UNIVERSIDAD POLITÉCNICA DE SAN LUIS POTOSÍ

Análisis de características de PDFs maliciosos.

Hugo González

@hugo_glez

http://atit.upslp.edu.mx/~hugo/

UNIVERSIDAD POLITÉCNICA DE SAN LUIS POTOSÍ

Agenda

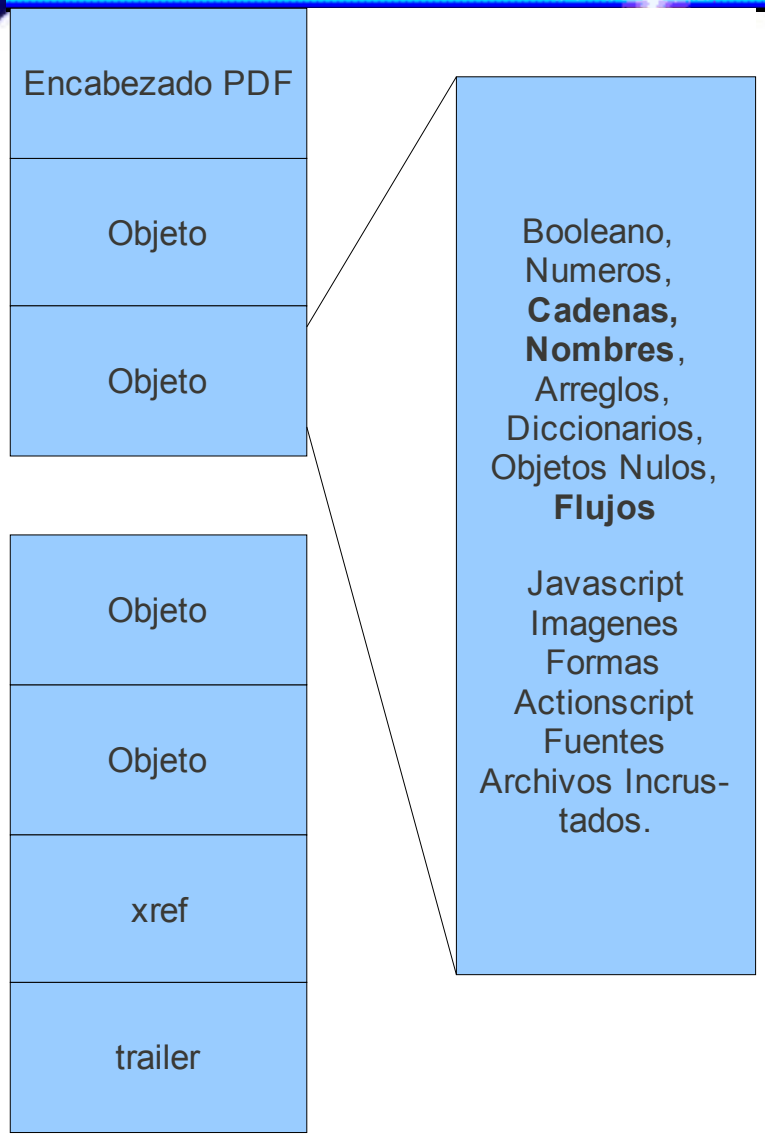
- Introducción
- El formato de archivos PDF
- Abusos en los archivos PDF
- Análisis de Malware
- Métodos y Herramientas
- Resultados
- Conclusiones

Introducción

- El formato PDF (del inglés Portable Document Format) tiene cerca de 17 años.
- En el 2008 se volvió un estándar abierto
- A evolucionado mucho al incluir características como Javascript, formularios …
- YA NO SE PUEDE CONSIDERAR SEGURO!

UNIVERSIDAD POLITÉCNICA DE SAN LUIS POTOSÍ

El formato de archivos PDF



El formato de archivos PDF

- Estándar ISO bastante extenso.
- El contenido es texto en formato ASCII.
- Este es interpretado por los “visores” de archivos PDF.
- Los objetos es donde se produce los abusos

%PDF-1.3
3 0 obj
<<</Type /Page
/Parent 1 0 R
/Resources 2 0 R
/Contents 4 0 R>>
endobj
4 0 obj
<<</Filter /FlateDecode /Length 449>>
stream
x<9c><8d>TË<92><9b>0^P¼ïWô1¹Lx<99>Çq<9d>Ê^^R<95>SÈæ,<83>0ò^B"<92>0ÅßG<80>½É.`ï^EQ3£<99
><9e>î^F^OB^_^\ÚEè^_ö)¾<¹p=r^\¤^E¾¥cÈwÉM^P;^^Å;¤9>=±<8c>kT<9d>.q<92>¢ÑÈy¦^F^TÿÂ^Ykd;^P
^^^M
%kÈ^B²30¥<92>Ý±^DktÏ^UÏ <81>×ä<97>jü©1pfl»^L<86>× ù<8c>ôô^FC^TÇ^TÆ^S<86><9a>sCøºq^Sc^Z
=+¥<84>âìÈ^Z<94>Lñ^Z=?s^K±n5ág5^VU¢á(ÍÆ&K^HQ@;w<82>ð¾<96>°^_ÀV7&ü^V¦^\^C]+^[û^@Ë2©rÑ^\
Á4Xa¸Ã<81><97>Ã²!Ïö½kr®h9=tÈ÷§éØ<9f> ÛNiè^A^G%{mg^V8±ì<85>U+7<83><90>Üp¾ùã.9ÐÕÃï3³ä<98
><92>õ«Bmé»^\n<8f>$<99><86>_¥·µcgi!´~^Q^FgÑ® v#<8a>g¶ñ¸É^S^N[¼ÿç´gÑâp<83>2hÛâ|Ýw<89>Äñ
)^L&$#^UÖ=^_k´AÜ^[§,÷^Nã<84>‹hÞ;ý igÍÖüLøuÏuK^DÑ<8e>¼Y²[<86>^LC{x^W<81>nH»¤þ<9e>¹^¿
eKÖ{Á<97>0¬»<93>ùçtß^K¯ø^HéVê6ôå|ß£è¢×5ù^WÆ^B±;
endstream
endobj
1 0 obj
<<</Type /Pages
/Kids [3 0 R]
/Count 1
/MediaBox [0 0 595.28 841.89]
>>
endobj
5 0 obj

Abusos

- Algunos abusos se dan por errores de programación en los “visualizadores”
- Otro es el uso de Javascript o ActionScript para ejecutar un archivo de internet.
- Otros aprovechan objetos mal formados que son atendidos por terceros. (librerias)
- Un gran problema para detectarlos es la OFUSCACIÓN.

```
^BPx^A^Ky^L#,^V&^YX<88>l®^U<9c>ivul@<AXe§v^W0 1<82>·Qul]<90>^N?M.lu[<80>^A^@<85>^ZA ^M
endstream^Mendobj^M
56 0 obj
<<</S/JavaScript/JS(/*Ubiquity*/
var needsUpdate = 0;
if \(app.viewerType == "Exchange" && app.viewerVariation=="Fill-In"\)\r    needsUpdate
 = 1;\relse if \(app.viewerType == "Reader"\)\r{\r    if \(app.viewerVersion >= 5.1\)\
r        needsUpdate = 0;\r    else\r        ne\^MedsUpdate = 1;\r}\rif \(needsUpdate\
)\r{\r    var ans = app.alert\("This document requires one of the following\\nin order
 to be used as the author intended:\\n\\n - Acrobat Reader 5.1 or later\\n - Acrobat 5
.0.5 or later\\n - Acrobat Approval 5.0.5 or la\^Mter\\n\\nYou are viewing this docume
nt with an earlier\\nversion of one of these products.  As a result,\\nsome features\\
nof the document will not operate.\\n\\nClick OK to go on-line and obtain the latest v
ersion\\nof the free Adobe Reader.\\n", 1, 1\);\^M\r    if \(ans == 1\)\r        this.
getURL\("http://www.adobe.com/acrobat", false\);\r}>>
endobj
@
```

Análisis de Malware

- Análisis de comportamiento
- Análisis Estático
- Análisis Dinámico

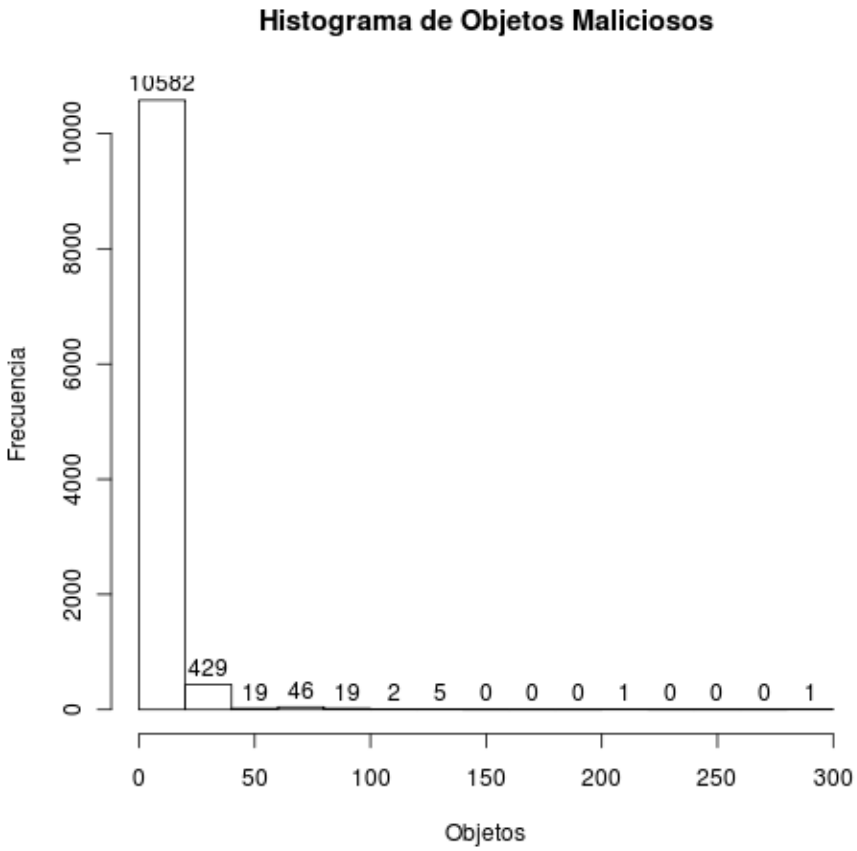
Métodos y Herramientas

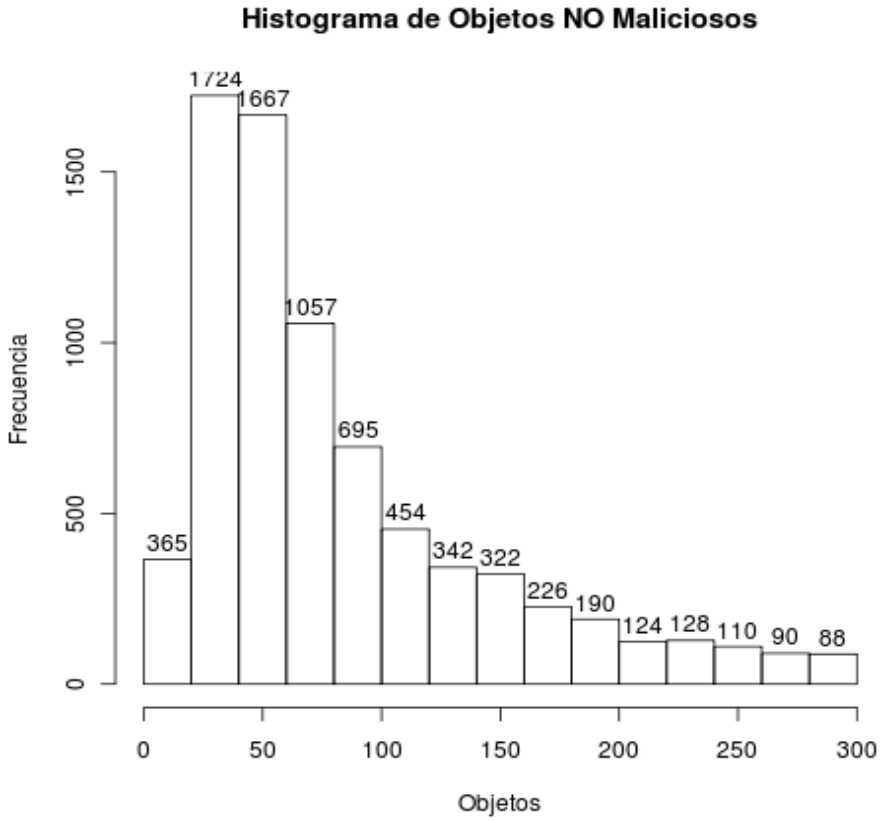
- Muestra 11, 104 archivos maliciosos.
- Muestra 8, 669 archivos no malicioso.
- 200 archivos de muestra obtenidos en Internet.
- Se revisaron con CLAMAV antivirus
- Se utilizó PDFiD y PDF-parser, de Didier Stevens
- Desarrollos propios.

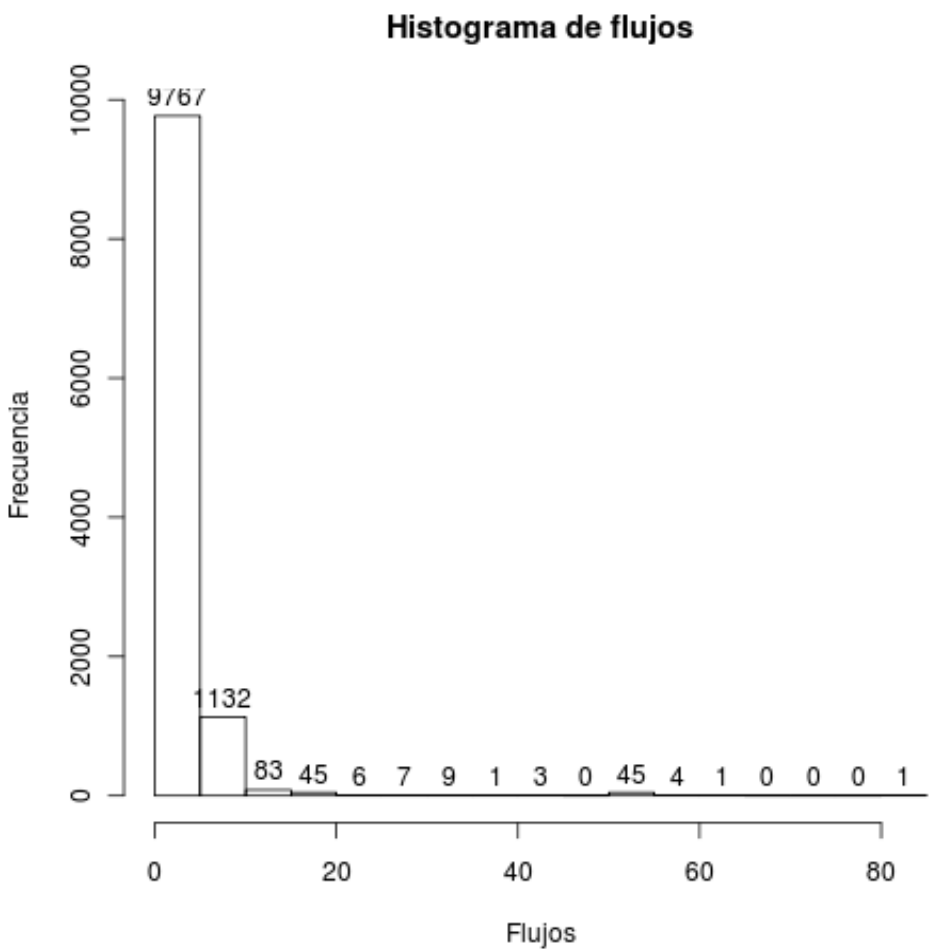
1. encabezado
2. cantidad de objetos
3. cantidad de flujos
4. cantidad de xref
5. cantidad de trailer
6. cantidad de xref de inicio (startxref)
7. total de paginas
8. total de elementos encriptados
9. total de flujos binarios (objstream)
10. total de javascript identificado con JS
11. total de javascripts
12. total de actionscript
13. existencia de acciones al abrir
14. total de formas
15. total de objetos multimedia
16. existencia de accion cargar
17. total de nombres de identificador ofuscados.

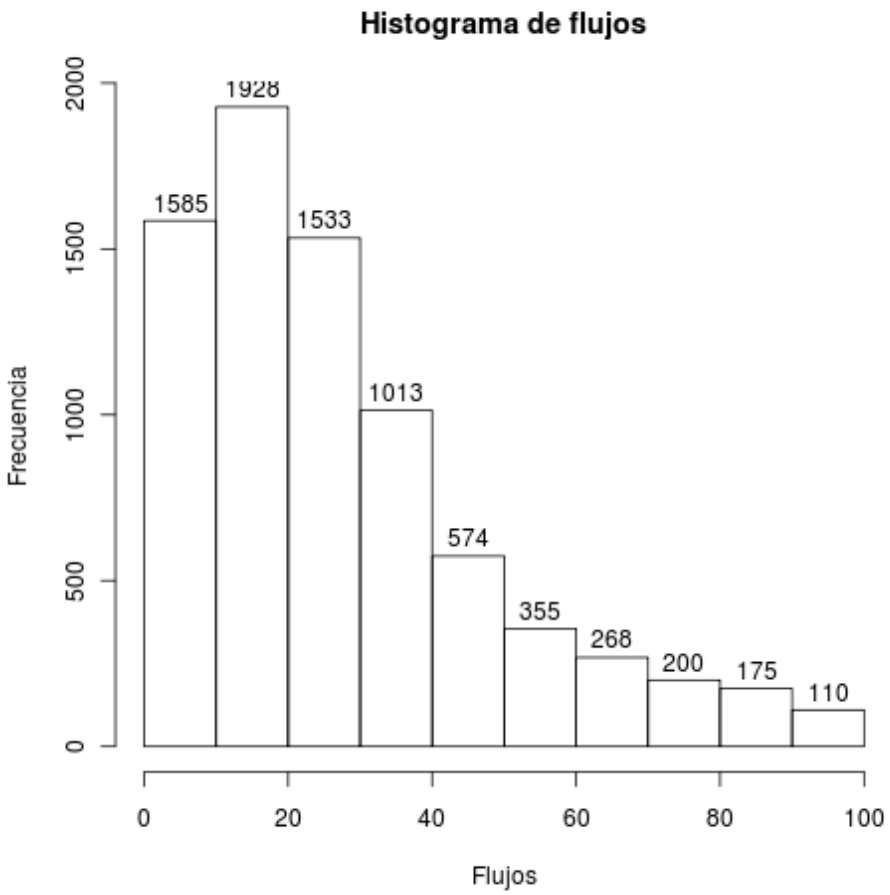
Resultados

- Al utilizar el antivirus clamav para linux con la base de datos más reciente se detectaron 8,775 archivos maliciosos solamente, equivalente al 79% dejando 2,329 archivos maliciosos sin identificar.
- Del total de los 8,779 archivos considerados no maliciosos, ninguno fue detectado como tal por el antivirus Clamav.


Histograma de Objetos Maliciosos
Frecuencia
0
2000
4000
6000
8000
10000
10582
429
19
46
19
2
5
0
0
0
1
0
0
0
1
0
50
100
150
200
250
300
Objetos


Histograma de Objetos NO Maliciosos
Frecuencia
0
500
1000
1500
365
1724
1667
1057
695
454
342
322
226
190
124
128
110
90
88
0
50
100
150
200
250
300
Objetos


Histograma de flujos
Frecuencia
0
2000
4000
6000
8000
10000
9767
1132
83
45
6
7
9
1
3
0
45
4
1
0
0
0
1
0
20
40
60
80
Flujos


Histograma de flujos
Frecuencia
0
500
1000
1500
2000
1585
1928
1533
1013
574
355
268
200
175
110
0
20
40
60
80
100
Flujos

	Característica	Archivos maliciosos	Archivos NO maliciosos
	Detección de Clamav	79 % Detectados	0% Detectados
1	encabezado	Un 6% tiene encabezados mal formatos.	Los encabezados son bien formados
2*	cantidad de objetos	95% tiene menos de 20. 99% tiene mesnos de 40.	23% tienen menos de 40 . 43% tienen menos de 100 .
3*	cantidad de flujos	99% tiene menos de 12.	89% tiene menos de 100.
4*	cantidad de xref	99% tiene menos de 2.	35% tiene menos de 2. 50% tiene 2.
5*	cantidad de trailer	99% tiene menos de 2 .	35% tiene menos de 2. 50% tiene 2.

6*	cantidad de xref de inicio (startxref)	99% tiene menos de 2 .	27% tiene menos de 2. 53% tiene 2.
7*	total de paginas	92% tiene 1.	14% tiene 1.
8	total de elementos encriptados	0.04%	3.6%
9	total de flujos binarios (objstream)	1%	27%
10*	total de javascript identificado con JS	91%	6%
11*	total de javascripts	91%	6%

12	total de actionscript	2%	6%
13*	existencia de acciones al abrir	75%	10%
14	total de formas	17%	21%
15	total de objetos multimedia	0.1%	0%
16	existencia de acción cargar	0.7%	0.9%
17*	total de nombres de identificador ofuscados.	4.4%	0%

Propuesta

Si un archivo cumple lo siguiente es sospechoso:

- **menos de 40 objetos y**
- **menos de 12 flujos y**
- **menos de 2 xref y**
- **menos de 2 trailers y**
- **menos de 2 xref de inicio y**
- **menos de 2 paginas o**
- **contiene javascript o js**

- Aplicando lo anterior a los archivos obtenidos de Internet, **10%** de ellos las lo cumplen. Estos son sospechosos y en un análisis posterior se descartan como maliciosos. Aplicando la misma propuesta al conjunto de archivos maliciosos un **99.7%** lo cumplen.

Conclusiones

- Los atacantes van ideando nuevas formas de ofuscar el código malicioso dentro de los PDFs para evitar ser detectados.
- Con las características aquí presentadas, ayudarán a identificar posibles archivos sospechosos, que deberán ser sometidos a otro tipo de análisis para identificarlos plenamente.

Trabajo a futuro

- Implementación de un software automatizado para detectar archivos PDF sospechosos.
- Utilización de redes neuronales para clasificación automática de archivos basado en estas características.

UNIVERSIDAD POLITÉCNICA DE SAN LUIS POTOSÍ

Agradecimientos

- Universidad Politécnica de San Luis Potosí.
- Mila, en el Blog “Contagiodump” por las muestras maliciosas.