题目：恶意 PDF 文件解析思路
作者：Cryin'
时间：2010/12/03
链接：https://github.com/Cryin/PDFTear

概述

对于 PDF 文件的解析，必须先要熟悉 PDF 文件各式，貌似所有官方的 PDF 文件各式文档都是英文的。这样就没办法了，硬着头皮去读呗，如果你对自己的英语构自信那就到这里看吧参考[1]。除此之外就只能找一些国内撰写的相关资料了。熟悉了 PDF 文件各式之后，怎么解析 PDF 文件呢？我目前的办法是查找 PDF 文件里面的关键字段，这样做的弊端是对于 Obj 对象里的数据流对象(stream)包含的内容是没办法查找的 。另外也有一些 PDF 漏洞文件使用了一些混淆技术，这样的 PDF 文件暂时就没有什么好办法解析了。如下面的情况：

```
%PDF-1.5
1 0 obj
<</#54#79P#65 R 0 5 O#70e#6e#41c#74i#6fn 3 Pages C#61ta#6c#6f#67>>
endobj
```

关键字

这里就考虑一般的恶意 PDF 文件，主要是对以下关键字段(个人认为和漏洞不牵扯关系的就不考虑了)进行查找和解析，如下面所示：

```
obj
endobj
stream
endstream
xref
trailer
startxref
/Page
/Encrypt
/ObjStm
/JS
/JavaScript
/AA
/OpenAction
/AcroForm
/URI
/Filter
/JBIG2Decode
/RichMedia
/Launch
```

解析思路

这里要说的是几乎每一个 PDF 文件都包含有前 7 个字段，也有可能会不包含 stream 和 endstream。据说也有一些 PDF 文件没有 xref 或者 trailer，但是这种情况比较少见。如果一个 PDF 文件没有 xref 或者 trailer 关键字段，那么可以确定它不是恶意的 PDF 文件。

/xref 交叉引用表，描述每个间接对象的编号、版本和绝对的文件位置。PDF 文档中的第一个索引必须从版本为 65535 的 0 号对象开始，标识符/xref 后面的第一个数字是第一个间接对象(即 0 号对象)的编号，第二个数字是/xref (交叉引用表)的大小。

/Page 指明 PDF 文件的页数，大多数恶意 PDF 文件仅仅只有一页

/Encrypt 指明 PDF 文件有数字水印或者是被加密过的

/ObjStm 是 object streams 的数量。这里要明白 object streams 是一个可以包含其它 Object 对象的数据流对象

/JS 与/JavaScript 指明 PDF 文件中含嵌有 JavaScript 代码。我所见过的 PDF 恶意文件几乎全部嵌有 JavaScript 代码，这里一般都是利用 JavaScript 的解析漏洞或者使用 JavaScript 来实现堆喷射(heap spray)。当然要注意，在很多正常的 PDF 文件里也可以发现含有 JavaScript 代码

/AA、/OpenAction 和/AcroForm 指明当查看 PDF 文件或者 PDF 的某页时会有自动的动作随其执行，几乎所有嵌有 JavaScript 代码的恶意 PDF 文件都有自动执行 JavaScript 代码的动作(action)。如果一个 PDF 文件包含/AA 或/OpenAction 自动执行动作的关键字段，而且又含有 JavaScript 代码，那么这个 PDF 文件就极可能是恶意的 PDF 文件

/URI 如果你要在 PDF 文件中执行打开网页的动作的话就需要这个关键字段

/Filter 一般为 FlateDecode 即使用 zlib 压缩解压缩算法，详见参考[2]

/JBIG2Decode 指明 PDF 文件使用了 JBIG2 压缩。虽然 JBIG2 压缩本身可能会存在漏洞(CVE-2010-1297)。但/JBIG2Decode 关键字段并不能说明 PDF 文件是否可疑

/RichMedia Flash 文件

/Launch 执行动作(action)数量

最后的工作就是检查 PDF 文件的各个对象以及对象之间是否符合 Adobe 的 PDF 文件格式规范。并综合上面描述的各关键字段信息分析该 PDF 文件是否可能为恶意文件。

结束语

使用上面的思路，按照我目前的试验。对恶意的 PDF 文件检测的准确率还是蛮不错的，不过并不能百分百准确的检测出恶意的 PDF 文件，特别是分析一些经过所谓混淆技术或者特殊处理的 PDF 文件。暂时还没什么好办法。如果你有什么好的想法和思路，非常欢迎与我交流学习。我肯定还有更好的思路及方法来更准确的检测出恶意的 PDF 文件。对于这一点，杀毒软件是如何来做的我非常好奇。或许有一天我能进杀软公司的话，那就能探个究竟了！

参考

[1] http://www.adobe.com/devnet/pdf/pdf_reference.html
[2] http://www.zlib.com/