第 3 章 文件与文件格式

恶意软件分析人员每天要分析数百个文件，分析人员必须要了解各种文件格式以及识别文件格式的方法。本章将会介绍如何识别各种文件以及扩展名和文件格式。

十六进制可视化文件

计算机最终能够理解的都是二进制。二进制转换为比特（Bit），最终由 0 或者 1 表示。实际上，操作系统上的每个文件都是二进制的，认为只有可执行文件才是二进制文件是一个典型的误解。各种文件和数据，可执行文件、文本文件、HTML 文件、应用程序、PDF 文件、Word 文档文件、PowerPoint 幻灯片文件、音频文件、视频文件、游戏文件或者以文件形式存储在计算机中的数据，其实都是二进制文件。在打开文件时，每个文件都会根据文件的扩展名或数据格式的不同，以不同的方式运行呈现给用户。文件的每个字节都可以以十六进制形式进行可视化，如图 3- 2 所示。

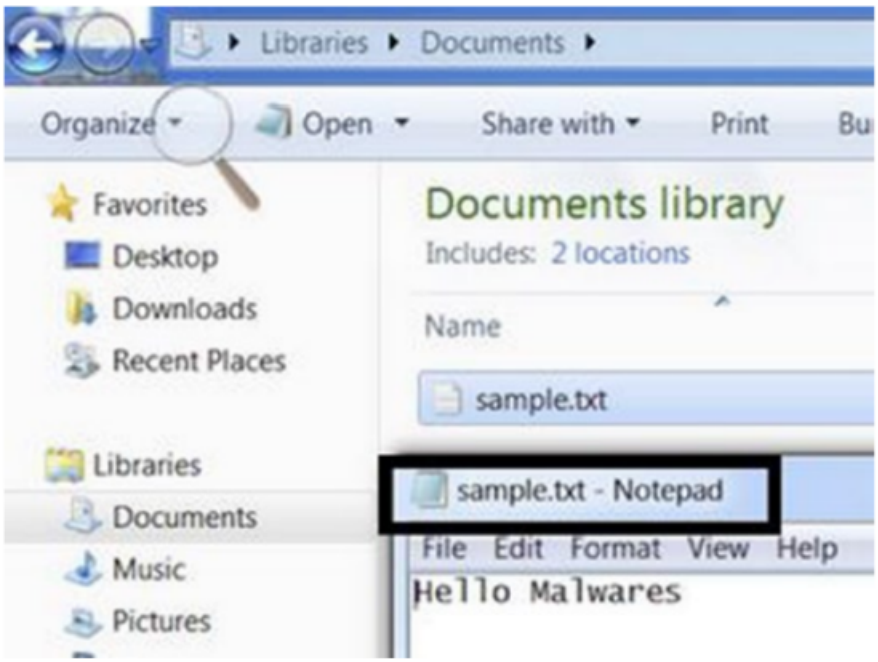

图 3- 1 在 Windows 上使用 Notepad 创建文本文件

使用记事本（Notepad）创建文本文件并输入一些文本，如图 3- 1 所示。尝试使用十六进制编辑器打开新创建的文件，在 Windows 上可以使用 Notepad++的十六进制视图，如图 3- 2 所示。当然，也可以使用其他十六进制编辑器。

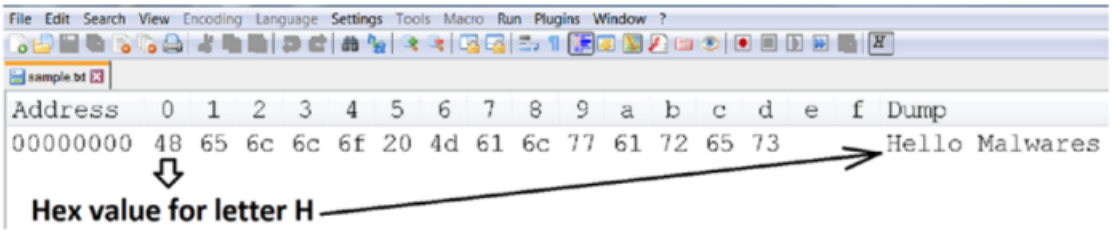

图 3- 2 使用 Notepad++的十六进制编辑插件查看十六进制视图

图 3- 2 中间部分是以十六进制显示文件的字节，如果是可见字符则在右侧会显示与十六进制相对应的 ASCII 可打印字符。中间部分的任何字符都是十六进制字符，十六进制字符的范围为 0-9 和 A-F。为什么不是 0 或 1 呢？十六进制是二进制的另一种表示方法，就像十进制表示法一样。在图 3- 2 中，字母 H 的十六进制为 0x48、十进制为 72、二进制为 0100 1000。十六进制编辑器以十六进制显示二进制内容，主要是为了便于分析人员阅读。

现在，大多数程序员已经不需要处理十六进制或者二进制的文件。但由于恶意软件分析人员还需要深入分析恶意软件样本文件，因此分析人员不能放弃查看原生二进制形式文件的能力。通常分析都以十六进制可视化的形式开展，作为恶意软件分析人员、逆向分析工程师和检测工程师，都必须熟悉十六进制。

哈希：标识文件唯一性的指纹

世间这些数不尽的文件需要一种唯一标识的方法，而文件名并不能作为唯一标识符，两台不同的计算机甚至同一台计算机上的两个文件都可以具有相同的文件名。哈希函数闪亮登场，在恶意软件分析领域中经常被用于唯一标识恶意软件样本。

任何数据都可以通过哈希函数为本身生成可以作为唯一标识的哈希值，几个字节甚至是全部文件内容都可以计算哈希值。文件哈希的工作原理是获取文件内容并通过哈希算法为这些文

件内容生成一个唯一的字符串，如图 3- 3 所示。

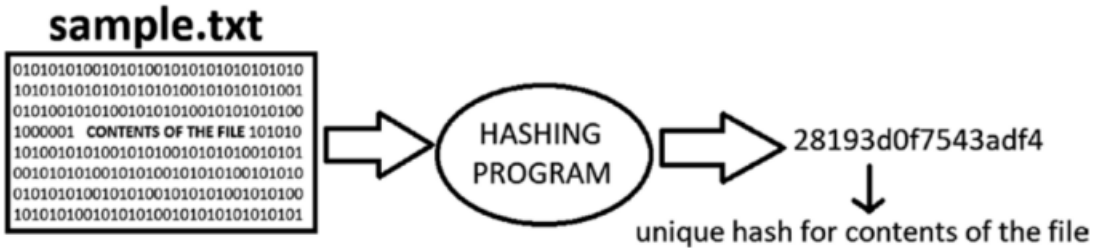

图 3- 3 生成文件哈希的说明

关于文件哈希的常见误解就是更改文件名会使哈希发生变化，但其实哈希仅与文件内容有关。文件名并不是文件内容的一部分，并不会在哈希函数计算生成哈希值的过程中。另一个需要注意的是，文件内容中即使改变了一个字节也会导致哈希发生变化，如图 3- 4 所示。

图 3- 4 修改文件单字节会影响哈希值变化

恶意样本的哈希也是恶意软件分析领域用于识别和引用样本文件的值。后续章节中也会介绍到，一旦发现了恶意软件样本文件，通常研究人员就会生成文件的哈希值并在互联网上进行检索。又或者，如果只有恶意软件样本文件的哈希值，可以根据哈希获取更多信息进行进一步分析。

在恶意软件分析领域主要使用三种哈希函数（MD5、SHA1 和 SHA256），每一种哈希函数对应的哈希值都可以由对应的哈希函数计算工具生成。代码 3- 1 显示了同一的对应 MD5、SHA1 和 SHA256 的哈希值。

代码 3- 1 根据相同文件计算生成的 MD5、SHA1 与 SHA256 哈希值

```
MD5 - 28193d0f7543adf4197ad7c56a2d430c
SHA1 - f34cda04b162d02253e7d84efd399f325f400603
SHA256 - 50e4975c41234e8f71d118afbe07a94e8f85566fce63a4e383f1d5b
a16178259
```

在 Windows 系统上查看文件的哈希，可以使用 HashMyFiles 程序，如图 3- 5 所示。为位于 C:\Windows\notepad.exe 的记事本程序计算哈希，该文件是 Windows 系统上打开文本文件的处理程序。此外，也可以使用 QuickHash 等工具进行哈希的计算。

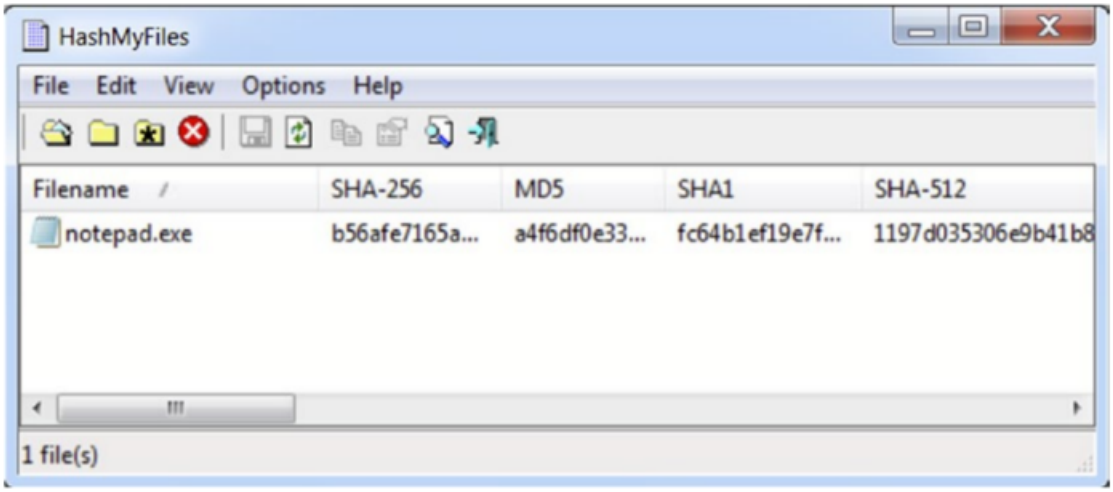

图 3- 5 HashMyFiles 能够生成 MD5、SHA1 与 SHA256 等其他哈希值

在 Windows 上也可以使用 md5deep、sha1deep、sha256deep 程序通过命令为文件生成对应的 MD5、SHA1 与 SHA256 哈希值，如代码 3- 2 所示。

代码 3- 2 通过命令行使用 md5deep、sha1deep 与 sha256deep

```
C:\>md5deep C:\Windows\notepad.exe
a4f6df0e33e644e802c8798ed94d80ea C:\Windows\notepad.exe
C:\>sha1deep C:\Windows\notepad.exe
fc64b1ef19e7f35642b2a2ea5f5d9f4246866243 C:\Windows\notepad.exe
C:\>sha256deep C:\Windows\notepad.exe
b56afe7165ad341a749d2d3bd925d879728a1fe4a4df206145c1a69aa233f68b
C:\Windows\notepad.exe
```

识别文件

识别文件主要有两种方法：文件扩展名与文件格式。本节中将会挨个介绍这些文件识别技术，并且会介绍部分攻击者用来欺骗用户运行恶意软件的技术。

文件扩展名

操作系统识别文件的主要方式是使用文件的扩展名。在 Windows 上为标识文件类型，文件

扩展名作为文件名的后缀，通常是句点字符.后紧跟着的三个字母，例如.txt、.exe 与.pdf。文件扩展名最短可以为一个字符，也可以为数十个字符。默认情况下，Windows 文件资源管理器不会显示文件扩展名，如图 3- 6 所示。但是，用户可以手动配置显示系统上所有文件的扩展名，如第 2 章中所述。

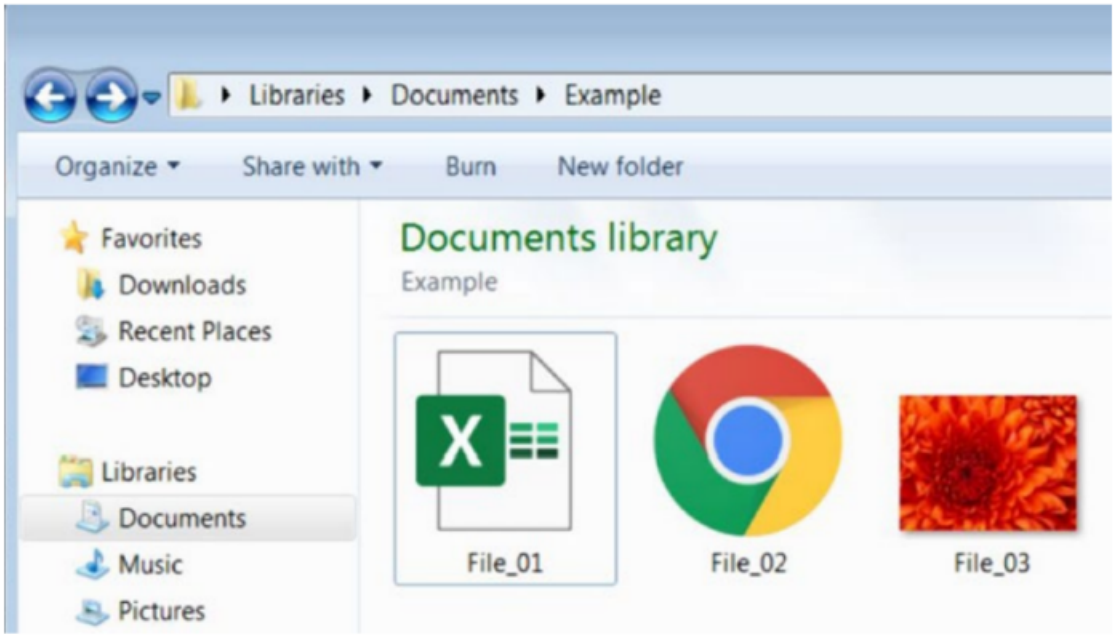

图 3- 6 文件扩展名默认不显示

禁用文件扩展名隐藏后，用户就能够看到文件扩展名，如图 3- 7 所示。

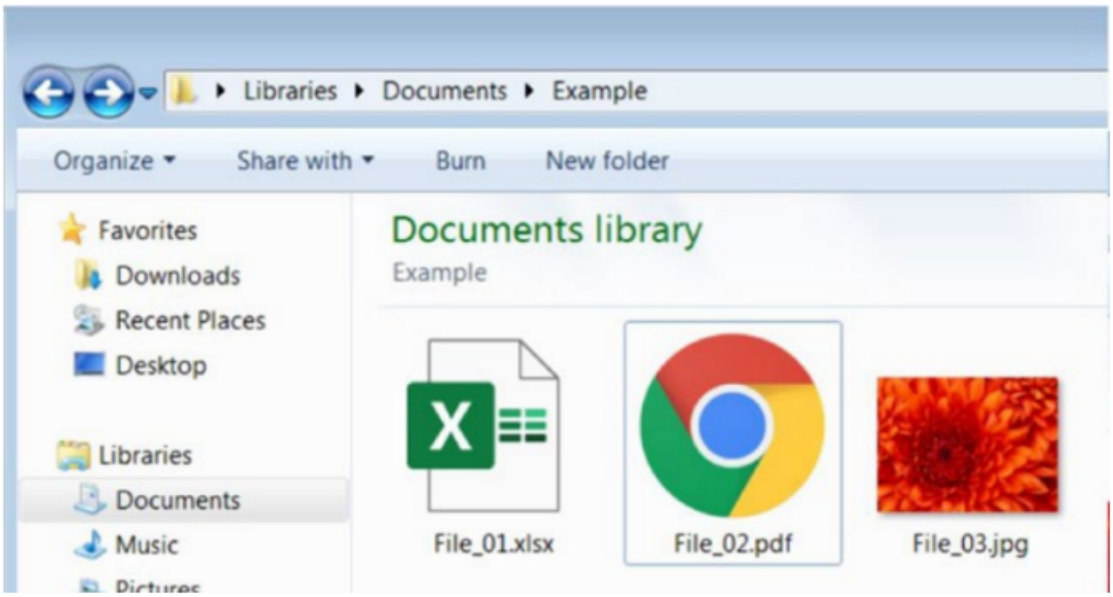

图 3- 7 禁用文件扩展名隐藏后，文件扩展名可见

表 3- 1 中列出了部分常见的文件扩展名和每个扩展名对应的文件类型。

表 3- 1 部分常见的文件扩展名和每个扩展名对应的文件类型

文件扩展名	简介
.pdf	Adobe 便携文档格式文件
.exe	微软可执行文件
.xslx	微软 Office Excel Open XML 格式文档
.pptx	微软 Office PowerPoint Open XML 格式文档
.docx	微软 Office Word Open XML 格式文档
.zip	ZIP 压缩文件
.dll	动态链接库文件
.7z	7-Zip 压缩文件
.dat	数据文件
.xml	XML 文件
.jar	Java 归档文件
.bat	Windows 批处理文件
.msi	Windows 安装包文件

文件关联：操作系统如何使用文件扩展名

操作系统通过文件关联将应用程序与特定的文件类型或者扩展名相绑定，通常应用程序都依赖文件扩展名建立关联关系。

在从未安装Microsoft Office的操作系统上，尝试打开任意Microsoft PowerPoint文件(其扩展名为.ppt或.pptx)。由于缺少与 Microsoft PowerPoint 文件或者与.pptx 文件扩展名相关联的软件，操作系统都会提示错误：“无法打开该文件”，如图 3- 8 所示。系统中如果没有与.pptx 文件扩展名相关联的软件，在用户尝试打开对应的文件时，Windows 操作

系统由于不知道如何处理这些文件就会弹出错误提示。

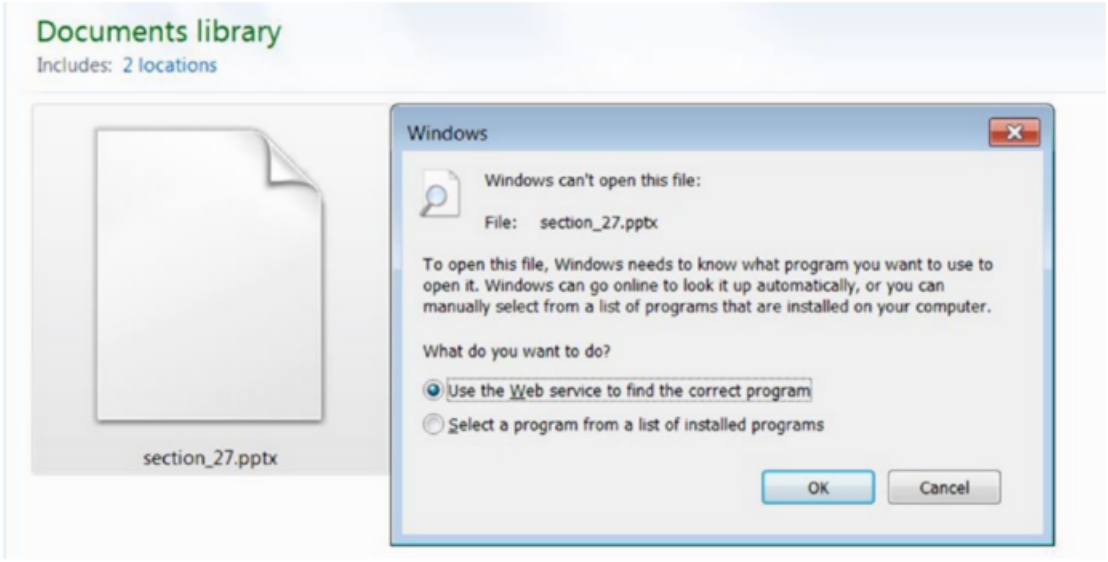

图 3- 8 没有安装扩展名关联的特定应用程序

在同一台计算机上尝试打开.jpeg 或者.png 图片文件，如图 3- 9 所示可以成功打开。这是由于 Windows 系统默认安装了与.jpeg 和.png 文件扩展名相关联的图像查看程序。

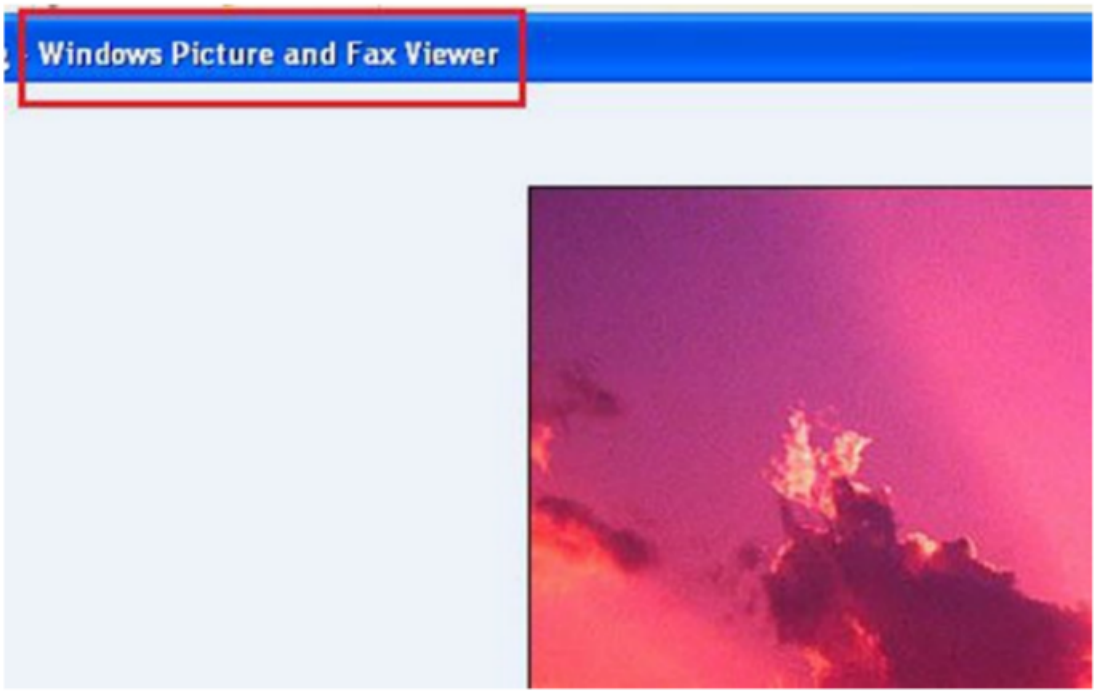

图 3- 9 Windows 使用图片查看程序打开扩展名为.jpeg 的图片文件

为什么禁用扩展隐藏?

在分析恶意软件时，根据文件扩展名就可以快速了解正在分析的文件类型。此外，恶意样本

在运行时可能会创建多个文件，根据文件扩展名可以快速定位创建的文件类型。如后续内容所述，攻击者还会使用扩展名伪装和缩略图伪装技术来欺骗用户点击恶意软件，了解文件的正确扩展名可以帮助分析恶意软件。

扩展名伪装

一些已知的恶意软件会利用扩展名隐藏来欺骗用户点击执行，从而感染系统。示例样本库中的样本文件 Sample-3-1，如图 3- 10 所示。如左侧所示，样本文件乍一看似乎是.pdf 文件。但实际上该文件为可执行文件，其真实文件扩展名.exe 被隐藏了。攻击者利用实际扩展名未被显示的这一点来对文件重命名，通过在文件名中间增加.pdf 来诱使用户认为该文件是可以安全打开的 PDF 文档文件。如右侧所示，在 Windows 中禁用了扩展名隐藏后就可以看到实际的文件扩展名.exe。

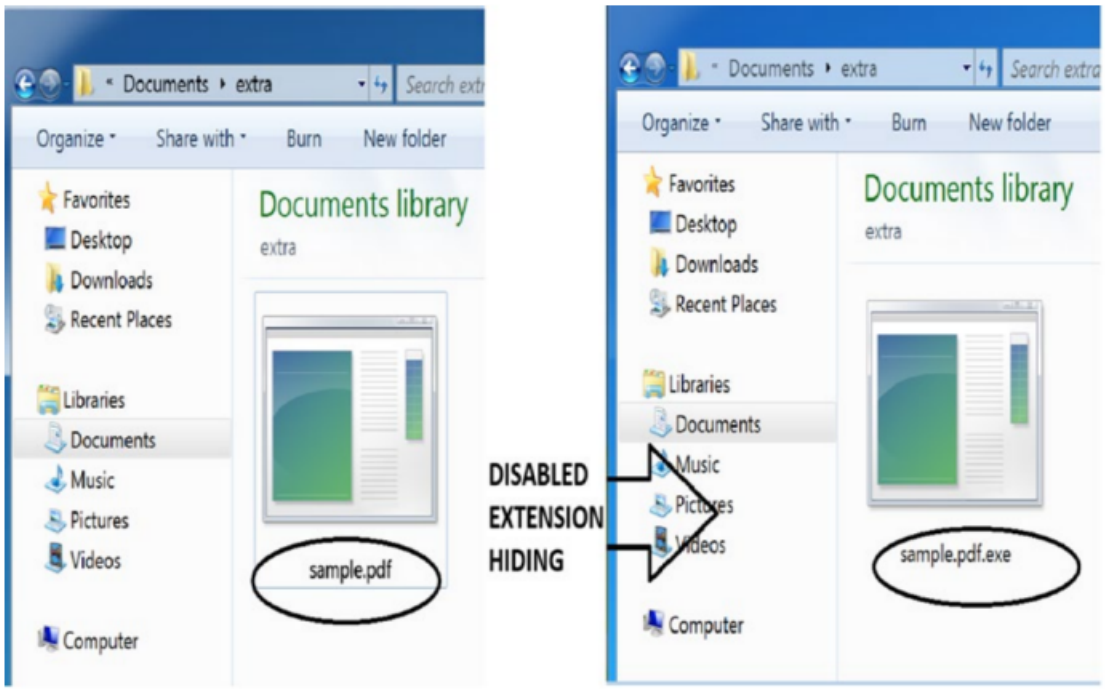

图 3- 10 恶意软件使用扩展名伪装欺骗用户

缩略图伪装

攻击者经常使用的另一种方法是伪造显示的缩略图来欺骗用户点击执行。如图 3- 11 所示，左侧看起来该文件是一个 PDF 文件。但其实文件的缩略图可以任意修改，攻击者就是修改了该可执行文件的缩略图。如右侧所示，在禁用扩展名隐藏后可见真正的文件扩展名为.exe。攻击者为该文件定制了 PDF 文件的缩略图，在扩展名隐藏的情况下诱使用户认为该文件为 PDF 文件，从而点击执行。

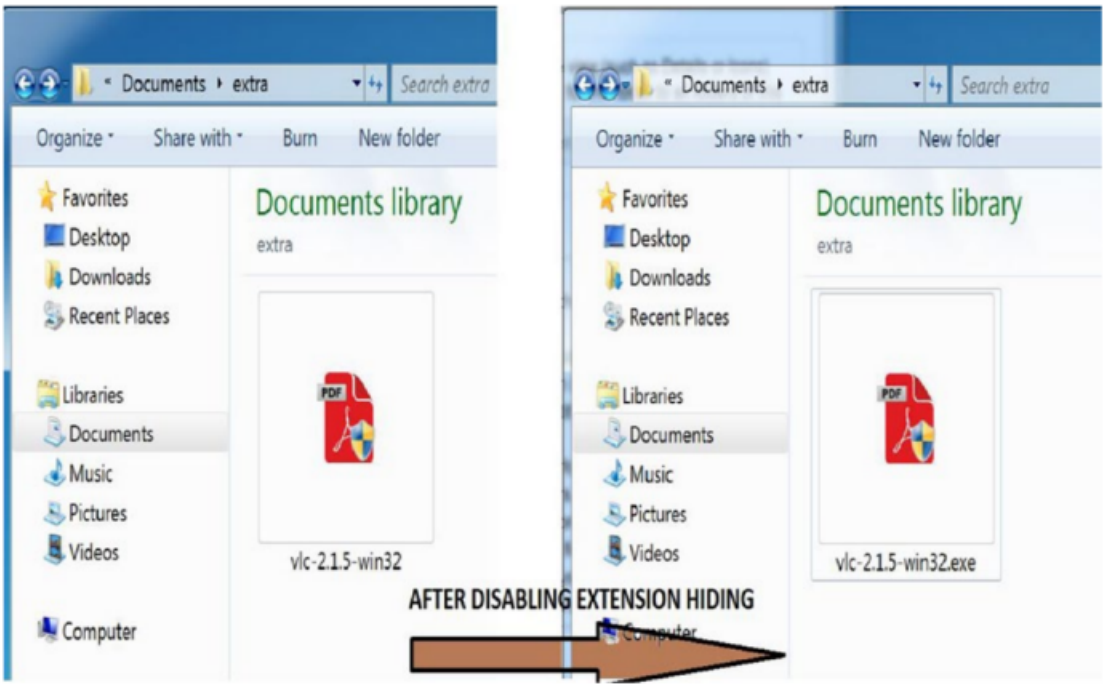

图 3- 11 使用 PDF 文件缩略图伪装的恶意可执行文件

常见文件扩展名

中列出了一些常见的文件扩展名与默认与之关联的程序，用户可以自行更换与特定文件扩展名关联的程序。例如，.pdf 扩展名的文件可以与 Adob e Acrobat PDF Reader、Foxit PDF Viewer 或者其他任何程序进行关联。

表 3- 2 部分常见的文件扩展名和默认与之关联的程序

文件扩展名	程序

.png、.jpeg、.jpg	Windows Photo Viewer
.pdf	Adobe Acrobat Reader
.exe	Windows loader
.docx、.doc、.pptx、.xlsx	Microsoft Office tools
.mp3、.avi、.mpeg	VLC Media Player

可以只依靠文件扩展名吗?

可以只依靠文件的扩展名来决定文件的类型吗? 答案是否定的。例如，将扩展名为.pptx 的文件的扩展名修改为.jepg 并不会使文件本身从 Microsoft PowerPoint 文件变更为 JPEG 图片文件。由于文件内容未发生更改，该文件仍然是 PowerPoint 文件。尽管扩展名错误，用户仍然可以强制 Microsoft PowerPoint 加载此文件。

作为恶意软件分析人员，这个问题是至关重要的。通常，恶意软件样本文件会在没有可读文件名和扩展名的情况下出现。此外，恶意软件还会使用虚假文件扩展名进行伪装来欺骗用户。后续将会介绍通过文件格式来识别文件类型的方法，这种方法相比扩展名更加可靠。

文件格式：真正的扩展名

此前使用 Notepad++十六进制模式打开过位于 C:\Windows\Notepad.exe 的文件，当然可以使用相同的方式打开系统上其他类型的文件，如 ZIP 压缩文件、PNG 图片文件等。值得注意的是，具有相同扩展名的文件在文件的起始处通常有相同的特定字符。例如，ZIP 压缩文件起始为 PK；PNG 图片文件的第 2-4 个字符是 PNG；而 Windows 可执行文件起始为 MZ。在图 3- 12 中可见，MZ 即为十六进制 4D 5A 的 ASCII 等价表示。

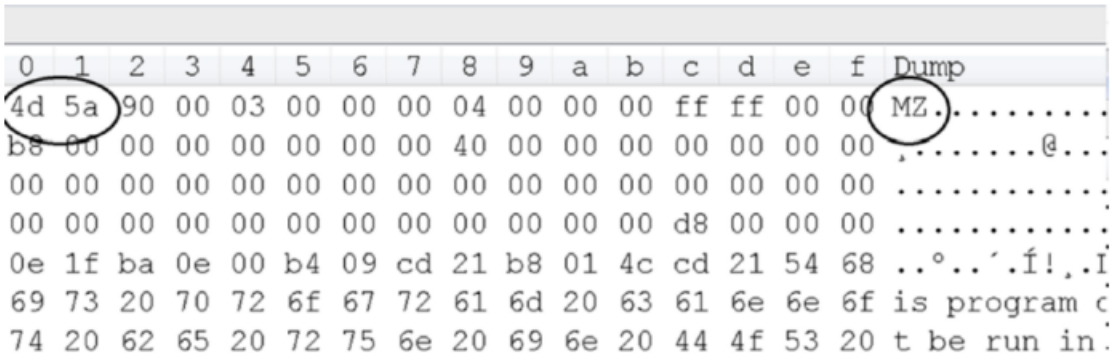

图 3- 12 可执行文件的 Magic Number

这些字节被称为 Magic Number，它们不是随机分布在文件中的，而是文件头的一部分。每类文件都有一个约定的结构或者格式，定义了数据应该如何存储在文件中。文件结构通常由文件头定义，其中包含有关存储在文件中的数据的元信息。解析文件头与 Magic Number 就可以识别文件的格式或者所属类型。

不论是音频文件、视频文件、可执行文件、PowerPoint 文档文件、Excel 文档文件、PDF 文档文件，都有对应的文件结构来存储其数据，这种文件结构被称为文件格式。解析文件格式可以获取更多关于文件的信息，例如 Windows 可执行文件的文件格式除了 MZ 的 Magic Number 之外，还保存着文件的其他特征（文件是 DLL 文件、EXE 文件还是 SYS 文件；文件是 32 位还是 64 位等）。通过文件格式来确定文件的实际类型，相比文件扩展名更加可靠。

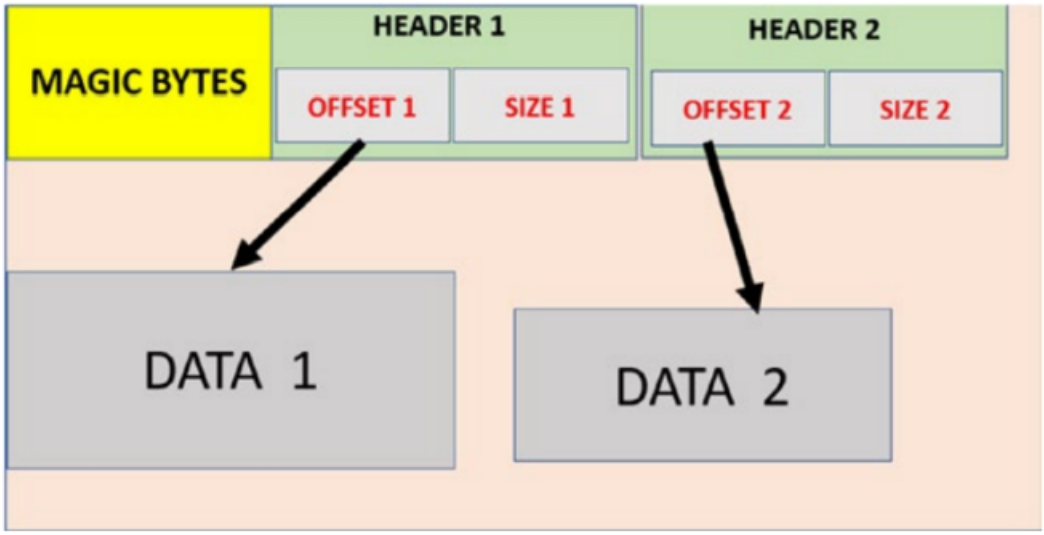

图 3- 13 文件结构与头信息的高层级概览

图 3- 13 给出了文件结构与头信息的一般性概览。如图所示，文件格式可以由多个文件头

定义，其中包含文件中保存的数据块的偏移量、大小和其他属性信息。

表 3- 3 与表 3- 4 中列出了常见的可执行文件与其他文件格式对应的 Magic Number。

表 3- 3 部分常见的可执行文件格式与对应的 Magic Number

操作系统	文件类型	Magic Number（十六进制）	Magic Number（ASCII）
Windows	Windows Executable	4D 5A	MZ
Linux	Linux Executable	7F 45 4C 46	.ELF
Mach-O	Mach-O Executable	FE ED FA CE	....

表 3- 4 部分常见的非可执行文件格式与对应的 Magic Number

文件类型	文件扩展名	Magic Number（十六进制）	Magic Number（ASCII）
PDF 文档文件	.pdf	25 50 44 46	%PDF
Adobe Flash 文件	.swf	46 57 53	FWS
Flash 视频文件	.flv	46 4C 56	FLV
AVI 视频文件	.avi	52 49 46 46	RIFF
ZIP 压缩文件	.zip	50 4B	PK
RAR 压缩文件	.rar	52 61 72 21	rar!
DOC 文档文件	.doc	D0 CF	

识别文件格式

有很多工具都可以识别文件格式，但有两个典型工具经常被分析人员使用。一个是 Linux 中的 file 命令，另一个是在 Windows、Linux 和 macOS 上都可用的 trid 命令（对应图形化界面工具为 TriDNet）。输入文件的路径作为命令行参数，这两个命令行工具即可给出文

件格式的判定结果。

TriD 和 TriDNet

在 Windows 中打开命令提示符并输入代码 3- 3 中所示的命令。

代码 3- 3 通过命令行使用 md5deep、sha1deep 与 sha256deep

```
c:\>trid.exe c:\Windows\notepad.exe
TrID/32 - File Identifier v2.24 - (C) 2003-16 By M.Pontello
Definitions found: 12117
Analyzing...
Collecting data from file: c:\Windows\notepad.exe
49.1% (.EXE) Microsoft Visual C++ compiled executable (generic)
(16529/12/5)
 19.5% (.DLL) Win32 Dynamic Link Library (generic) (6578/25/2)
 13.3% (.EXE) Win32 Executable (generic) (4508/7/1)
 6.0% (.EXE) OS/2 Executable (generic) (2029/13)
 5.9% (.EXE) Generic Win/DOS Executable (2002/3)
```

在代码 3- 3 中，trid 列出了文件可能的文件格式。trid 认为分析的文件 notepad.exe 有

49.1%的准确率是使用 Microsoft Visual C++编译的可执行文件，给出的概率越大是该文

件格式的可能性也就越大。

或者也可以使用 trid 命令行工具的图形化界面版本 TridNet，对相同的 notepad.exe 文件

分析如图 3- 14 所示。

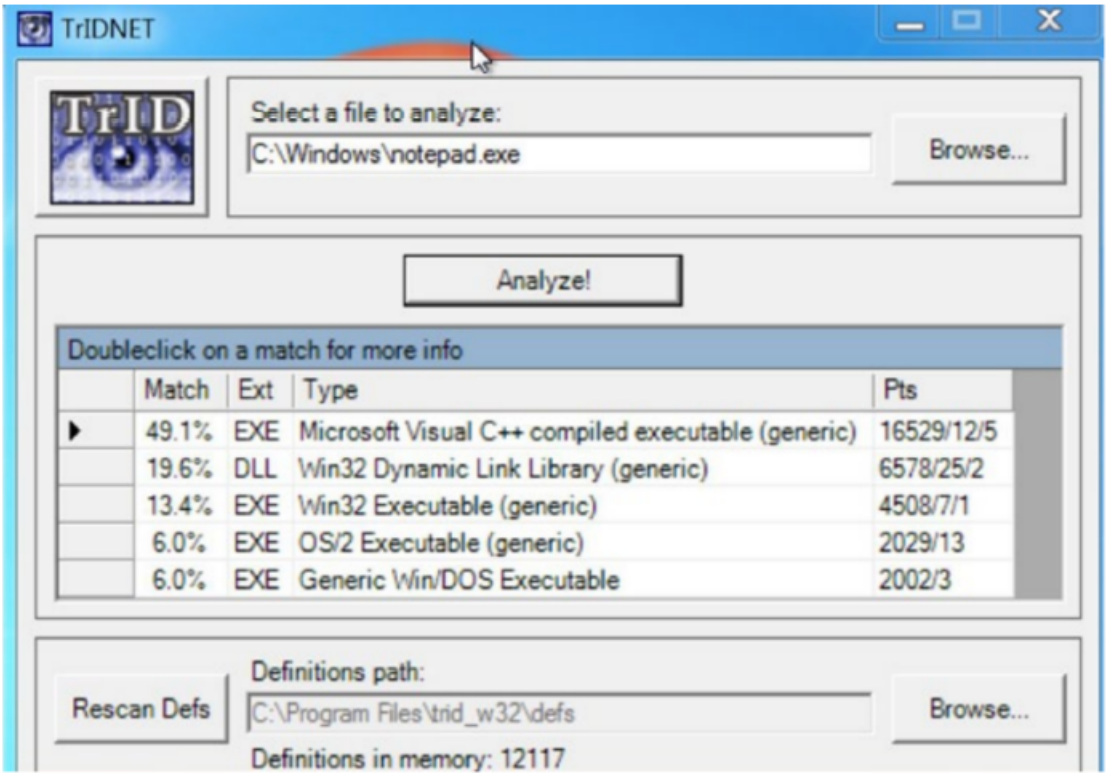

图 3- 14 文件识别工具 trid 的图形化界面版本 TriDNet

file 命令

另一个文件识别工具 file 在 Linux 上非常常用，该程序基于 libmagic 实现文件格式识别。

不仅是 file，libmagic 库也被许多其他检测工具用于识别文件格式。与 TriD 类似，file 也是将文件的路径作为参数提供，返回检测的文件格式，如代码 3- 4 所示。

代码 3- 4 Linux 上使用 file 命令识别可执行文件的文件格式

```
@ubuntu:~$ file notepad.exe
notepad.exe: PE32+ executable (GUI) x86-64, for MS Windows
```

手动识别文件格式

前文中介绍了 Magic Number、文件头与文件结构，分析人员可以利用这些特征手动识别文件格式。但随着 TriD 的分析工具的出现，似乎分析人员已经不再需要记住这些文件格式的详细信息再使用十六进制编辑器手动识别文件格式。

但必须强调的是，了解常见文件格式的 Magic Number 是十分有帮助的。作为恶意软件分

析人员，处理的数据可能来自网络包而且数量很大。在某些情况下，要分析的文件就被包含其中。通常来说，网络数据包中会包含攻击者的恶意软件或者包含嵌入外部父文件的其他文件。了解常见的文件格式的 Magic Number 和文件头结构有助于分析人员在海量的数据中快速识别其中是否存在文件，避免大海捞针，提高分析效率。例如，图 3- 15 为 Wireshark 分析的数据包文件，在 HTTP 响应数据中带有 ZIP 文件。根据 Payload 中的 Magic Number 为 PK 可以帮助分析人员快速判断：来自服务器的响应中包含一个压缩文件。在表 3- 4 中也能够进一步确认 ZIP 压缩文件的 Magic Number 为 PK。

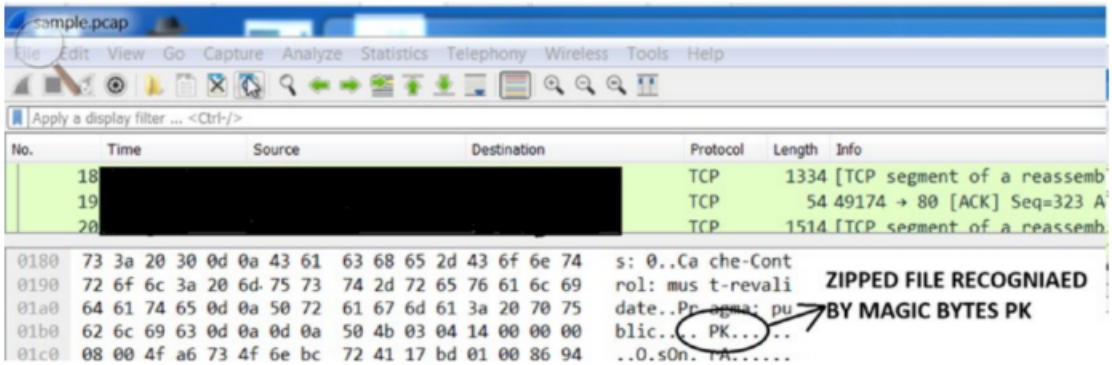

图 3- 15 使用 Magic Number 快速识别数据包中包含的文件

总结

本章中介绍了文件扩展名、文件格式、文件结构、Magic Number 与文件头等概念。使用 file 等命令行工具，即可以快速识别恶意软件的文件类型，并且能够根据其类型为文件配置正确的分析环境。Magic Number 可以帮助分析人员，在手动分析时识别数据（例如数据包与加壳文件等）中是否存在文件。