Malware and fileformats

Maximilian Heim

October 23, 2022

Contents

1 Abstract

One of the biggest problems in cybersecurity these days is malware. In this paper we will discuss which kinds of malware can be spread by which file formats, what action the user has to perform to execute the malware and how to configure the corresponding programs to prevent infection.

2 Introduction

The word malware, a portmanteau of the words malicious and software is a special type of computer software which, as the name implies aims to produce damage. There are several ways in which malware can be distributed, in this paper we will discuss the distribution via files and which role the different file formats have in the distribution of malware. We will analyse which common methods exist to implement malware into the different file formats, which file formats can carry what kind of software, what actions are necessary to execute the malware and which configurations can prevent the infection of the system.

3 Definition of terms

3.1 Malware

The motivations for the distribution of malware can be diverse. To name a few there is greed, curiosity, spying, revenge and conspiracy. [1] When talking about malware it is important to discuss the how does the malware spread itself and what does it do. Malware itself can be divided into different types, depending on their method of spreading, here are listed some of the most common methods.

Type	Spreads by
Virus	Oldest kind of malware, they spread themselves by infecting programs, documents and drives
Worms	Similar method to that of viruses but spreads over networks like the internet and then tries to infect systems
Trojan	Is a combination of a seemingly useful program and a hidden malicious part, it doesn't spread itself, it rather gets distributed by the people that download these programs

Additionally, there are different categories into which malware can be divided into, depending on what they actually will do.

Type	Causes
Backdoor	They will grant third parties access to a system by using, as the name implies a backdoor, this kind of malware often gets used for spamming mails and denial of service attacks
Spyware	Collects information about the actions a user performs and sends them to a 3rd party
Adware	Causes to display unwanted advertisements to the user
Scareware	Tries to unsettle the user and trick him into downloading malware, pay for unnecessary software. This could include fake messages which pretend that the system has been infected with malware and that one have to pay for a software to remove it
Ransomware	Blocks the system of the user by encrypting the drive(s) and demands payment via payment systems like Bitcoin, Paysafecard, uPay and other payment systems to unlock the system again
Keylogger	Logs the input from the keyboard to fetch usernames and passwords for accounts of the user and sends them to a third party
Mineware	Causes the system to start mining crypocurrency in the background and transfering the money to a third party

3.2 File formats

A file format refers to the format of files in computers. The format of a file determines in which way information is encoded to store it within the file and later how to decode it for further use. Because of that file formats could be refered to as the syntax in which data is stored within files, creating an uniform way for encoding/decoding data. For programs it is important to know in which way they have to decode the files, there are several ways in which the information which file format is present can be stored.

Method	Determines by
File extension	Interpretation of the string of characters which follows the files name, seperated by a full stop
External metadata	Storing information about the file format outside of the file at a different place
Internal metadata	Storing information about the files format within the file itself, either as a so called magic number directly at the beginning of the files content which refers to a specific format or in the files header which is an area within the file that contains metadata

The format of a file has to be distinguished from the file extension. The file extension is the suffix of a files name, seperated from the name by a full stop. The file extension usually corresponds to a specific file format. Different extensions can refer to the same format though, as example .jpeg and .jpg could be named, they refer to the same format but in earlier versions of Microsoft Windows the length of the file extension was limited to maximum 3 characters so jpg was used.

4 Results

4.1 Microsoft Word documents

4.1.1 Introduction

Microsoft Word is a text editing software developed and distributed by Microsoft. It is the most popular software of its kind and offers a big variety of features. Due to the amount of features it provides it also has many weaknesses which can be exploited by attackers. The main vectors over which these malicious Microsoft Word files get distributed are e-mails (Plain spam, phishing and spear phishing)[2] and downloads from the internet.

4.1.2 Current situation

To this day malware distributed via Word documents is still a problem and there isn't any improvement visible, the tendency is that it gets even worse.

4.1.3 Macros

Introduction The file formats supported by Microsoft Word are plenty - but the most important formats these days are docx and docm, the docm variant can - in contrast to docx contain macro scripts. Microsoft Word documents, with the file format .docm are a common way to spread malware. The aspect that makes these .docm files able to infect systems with malware is a function of text editing softwares called macros. Macros in text editing softwares are small scripts which allow the user to automate repetitive tasks like adjusting several settings with just one click or keyboard shortcut. But they aren't just a handy feature for users to make their life easier, they can be used to infect the system with serious malware.

Which malware can be contained? As these macro viruses are viruses they will infect other .docm files on the computer and this way they can have a high reproduction rate. The damage they cause depends on the methods and or exploits that are being used while creating the macro virus. It can range from simple twisting/removing/adding text in the document, to deleting files on the drives to downloading malware from the internet and infecting the system with said malware.

Which actions are necessary to execute the malware? For the execution of a macro virus all the user has to do is opening the .docm document with Microsoft Office - Word and enable the macros from this file.[2] With a default installation a popup will pop up and tell the user that there are macros which can be enabled, the macros will then be executed automatically and infect the system. It's important to note that different office software manufacturers use different macro scripting languages, hence macro viruses written for Microsoft Word only work in Word, not in other office text programs like Openoffice Writer or Libreoffice writer.

Is it possible to configure the software to prevent infection? It is possible to completely deactivate macros in Microsoft Word, this will reduce the risk of infection to near zero, but the user will lose functionality as the macros are still a great function in Microsoft Word. The reason why the risk isn't completely gone is because there could still be exploits which allow execution of macros, even with macros deactivated in the settings. It is also important to note that the Microsoft Word software and Microsoft Windows itself should always be updated to the newest version if possible, these often fix exploits which can be exploited by the macro function and reduce the risk of infection.

4.1.4 Exploits

Introduction Microsoft Word just like any other software has exploits - weaknesses/bugs which can be exploited for personal motives. These exploits get discovered from time to time[3] and Microsoft tries to fix them, but its in the nature of a software to be imperfect so it will probably never be 100 percent exploit free. Attackers will try to use these exploits for their own good, or just to cause destruction. There have been several major exploits in Microsoft Word and the tendency is that the attacks via exploits gain popularity since 2017, according to a report the amount of attacks which exploit weaknesses of Microsoft Word increased from roughly 2400 in the second half of 2017 to 20600 in the second half of 2018.[4] Additionally, in comparison to macro based malware the exploit based malware is way harder to identify, they aren't only hidden in files with .docm formats, they can also be hidden in the trustworthy looking .docx files, which could even lead a professional to think opening this document is safe.

Which malware can be contained? As exploits come in such a variety they can be used to infect computers with any kind of malware, which makes them very dangerous.

Which actions are necessary to execute the malware? For this kind of Microsoft Word malware, it is often sufficient to just open the file and the malicious code will be executed automatically.

Is it possible to configure the software to prevent infection? For this kind of malware it is very hard to guarantee safeness, the user can't just adjust a setting like for macros so they aren't executed on program start, so exploit based malware is uncontrollable and relies on Microsoft to patch them. The user should make sure to always keep their Microsoft Word installation up to date.

4.2 Microsoft Excel Tables

4.2.1 Introduction

Microsoft Excel is a spreadsheet software developed and distributed by Microsoft. It is the most popular software of its kind and offers a big variety of features. Due to the amount of features it also has many weaknesses which can be exploited by attackers.

4.2.2 Current situation

As Excel just like Word features macros and has many exploits which allow code execution it is a relatively common source of malware infections.

4.2.3 Macros

Introduction Like Microsoft Word Microsoft Excel supports many different file formats, but the most commonly used ones are xlsx and xlsm, where the m variant supports macros. Similar to the macro viruses for Microsoft Word documents described above, there are also macro viruses for Microsoft Excel, the file format is analogously to .docm .xlsm. As Excel uses the same macro scripting language as Word (Visual Basic for Applications), they can be used for exactly the same attacks and get distributed the same way, because of that we wont repeat the things named above and just refer to the subsubsection 4.1.1

4.3 PDF

4.3.1 Introduction

PDF, short for Portable Document Format is a file format which was developed to be a platform independent format for documents. It is widely used as it provides wide functionality. But with wide functionality there usually are many weaknesses which can be exploited by attackers for their own purposes. Because of that the .pdf is a common root of malware infections. It is also important to note that the attacks depend which pdf reader is used, each reader has different vulnerabilities, some more, some less. This table represents the seriousness of the vulnerabilities in pdf readers - only the Adobe Acrobat Reader, which is one of the most popular pdf readers has had 42 known and exploits in the year 2018.[5]

4.3.2 Current situation

Malware distributed by pdf files is still an active problem in it security. Those aren't just single cases from time to time, instead there are constantly cases which use new exploits in the pdf reader softwares which make this possible.

4.3.3 Java Script

Introduction PDF Files are able to contain JavaScript code which is a useful feature, but also a weakness which can be an attack vector for attackers. One of the most used methods is exploiting a weakness of the used reader/JavaScript interpreter and use heap spraying[6] - heap spraying will fill a processes heap with a certain sequence of bytes, in this case the JavaScript heap with a predetermined shell code script, therefore any given adress in the heap will contain a shell code and then the attacker just has to use an exploit to execute the code at that adress. This way the shellcode will be executed, which can contain any kind of code, it could download malware from the internet or open a file contained in the .pdf file. This is an example of a heap spray script written in JavaScript, thanks to danfujita from GitHub[7] which published this script under the MIT License (which is still the license for this code snippet!):

```
var shellcode = unescape('%uc031%u6850%u696e%u6c65...');
var nopslide = unescape('%u9090%u9090');
var headersize = 20;
var block = headersize+shellcode.length;
while(nopslide.length < block)nopslide += nopslide;
var fillblock = nopslide.substring(0,block);
var finalblock = nopslide.substring(0,nopslide.length-block);
while (finalblock.length + block < 0x40000)finalblock =
finalblock + finalblock + fillblock;
var memory = new Array();
for(i = 0; i<1000; i++){memory[i] = finalblock + shellcode}
```

Another technique used to exploit JavaScript is using XFA forms, there have been known attacks which chose a quite complicated route to attack, in one example the attackers did hide a .png file in the .pdf file, the .png file contained a XFA form, the XFA form in turn contained obfuscated JavaScript code.[8]

Which malware can be contained? The shellcode can be directly contained in the JavaScript file and executed by using an exploit and the heap spraying technique combined, the contained shell script then can download a file from anywhere in the internet, it can pretty much install any kind of malicious software making it really dangerous.

Which actions are necessary to execute the malware? It depends on what exploits are being used to hide the malicious part inside. The user can deactivate JavaScript in their .pdf reader which covers most of the vulnerabilities but isn't a guarantee to completely prevent malicious code execution. Most importantly the user should always keep their .pdf reader up to date.

Is it possible to configure the software to prevent infection? Most importantly the user should deactivate the JavaScript function within their pdf reader as most of the attacks via pdf files use some sort of JavaScript to execute the payload. As always it is important to keep the reader up to date to fix vulnerabilities.

4.4 JPEG

4.4.1 Introduction

The jpeg format is a file format used to store compressed pictures to save memory. The encoding of this data follows one of the several methods which are defined by the standard ISO/IEC 10918-1. There are lossless and lossy compression methods available to fit several needs, although the lossy methods are the dominating methods.

4.4.2 Current situation

The current situation of malware infections caused by jpeg is relatively

4.4.3 Buffer overrun/overflow

Introduction A buffer overflow is an anomaly in the memory of a computer, buffers are parts of the memory that temporarily hold data. If a buffer is now about to get filled with data, and the transferred data is bigger as the buffers size, the data could overwrite parts of the memory which hold executable code, this way malicious code could sit in the memory where other program code should sit, if using exploits, or by pure luck the payload next to the buffer gets executed it could download malware from the internet, compromising the system. There have been plenty known buffer overflow vulnerabilities for photo displaying software in the past few years which allowed malicious code execution. One example of an exploit that uses this technique is CVE-2004-0200 which affected many Microsoft products like Windows XP, Word, Excel, Powerpoint, Outlook and the .net Framework. Another example is the OpenJPEG codec which has had several vulnerabilities over the last few years.[9] But those are only few of the many exploits which can be used to execute malware hidden in jpeg files and new exploits are getting discovered constantly.

Which malware can be contained? As these exploits can allow arbitrary code execution any malware can be hidden inside this format which makes it a very dangerous source of malware.

Which actions are necessary to execute the malware? The only action necessary to execute the malware is opening the file with a photo viewer which is vulnerable to the exploit used from the attacker.

Is it possible to configure the software to prevent infection? It isn't possible to configure the software to prevent this kind of infection as it isn't part of any feature rather the the whole process of displaying the picture itself, hence the only option to prevent this kind of infection is to always keep the used photo viewer up to date and install the newest operating system updates.

4.4.4 Stegomalware

Introduction This method is a bit different from the others as it requires an actual software to later execute the hidden code, but it is still mentionable as it is a method to hide malware within seemingly clean software. Stegomalware or stegoware is a portmanteau of the words steganography and malware. Steganography is the practice of hiding information within other information while trying to make the changes made to hide the information not visible. For this, several different file formats can be used like text, picture, audio and video file formats. As this part is about jpeg files we will discuss hiding the information within jpeg files. Hiding informations within digital pictures isn't a very complex thing, one just has to change the least significant bits of the color encoding across the picture, which doesn't cause any visible change for humans because the color shift is so minor it isn't observable for the human eye. Later the informations can be encoded again by an algorithm, which in fact can hold any kind of data. Applying this idea to software, the malicious part of a malware could be hidden in pictures, then once the malware which is camouflaged as useful software is executed on the system all the software has to do is apply the algorithm onto the file and then execute the code, compromising the system.[10]

Which malware can be contained? As steganography allows hiding any kind of data within images it could hold any kind of malicious code.

Which actions are necessary to execute the malware? Executing the seemingly clean software is the only action necessary to compromise the system with stegomalware. If the attacker used a proper algorithm to hide the payload it is almost impossible for anti malware programs to detect this kind of software.

Is it possible to configure the software to prevent this kind of infection? It isn't possible to configure the software in order to prevent infection as the software itself is the malware. The only protection against this kind of malware is avoiding the download of software from untrusted parties.

4.5 MP3

4.5.1 Introduction

The mp3 format is a file format commonly used to compress audio files to save storage. Converting an audio file to the mp3 format will result in the removal of frequencies which humans aren't capable of hearing, as this it is a psychoacoustic method of audio compression and will yield massive decrease in file size. It is a lossy audio format for the reason stated above.

4.5.2 Current situation

Malware distributed by mp3 files isn't a real problem anymore as most known weaknesses/exploits have been fixed in the past. That doesn't mean it isn't possible, but that there are other, way more convenient ways for malware to be distributed.

4.6 MP4

4.6.1 Introduction

The mp4 format is one of the most common file formats to store videos. It is a container format which means it can contains different files within itself. The different files contained can be video, audio, pictures and text for subtitles.

4.6.2 Current situation

Malware distributed by mp4 files has been a thing from time to time. Such as in the year 2019 there has been an exploit in WhatsApp's code that allowed arbitrary code execution via mp4 files sent over the WhatsApp messenger, allowing to execute any kind of payload, which luckily got fixed relatively fast. Since then there haven't been any major problems with mp4

4.6.3 HTML E-Mail

4.6.4 Introduction

E-Mail is a method for exchanging messages between people over the internet. Over the past years it slowly but steadily replaces the paper mail because it is inexpensive, fast and reliable. But as it is a system which works via the internet certain risks come with it. In the early days of E-Mail the content of them was pure text. At some point HTML was started to become relevant in E-Mails to allow formatting like bold, italic, underlined and colored text, changed fonts implementing pictures, links and much more. But this did not only created advantages for E-Mails, it increased the attack surface for cyber criminals.

4.6.5 Current situation

4.6.6 Table representation of the results

In this section we compress the results gathered into a table for a fast and easy overview.

Format	Current situation	Possible malware	Actions necessary	Prevention
Word	Acute problem	Any malware can either be contained or downloaded from the internet	For the macro based malware the user usually has to activate said macros, but other methods using exploits could achieve code execution by just opening the document	Disabling macros and keeping the software up to date
Excel				
PDF				
JPEG				
MP3				
MP4				
HTML E-Mail				

5 References

References

[1] *Bedrohungen*, pages 49–83. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[2] N. Nissim, A. Cohen, and Y. Elovici. Aldocx: Detection of unknown malicious microsoft office documents using designated active learning methods based on new structural feature extraction methodology. *IEEE Transactions on Information Forensics and Security*, 12(3):631–646, March 2017.

[3] Microsoft " word : Security vulnerabilities.

[4] Wren Balangcod. Distributing malware, one "word" at a time techblog, Feb 2019.

[5] Adobe " acrobat reader : Vulnerability statistics.

[6] Didier Stevens. Malicious pdf documents explained.

[7] Danfujita. danfujita/heapspray, May 2016.

[8] Complex – pdf hides malware inside xfa which is inside png – not an image.

[9] Openjpeg : Security vulnerabilities.

[10] Guillermo Suarez-Tangil, Juan Tapiador, and Pedro Peris-Lopez. Stegomalware: Playing hide and seek with malicious components in smartphone apps. 12 2014.