A research on PDF protection mechanisms

Linfeng Jiang

Nankai Secondary School, Chongqing, 400000, China

yitai@cqyaotai.com

Abstract. Portable Document Format, also named PDF, can integrate images, text, tables, or other elements into a single file, making it an easy-to-store and transfer document format. Due to the biggest feature of PDF being its portability, the modification of such files is very difficult and can ensure the security of the text to a certain extent. Therefore, it has been welcomed by a large number of researchers, office workers, and students. But as the demand for PDF files continues to rise, its security vulnerabilities are gradually being exposed. Consequently, a plethora of security measures for PDFs has come into view. This article primarily summarizes a portion of the overarching principles governing PDF security. Subsequently, this paper will focus on introducing the principles of steganography and cryptography in PDF security protocols while also selecting meaningful and representative research papers to summarize the advantages and disadvantages of the proposed methods.

Keywords: steganography, watermark, AES, encryption.

1. Introduction

PDF file, as one of the most commonly used file types, is widely used in work and study because of its strong compatibility, it can insert pictures, text, hyperlinks, and other contents into the file, and it can also be viewed in a very simple way on different devices. It can also be non-printable and non-replicable for some files involving copyright.

And due to the frequent use of Portable Document Format (PDF), the functionality of PDF has been improved step by step with each new generation. Now the PDF, in files includes a large number of embedded technologies, such as Pretty Good Privacy (PGP), JavaScript, Extensible Markup Language (XML), Hyper Text Mark-up Language (HTML), Simple Object Access Protocol(SOAP), Compression will Encryption, etc. At the same time, however, the security of PDF files has been expanded [1]. An attacker can simply use a browser to bypass the external password restrictions of PDF or can use Optical Character Recognition (OCR) technology to analyze PDF text so that it can be copied and pasted. In general, the addition of these technologies makes PDF security issues gradually revealed in people's eyes.

Therefore, several security measures have been applied to PDF technology, such as steganography and cryptography. Steganography can be hidden from attackers because it hides the PDF ID in a PDF document and then accepts the steganography using a key to see if the document has been hacked or tampered with. While password encryption technology can effectively prevent attackers from attacking, such as Advanced Encryption Standard (AES) and Rivest Cipher 4 (RC4) passwords, which use different hard problems to encrypt documents, making it difficult for attackers to crack the password to obtain PDF content during the document validity period.

However, in recent years, most scholars have focused on the improvement of PDF steganography but rarely mentioned the overall security principle of PDF. Therefore, this paper mainly summarizes the security principles of PDF (encryption algorithm and steganography, etc.).

In the second part, this paper will focus on the basic knowledge of PDF security principles, which can be divided into two parts: steganography and cryptography. Steganography, with its excellent part of the hidden information about PDF ID steganography in PDF, and according to the location of the hidden writing can be divided into image steganalysis, text steganography, and watermarking technology. On the other hand, password encryption is also widely used in PDF documents, and RC4 and AES are still the two most popular passwords used in PDFs. This paper focuses on the definition of steganography and cryptography and the principles used in PDF documents. In the third part, this paper collects some research papers on steganography and cryptography and points out the advantages and disadvantages of the methods mentioned in these papers according to the content of the articles.

2. Preliminaries

In this part, this paper mainly introduces the concepts of steganography and cryptography, as well as their application methods and principles in PDF documents. This part is roughly divided into steganography and cryptography two parts, steganography is subdivided into picture steganography, text steganography, and watermarking technology. The password encryption section mainly introduces the common mainstream encryption methods and encryption principles for PDF documents.

2.1. Steganography

In PDF security principles, general approaches can be divided into two broad categories, steganography, and cryptography. Among them, steganography can be divided into text steganography, picture steganography, watermarking, and so on. To understand its basic principles, it needs to be familiar with steganography, a technique widely used to protect documents of all kinds.

The word “steganography” originally comes from the Greek word for "hiding information." Unlike cryptography, which focuses on making information invisible -- often referred to as "garble" -- but still known to exist, cryptography focuses on making the information unreadable. The main goal of steganography is to enhance the undetectability of secret information, and this is done on a plain, seemingly unrelated file called a carrier or container, in which all secret information can be embedded. Virtually all file formats can use steganography, including JPG, JPEG, MP3, PDF, and even HTML pages.

Compared with the encryption algorithm, steganography after the encryption of the information itself and the information carrier has a certain confusion, so the use of steganography in the file seems to be more difficult to crack than the encryption algorithm.

The principle of steganography itself can be roughly described as follows: firstly, the information to be hidden is encrypted, encoded, and random numbers are generated, and then the processed encrypted information is added to a carrier through embedded technology in the file. Finally, the receiver extracts the secret information from the carrier and decrypts it through the key [2].

According to the different encryption algorithms for information, steganography can be roughly divided into three types: pure steganography, private key steganography, and public key steganography. Pure steganography refers to the method of no key exchange in the hidden channel constructed by steganography, while key steganography refers to the method of encrypting information by using the key algorithm in the hidden channel constructed by steganography. The two kinds of key steganography mean that the two parties transmitting secret information exchange keys before secret communication, and then the receiver will communicate with the steganography carrier. In pure steganography, there is no prior exchange of keys. Therefore, compared with pure steganography, key steganography is more likely to be hijacked or suspected by attackers. However, due to the development of various powerful public key algorithms such as Rivest-Shamir-Adleman (RSA), ElGamal, Error Checking and Correcting

(ECC), and so on, public key steganography has gradually flourished. At present, the two most popular PDF steganography are pure steganography and public key steganography

In addition, with the development of time, PDF also gradually used other kinds of steganography techniques, such as digital watermarking, text steganography, and picture steganography.

Digital watermarking can be divided into two types: emerging watermarking and hidden watermarking. The former is visible watermarking. Generally speaking, emerging watermarking usually contains the name or logo of the copyright owner. Hidden watermarking is a steganographic method of information hiding used to prevent unauthorized copying and use of files.

Text steganography is the addition of hidden informational content to text. For different carrier content, there are two categories: steganography for formatted text and concealment for unformatted text. Steganography with formatted text. Text content can be expressed in different ways by adding various grammar, so there are many steganographic methods. Examples include setting the font color to be the same as the background color for obfuscation, making the font too small to be detected by the naked eye, hiding information in a different language, hiding information in comments, and other methods. At the same time, information can be broken up and re-encoded into smaller and smaller chunks of information, or even into binary numbers, and represented as zeros and ones in various ways, such as whether there is a space at the beginning of each line, whether the spacing of each line is single or 1.5 times, the spacing of each word, the size of the font and the color can be used as ways of hiding information. Steganography of plain text, such as spacing between words, the layout of punctuation marks, initial groups of words, and pinyin, may also be large amounts of readable or unreadable text hidden in a small sentence of key information.

Picture steganography is to add hidden information to the PDF image. Information can be hidden at the head, middle, and end of the image. Hidden in the title can be explanatory information such as image information abstract, date, author, title, etc. Hidden in images, according to the unique format of the image, hidden information can be added at the end of each module without damaging the display of the original image, or special functions that are not displayed in the module can be added to achieve the effect of hiding. Meanwhile, some steganography algorithms can also be used for writing, such as LSB steganography in typical spatial steganography, which converts hidden information into binary data and writes it bit by bit at lower levels, thereby writing data without causing significant color differences [2].
Vulnerability Exploitations Using Steganography in PDF Files

2.2. Encryption algorithm

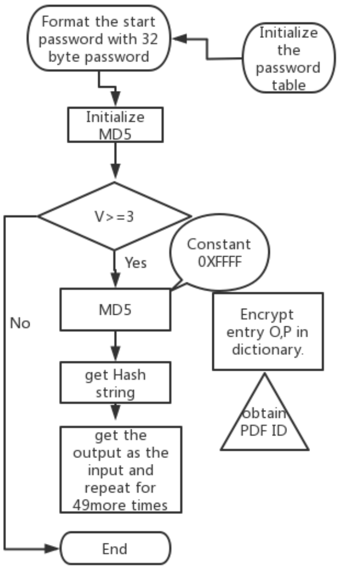
Format the start password with 32 byte password
Initialize the password table
Initialize MD5
V>=3
Constant 0XFFFF
Yes
No
MD5
Encrypt entry O,P in dictionary.
get Hash string
obtain PDF ID
get the output as the input and repeat for 49more times
End

Figure 1. Obtaining global encryption keys.

Another common method in PDF security principle is password encryption.PDF file is one of the files that users often use for transmission. It is necessary to consider whether the user and the text information can be copied or printed. Therefore, it is essential to encrypt PDFs both externally and internally.

As shown in Figure 1, earlier versions of PDF encrypt content with RC4 encryption (40-128 bits) and since PDF1.6, the AES encryption algorithm (128 bits) has been introduced with ACROBAT7.0. Therefore, the encryption algorithm in PDF is fixed and there is no other third encryption algorithm [3].

Therefore, the encryption algorithm of PDF can be divided into the generation of the encryption key and encryption dictionary, an object encryption key, and encryption using the key [4][5].

2.2.1. Generate an encryption key and dictionary.

The figure gives a brief flow path of how to obtain the encryption keys. The basic steps include the following steps:

(1) Generate entry O

First, the supple or truncated permission password contains 32 characters. If the entered password is larger than 32 bytes, only the first 32 bytes are kept. If it is less than 32 bytes, the missing bytes are filled in as follows:

$$\begin{gathered} < \mathrm{0x28, 0xbf, 0x4e, 0x5e, 0x4e, 0x75, 0x8a, 0x41,} \\ \mathrm{0x64, 0x00, 0x4e, 0x56, 0xff, 0xfa, 0x01, 0x08,} \\ \mathrm{0x2e, 0x2e, 0x00, 0xb6, 0xd0, 0x68, 0x3e, 0x80,} \\ \mathrm{0x2f, 0x0c, 0xa9, 0xfe, 0x64, 0x53, 0x69, 0x7a} > \end{gathered}$$

But if there is no permission password, use a user password instead. Then Initialize the MD5 function and enter the Message-digest algorithm 5(MD5) function for the result generated in the last step. Thirdly, repeat the previous stpes50 times consecutively, and take the output as the new input for the MD5 hash function. Then create the RC4 key using the first n bits of the HASH sequence. For version 2, n is always 5, but for version 3 or later, it depends on the value of Length in the encryption dictionary, length/8. Again, repeat the first step to get a 32-byte string from the user password. Encryption of the 32-bit string generated uses the key generated using the RC4 algorithm. And repeat 19 times: encode the output of the previous time as the input of the next time; The key is obtained by performing the XOR operation on the single byte and number of cycles of the original key generated previously. Last, the 32-byte string is the value of the encryption dictionary object entry O.

(2) Get the global encryption key

First, enter entry O of the encryption dictionary into MD5. And the P entry is treated as an unsigned 4-byte integer and the 4 bytes are entered into the MD5 function. Next, enter the first element of the ID identification array for this PDF document (that is, the first string of the ID entry in the trailer dictionary for this PDF document) into the function MD5. (See the article "PDF Document ID" for a description of PDF document ids). But if the document's metadata is not encrypted, enter 4-byte 0xFFFFFFFF into the hash function MD5. Then, End the hash. After that, input the first n bits of the previous MD5-generated HASH sequence, which is the Length/ 8-bit string in the encryption dictionary, into the new MD5 hash function, after which the output as input is done 49 times again in a row. Additionally, the encryption key is the first n bits of the HASH series, and for version 2, n is always the value of length/ in the encryption dictionary

(3) Generate entry U

First, it generates an encryption key based on the user password string. Then, it uses the encryption key generated in the preceding step to encrypt the 32-bit string generated in step 1 of Algorithm 2. Thus, the output of Step 2 is the value of entry U in the encryption dictionary object.

2.2.2. Generate the object encryption key and encrypt the object.

First to get the object number and generation number of a string object or stream object, or if the string object is a direct object, it is used to contain the object's identifier. Then the object number and the generation number are treated as binary integers, and the original n-byte long key is extended to n+5 bytes. That is, the lower 3 bytes of the object number and the lower 2 bytes of the generation number are connected to the first N-byte long encryption K key in the order of the lower byte first. (n is 5 if the Length of the key is 40, and length is divided by 8 if the value of V is greater than 1). Next, initializes the MD5 hash function, and then enter the string generated into MD5. With the first (N+5) bytes, if N+5&gt; 16 Then intercept the first 16 bytes and use the resulting hash result as the key of RC4 and AES symmetric encryption algorithms to encrypt the string or stream object.

3. Literature review

In the third part, this paper selects some representative research papers on PDF security principles and classifies them according to steganography and cryptography. Finally, the paper summarizes the motivation, the main methods, and the advantages and disadvantages of the proposed methods, as shown in Table 1.

Table 1. Analysis of related works.

Name	Advantages	Disadvantages
Justified texts	Large amount of hidden information Difficult to detect Good compatibility	Complex extraction Text format restriction
Chinese Remainder	Large amount of hidden information No establishment Good flexibility	Easy to reverse crack No Compression Data corruption in editing and processing
Watermarking	Easy to implement High accuracy Good security	Poor scalability Limited information capacity Limited targeting attack methods
RC4	Strong feasibility Easy to understand High practicality	Method limitation Required Partial knowledge of plaintext Fatal flaw of RC

3.1. Steganography

The PDF steganography method based on text alignment technology proposed in this paper [6] has the following advantages:(1) a large amount of hidden information: Compared to other PDF steganography methods, this method can embed more hidden information in PDF files. (2) Difficult to detect hidden positions: Text alignment technology is used to embed hidden information into the gaps between text lines without affecting the readability of the document, making it difficult to detect hidden positions. (3) Good compatibility: This method can be applied to all types of PDF files without the need for file conversion or encryption.

However, there are still some disadvantages of this method occurs: (1) The extraction process is relatively complex: to extract hidden information, it is necessary to determine the number of bits and content of hidden information by calculating the blank distance between each character. The extraction process is relatively complex. (2) Restricted by text formatting: As this method is based on text alignment technology, it is only applicable to PDF files that use text alignment technology. For PDF files that cannot be text aligned, this method cannot be used for steganography

PDF steganography [7] is a technique that embeds secret information into PDF files based on the Chinese Remainder Theory. Specifically, this method uses the Chinese remainder theorem to decompose embedded information and embed it into different parts of PDF files to improve the security and reliability of information hiding. This method also uses error correction codes to enhance the robustness of the information to help recover errors caused by the embedding process. The main advantages of this method include: it can hide a large amount of information without affecting the quality and size of PDF files; It has high security because the embedding process is based on mathematical algorithms and can be protected with passwords; Simultaneously robust, able to recover information from damaged PDF files.

And the advantage of the method this paper mainly talks about includes: (1) This method can hide a large amount of data because it uses the Chinese remainder theorem to segment the data and embed it into different PDF objects. (2) This method does not require the establishment of a key, so it is simple and fast to implement. (3) The steganography based on this method can embed secret information anywhere in the PDF file, so it is flexible.

However, there are still some disadvantages of this method occurs:(1) The embedded data is easily detected by attackers and may be decoded by reverse engineering because the data is not encrypted. (2) If the PDF file is compressed, this method may not function properly. (3) If a PDF file is edited or modified, the embedded data may be lost or damaged, so special care needs to be taken when handling it.

This paper [8] proposes a secure digital text watermarking algorithm for Portable Document Format (PDF). The main advantage of this algorithm is that it can protect the textual information in PDF files from being tampered with or copied, thereby preventing information leakage and theft. In addition, the algorithm also has high fault tolerance and robustness, and can effectively recognize and restore watermark information for some basic modification operations (such as rotation, scaling, cropping, etc.).

However, the implementation of this algorithm requires complex processing and analysis of PDF files, which may introduce a certain computational burden and time cost. At the same time, the cryptographic technology used in this algorithm needs to ensure the security of the key, otherwise, it may be cracked by an attacker, leading to the leakage of watermark information.

And the advantage of the method this paper mainly talks about includes:(1) Easy to implement: This algorithm adopts a simple encryption method and can be quickly implemented. (2) High accuracy: This algorithm can ensure that the embedded watermark information achieves high accuracy without affecting the quality of PDF files. (3) Good security: This algorithm uses the SHA-256 algorithm for hash value calculation and adopts a key extension mechanism, which can provide high security.

However, there are still some disadvantages of this method occurs:(1) Poor scalability: Due to its design based on PDF file structure, this algorithm may not apply to other types of document formats. (2) Limited watermark information capacity: Due to the use of hiding short text information, the embedded watermark information capacity is limited and not suitable for scenarios that require a large amount of embedded information. (3) Limited targeting attack methods: This algorithm can only prevent some simple attack methods, and may have vulnerabilities for more complex attack methods.

3.2. Encryption

This paper [9] mainly analyzes the security vulnerabilities that may exist when using the RC4 algorithm to encrypt PDF documents and proposes corresponding attack methods and defense measures. The author found that due to the unique structure of PDF documents, as well as the possible bias issues in the RC4 algorithm when generating pseudo-random number streams, attackers can recover the key stream of the entire document by obtaining a small amount of known plaintext, thereby further cracking document encryption and obtaining sensitive information.

This paper introduces two methods based on deviation attacks, one is a pure text attack (PTA), and the other is a selective plaintext attack (CPA). The author also compares the efficiency and feasibility of these two attack methods and provides some suggestions and preventive measures to mitigate such attacks, such as increasing password strength, using more secure encryption algorithms, and regularly changing keys. This paper provides an important reference and guidance for studying the security of PDF document encryption.

And the advantage of the method this paper mainly talks about includes:(1) Strong feasibility: This method can successfully crack PDF documents encrypted with RC4 in a short period. (2) Easy to understand: The method is based on known attacks, applies and improves the algorithm, and the operation is simple and easy to understand. (3) High practicality: The RC4 algorithm is widely used in various software and protocols, so this method has wide practicality.

However, there are still some disadvantages of this method occurs: (1) Limitations of the algorithm: This method is only applicable to PDF documents encrypted using the RC4 algorithm and cannot crack documents using other encryption algorithms. (2) Partial knowledge of plaintext is required: This method requires a small amount of known plaintext to successfully crack the document, which may limit its practicality. (3) Fatal flaw of the symmetric cryptographic algorithm: RC4 algorithm itself has the fatal flaw of the symmetric cryptographic algorithm, which can be cracked by key reuse attacks, so this algorithm is no longer recommended for use.

The main content of this paper [10] is to introduce a method for data encryption and decryption using Advanced Encryption Standard (AES) algorithms. The author introduces the principle and implementation process of the AES algorithm in detail and discusses how to choose an appropriate key length, as well as how to use the AES algorithm to protect sensitive data. The paper also provides an example of an encryption program based on the AES algorithm, demonstrating how to use the AES algorithm for data encryption and decryption in practical applications. The contribution of this paper is to combine the theoretical knowledge of the AES algorithm with practical operation

4. Conclusion

This paper mainly summarizes the overall security principles of partial PDF, including steganography, encryption, and other principles, and analyzes some PDF security methods involved in the paper, pointing out their advantages and disadvantages. Specifically, this paper discusses text steganography, image steganography, and watermark steganography to hide information in PDF and PDF ID to enhance the security of PDF, and then this paper also discusses the encryption principle of RS4 and AES ciphers for PDF in PDF1.6 system. Then, the paper on text steganography, Chinese remainder theorem image steganography, and watermark technology, as well as RC4 encryption algorithm, is cited respectively, and the advantages and disadvantages of these methods are analyzed. With the continuous development of technology, The PDF security direction generally further improves and optimizes the processing of encrypted data, and the decryption method of the receiver will be as simple as possible without reducing its security.

References

[1] Jens Müller, Fabian Ising, Vladislav Mladenov, Christian Mainka, Sebastian Schinzel and Jörg Schwenk 2019 *CCS '19* Practical Decryption exFiltration: Breaking PDF Encryption pp 15-29

[2] Istteffanny Araujo and Hassan Kazemian 2020 *International Journal of Computer Networks And Applications* Vulnerability Exploitations Using Steganography in PDF Files vol 7

[3] Adobe PDF Reference 1.6

[4] Morris J. Dworkin, et al. 2001 Advanced Encryption Standard(AES)

[5] Poonam Jindal and Brahmjit Singh 2015 RC4 Encryption-A Literature Survey vol 46 pp 697-705

[6] Behrooz Khosravi, et al. 2019 *Journal of Information Security and Applications* A new method for pdf steganography in justified texts vol 45 pp 61-70

[7] Rene Ndoundam, Stephane Gael and Raymond Ekodeck 2016 *Journal of Information Security and Applications* PDF steganography based on Chinese Remainder Theorem vol 29 pp 1-15

[8] Umair Khadim, Munwar Iqbal and Muhammad Awais Azam 2022 *Mehran University Research Journal of Engineering and Technology* A Secure Digital Text Watermarking Algorithm for Portable Document Format (PDF) vol 41 pp 100-110

[9] Cryptanalysis of RC4 Encryption in PDF Documents

[10] ko Muhammad Abdullah 2017 *Algorithm to Encrypt and Decrypt Data* Advanced Encryption Standard (AES)