt Secret Message through *HC*
Encrypted Message is Converted into Binary
Binary
Conversion
docx
.docx Component
Embedding Watermark in (.docx custom components)
Watermarked Document
docx
Watermarked Document
D′ (.docx custom components)
Preprocessing
D′ (.docx custom components)
Wg[]
Extracted Watermark Information
Binary String
Binary String is converted into Characters
Binary
To Characters
Encrypted Message
Decrypt Message through *HC*
Secret Message

Fig. 2. The proposed model for digital watermarking

2. Proposed Method

A novel framework is proposed here for text document ownership verification and copyright protection based on Microsoft Word (*MSW*) document custom components, as shown in Fig. 2. Furthermore, this section covers the watermark (secret information) embedding and extraction process. Today, the *MSW* document is a critical part of all public and private organizations. The copyright protection and ownership verification of these documents are essential. We introduced a novel content free watermarking technique for text document security and privacy on local and cloud paradigms. We use the custom components of *MSW* document for concealing the watermark information. The *MSW* custom components are appropriate for watermarking, because these components are not part the main document. The process of watermarking does not change the document original contents. The main reason for using these components is that it can mask an enough watermark. In addition, the proposed technique is robust against format-based attacks, semantic-based attacks, and content-based attacks. Huffman Coding is used for encryption and decryption. Secret Message (M_S) and original document (D_O) are given as input in the proposed model, M_S is encrypted through Huffman Coding. The detailed procedure of encryption is presented in the next section. The encrypted message M_E is transformed into a binary string, then divided it into n groups. The custom components of text document D_O are checked, and the groups of M_E are inserted into that components. The watermark information does not affect the document original content and does not interfere with imperceptibility. After hiding the watermark information, the watermarked document is generated and shared via different communication technologies.

2.1 Huffman Coding

Huffman Coding is one of the high data compression rate algorithms, which gives the variable-length code to input characters. On the bases of the character's frequencies Huffman tree is constructed, which determines the length of the code. Small codes are given to most frequent characters, and bulky code is assigned to less frequent characters. Huffman tree is responsible for the coding and decoding process from the sequence of characters to bitstream or vice versa. To avoid ambiguity a unique code is assigned to each character that should not be used with other characters [11]. In Huffman Coding data compression is achieved through binary allocation codewords of different lengths.

Let W belongs to the possible plain text, and $M = \{M1, M2, M3 \ldots\ldots Mn\}$, the plain text alphabet of P, and P belong W such that $P = P1, P2$. where P_i belongs to M. If W_i is the probability of P_i appearing in the plain text P, we have the Entropy of P defined by (1):

$$H(P) = -\sum_{m}^{i=1} W_i * \log W_i \qquad (1)$$

As the average number of bits to represent each symbol M_i belongs M. Moreover, H (P) leads to zero redundancy, that is, has the exact number of significant bits to represents P. The encoding produced by Huffman Coding is prefix-free and satisfies through (2):

$$H(P) = 1(HC) < H(P) + 1 \tag{2}$$

Where 1 is the weighted average length.

2.2 Watermark Embedding

The *MSW* document is a common type of text document throughout the world. It comprises a lot of custom components. These components are suitable for watermarking and authorized users to manipulate with it through programming. There are three main reasons that are why these components are appropriate for watermarking. Firstly, the watermark information is stored in custom components, which cannot affect the contents of the document. Secondly, without affecting the imperceptibility a large amount of watermark information is stored. Thirdly, it is robust, any command of *MSW* will not interrupt or delete the watermark. The Microsoft Visual Basic (VB) is used to store and retrieve the watermark information from the MSW document. The M_E is divided into groups before embedding using (1).

$$W_g = \left\{ \frac{i_w, \{w \mid w = 1, 2, \ldots.., n\}}{N_w} \right. . \tag{1}$$

Where W_g is total groups of watermark information, i_w is watermark information, N_w is the number of groups. The groups of watermark information W_g is dependent on $W_{obj}(D)$. $W_g[i]$ in embedded into the value attribute of custom objects. When all $W_g[n]$ is embedded into the D_O then D_w is generated and shared on the cloud.

2.3 Watermark Extraction

The objective of extraction is to extract the M_S and verify the document originality. In our system, the second phase of the proposed model describes the watermark extraction complete procedure, as shown in Fig. 2. The D_w document is given to the system as input, the list of interrupted components D□ are utilized for collecting the groups of watermark information $W_g[n]$. The groups of watermark information are concatenated and then converted into a binary string then characters. The Huffman Coding is used to decrypt the message, and finally, the secret message is recovered.

3. Experimental Classification Results and Analysis

In this section, the results of our proposed technique are analyzed on the bases of digital watermarking evaluation criteria. Which can be categorized into robustness, capacity, and imperceptibility.

3.1 Robustness Analysis

Robustness is a critical factor in digital watermarks, and it indicates that after applying various attacks, either 100% watermark information is restored or not. Different types of brute force attacks are applied to the watermarked document to verify its robustness. These attacks include attacks based on content and format. Table 1 presents the comparison of the proposed method with [14], [15, 16] against content and format-based attacks. The proposed technique is based on MSW custom components and any mutual commond cannot distrub the watermark. Table 1 shows that the proposed technique resistant against content and format-based attacks. The comparison demonstrates that the proposed algorithm presents improved results.

Table 1. The comparison of the proposed technique with existence techniques against content and format-based attacks

Techniques	**Formatting attacks**										
	Insert	**Delete**	**Replace**	**Copy**	**Paste**	**Font Size**	**Font Color**	**Font Weight**	**Paragraph Alignment**	**Line and Paragraph Spacing**	**Text Highlight**
Zhang et al. [14]	✓	✓	✓	×	×	✓	×	×	×	×	×
Aniello et al. [15]	✓	✓	✓	×	×	✓	✓	✓	×	×	×
Liu et a. [16]	×	×	×	×	×	✓	✓	✓	✓	×	×
Proposed Method	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

3.2 Capacity Analysis

Hiding Capacity is one of the significant parameters that measure the watermarking algorithm's strength. The capacity specifies the maximum number of bits called a secret message can be stored in the original text. The analysis of the existing technique summarized that an efficient system is required that maximize the hiding capacity, without affecting the original content of the text and conflicting other parameters. Equation (2) can be used to measure the hiding capacity of the proposed system.

$$HC = \frac{\text{Secret } information\ (bits)}{Size\ of\ \text{cov}er\ file(Kb)} \qquad (2)$$

In this experiment, we select 50 different text documents with different sizes. When we compared the proposed algorithm with [17] and [18], it improves the hiding capacity dramatically, as shown in Fig. 3.

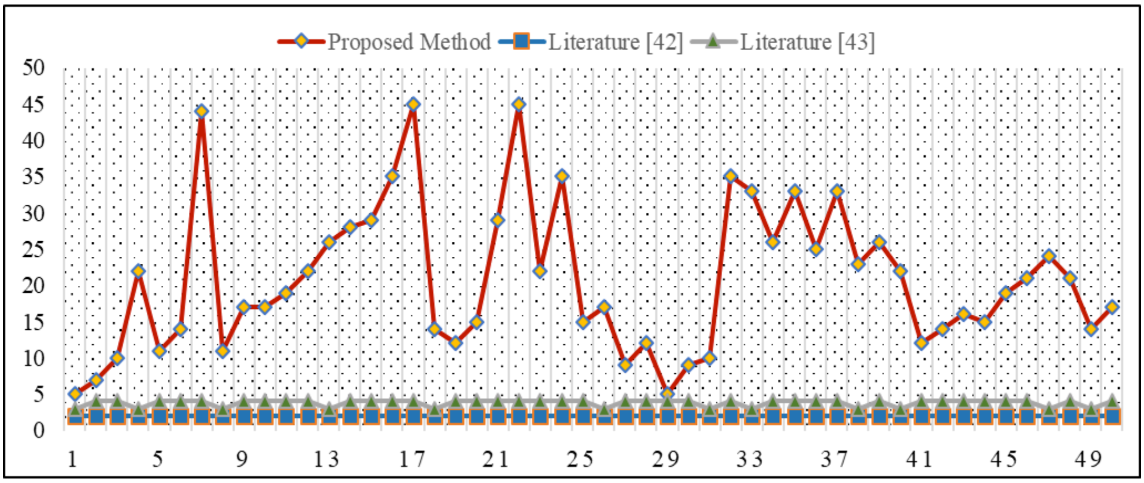

Fig. 3. The comparison of capacity analysis

3.3 Imperceptibility Test

Imperceptibility defines as the watermark information that will not alter the original content and cannot be seen through human eyes. Only the authorized persons can extract the watermark through special processing or dedicated circuits. We use 15 different strings to measure the imperceptibility, the former technique differs from 0.83 to 0.97, but the average similarity of the proposed system is 1 as shown in Fig. 4. As mentioned above, we use the custom components to embed the watermark, so the watermark does not affect the original content that’s why our scheme has 100% results on imperceptibility.

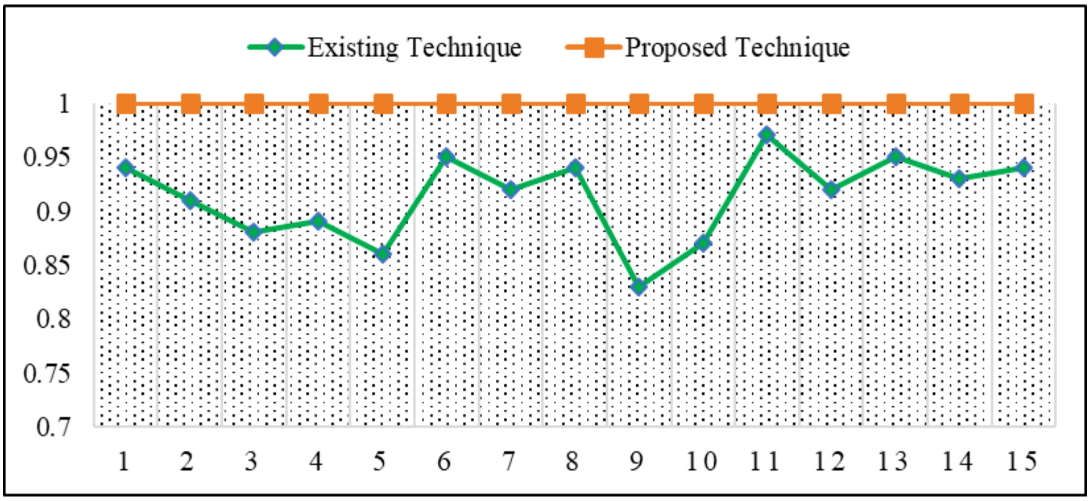

Fig. 4. The comparison of imperceptibility test with the previous technique [19]

As demonstrated in the experimental results, our proposed technique achieves excellent results against robustness, capacity and imperceptibility. The proposed method is robust against all formatting attacks and more secure compared to the previous techniques. Our system can be applied for text documents authentication and copyright protection. It can also protect text documents from illegal use.

4. Conclusion

In this investigation, we proposed a content free watermarking technique that is based on Microsoft word document custom components. These custom components are not referred to in the main document. Therefore, no changes were made to the content and format of the original document when we embed the watermark information. The experimental results and analysis prove that the proposed technique is robust against attacks based on content and formatting, imperceptible, and improves the capacity as compared to the previous techniques. The watermark information can be extracted with high probability after applying various formatting attacks. In the future, Microsoft Word and Excel documents, other properties will be examining for watermarking. Moreover, we also investigated the Portable Document Format (PDF) document that is the most popular document format in the world. The handwritten text, fingerprints, and manual signatures can also be taken as watermark.

References

1. Yesilyurt, M. and Y. Yalman, New approach for ensuring cloud computing security: using data

hiding methods. Sādhanā, 2016. 41(11): p. 1289-1298.

2. Khadam, U., et al., Digital Watermarking Technique for Text Document Protection Using Data Mining Analysis. IEEE Access, 2019. 7: p. 64955-64965.

3. Naz, F., et al., Watermarking as a service (WaaS) with anonymity. Multimedia Tools and Applications, 2019: p. 1-25.

4. AlKhamese, A.Y., W.R. Shabana, and I.M. Hanafy. Data Security in Cloud Computing Using Steganography: A Review. in 2019 International Conference on Innovative Trends in Computer Engineering (ITCE). 2019. IEEE.

5. Naqvi, N., et al., Multilayer partially homomorphic encryption text steganography (MLPHE-TS): a zero steganography approach. Wireless Personal Communications, 2018. 103(2): p. 1563-1585.

6. Khadim, U., et al., Information hiding in text to improve performance for word document. International Journal of Technology and Research, 2015. 3(3): p. 50.

7. Khairullah, M. A novel text steganography system using font color of the invisible characters in microsoft word documents. in 2009 Second International Conference on Computer and Electrical Engineering. 2009. IEEE.

8. Bhaya, W., A.M. Rahma, and D. Al-Nasrawi, Text steganography based on font type in MS-Word documents. 2013.

9. Kumar, R., et al. A high capacity email based text steganography scheme using Huffman compression. in 2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN). 2016. IEEE.

10. Khosravi, B., et al., A new method for pdf steganography in justified texts. Journal of information security and applications, 2019. 45: p. 61-70.

11. Alghamdi, N. and L. Berriche. Capacity Investigation of Markov Chain-Based Statistical Text Steganography: Arabic Language Case. in Proceedings of the 2019 Asia Pacific Information Technology Conference. 2019. ACM.

12. Long, Y., et al., Coverless Information Hiding Method Based on Web Text. IEEE Access, 2019. 7: p. 31926-31933.

13. Rizzo, S.G., F. Bertini, and D. Montesi, Fine-grain watermarking for intellectual property protection. EURASIP Journal on Information Security, 2019. 2019(1): p. 10.

14. Zhang, J., et al. Text Information Hiding Method Using the Custom Components. in International Conference on Cloud Computing and Security. 2018. Springer.

15. Castiglione, A., A. De Santis, and C. Soriente, Taking advantages of a disadvantage: Digital forensics and steganography using document metadata. Journal of Systems and Software, 2007. 80(5): p. 750-764.

16. Liu, T.-Y. and W.-H. Tsai, A new steganographic method for data hiding in microsoft word documents by a change tracking technique. IEEE Transactions on Information Forensics and Security, 2007. 2(1): p. 24-30.

17. Chen, X., et al. Coverless information hiding method based on the Chinese mathematical expression. in International Conference on Cloud Computing and Security. 2015. Springer.

18. Zhou, Z., et al. Coverless information hiding method based on multi-keywords. in International Conference on Cloud Computing and Security. 2016. Springer.

19. Wang, Z.-H., et al. Emoticon-based text steganography in chat. in 2009 Asia-Pacific Conference on Computational Intelligence and Industrial Applications (PACIIA). 2009. IEEE.