A Fast and Precise Malicious PDF Filter

Wei Xu, Xinran Wang, Huagang Xie, Yanxin Zhang

Palo Alto Networks

Sep 26, 2012

Outline

- Introduction
- Portable Document Format (PDF)
- Overview
- Design
- Evaluation
- Summary

paloalto NETWORKS

Introduction

- PDF documents have become a popular vector for malware distribution
  - PDF documents are less likely to be blocked by e-mail servers
  - Majority users are still using vulnerable versions of PDF readers
- Existing techniques are limited by scalability
  - MDScan, Wepawet …
- Goal: A PDF filter that can discard the benign PDFs very quickly with high precision

paloalto NETWORKS

Portable Document Format (PDF)

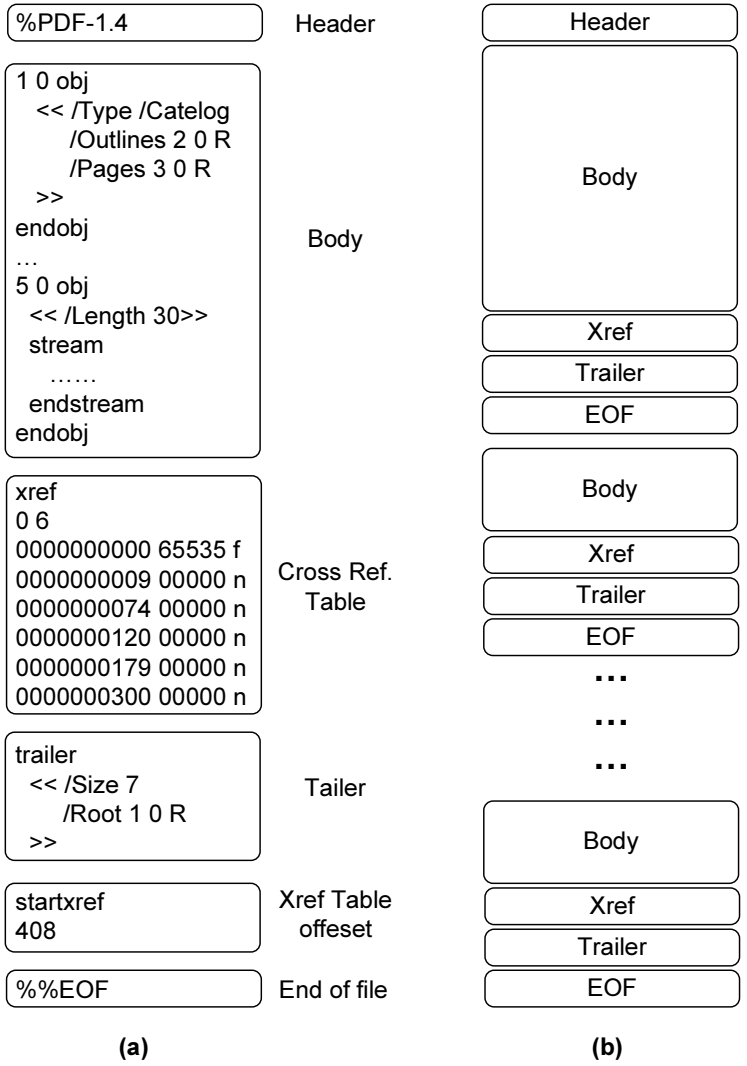
- **Format Specification**
  - **A 8-bit binary file format created by Adobe in 1993**
  - **"A complete description of a fixed-layout flat document, including the text, fonts, graphics, and other information needed to display it" [1]**
  - **Version**

Year	Version
2003	PDF 1.5 / Acrobat 6.0
2005	PDF 1.6 / Acrobat 7.0
2006	PDF 1.7 / Acrobat 8.0 (ISO 32000-1)
2008	PDF 1.7, Adobe Extension Level 3 / Acrobat 9.0
2009	PDF 1.7, Adobe Extension Level 5 / Acrobat 9.1

paloalto NETWORKS

Portable Document Format (PDF)

- PDF file structure
  - Header
  - Body
  - Cross-reference table
  - Trailer
  - EOF marker
- Linearization (optimization)
- Incremental update



paloalto NETWORKS

Portable Document Format (PDF)

- **Common PDF Exploits**
  - **JavaScript**
  - **Flash Objects**
    - **(ActionScript)**
  - **TIFF image objects**
  - **XFA Stream**

```
function heapspray() {
        var nop = unescape('%u9090%u9090');
        var shellcode = unescape("%u56e8.....%u246c");
        while (nop.length <= 0x10000/2) {
                nop += nop;
        }
        nop = nop.substring(0, 0x10000/2 – shellcode.length);
        memory = new Array();
        for (i=0; i<0x1000; i++) {
                memory[i] = nop + sc;
        }
}

heapspray();
try {
        this.media.newPlayer(null);
}
catch (e) {}
```

Shellcode

Heap spray

Exploit (CVE-2009-4324)

paloalto NETWORKS

Portable Document Format (PDF)

- Evasion techniques
  - String splitting
  - Split into various objects and combined later
    - Obj.getField()
    - Small data chunk concatenation
  - Encryption
  - Multi-level encoding
  - etc.

paloalto NETWORKS

Overview

- **PDF filter**
  - **Differentiate benign and malicious PDF documents?**
  - **Features**
    - **Structure of PDF**
    - **Functionalities**
    - **Embedded code**
  - **Machine Learning**

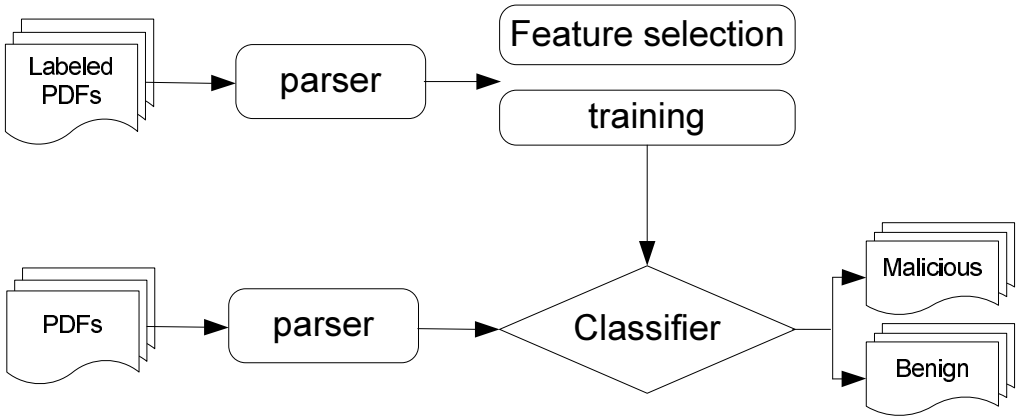


Figure 3: Architecture of the PDF filter

paloalto NETWORKS

Design

- **Features on Embedded Code**
  - **Number of occurrence of "/JavaScript" ("/JS") action**
    - **In clear-format**
    - **In encoded-format**
  - **Invocation of suspicious JavaScript functions**
    - **Obfuscation**
    - **To exploit vulnerabilities in JS functions**

paloalto NETWORKS

Design

- **Suspicious JavaScript Functions**

Table 1: List of Suspicious JavaScript Function Invocation

Suspicious JavaScript Function	Indication
eval()	obfuscation
str.concat()	obfuscation
str.replace()	obfuscation
str.fromCharCode()	obfuscation
str.split()	obfuscation
str.substr()	obfuscation
str.substring()	obfuscation
util.printf()	CVE-2008-2992
doc.media.newPlayer()	CVE-2009-4324

paloalto NETWORKS

Design

Features on PDF Functionalities

Potentially Harmful PDF Actions

Table 2: List of Potentially Harmful PDF Actions

Potentially Harmful PDF Action	Definition
/Action	A class of actions triggered by user
/OpenAction	A class of actions triggered by opening the PDF file
/GoTo	[F] redirection within the document
/GoToR	[F] redirection to external src
/GoToE	[F] redirection to embedded file
/Launch	[F] launch an application
/SubmitForm	[F] send interactive data to a URL
/URI	[F] Access remote URL
/ImportData	[F] Import external data

paloalto NETWORKS

Design

- **Features on PDF Functionalities**
  - **Misused PDF Filters**
    - **Encode malicious content**

Table 4: Stream Filter Related Features

Feature	Note
Number of filters > 2 for one stream object	excessive number of filters
Suspicious filters pipeline	e.g., [/JBIG2Decode /DCTDecode /ASCIIHexDecode]
Escaped filter name	#hh in a filter name, e.g., /A#53#43#49#49#48#65#78Dec#6f#64e
Unknown filter name	Filters that are not supported by PDF format

paloalto NETWORKS

Design

- **Features on PDF Structure**
  - **Malformed/Mismatched elements in PDF files**

Table 5: Structure Related Features

Feature
Malformed “startxref”
Malformed “trailer”
Malformed “xref”
Code after last “EOF”
Average Object size

  - **Statistic features of elements in PDF files**
    - **Avg. size of objects**
    - **Number of objects**

paloalto NETWORKS

Evaluation

- **Implementation**
  - PDF parser -> Feature extractor -> Classifier
- **Samples**
  - Training set: 25,204 (19,518 benign samples, 5,686 malicious samples)
  - Evaluation set: 157,842 (Download from Google)

paloalto NETWORKS

Evaluation

- Compares different machine learning model
  - Linear model (FP: 11.42%, FN:1.03%)
    - Light-weight, fast, adjustable (e.g., online filtering)
  - Other models
    - 8 different machine learning models

Table 6: FP and FN of different machine learning models

Classifier	FP (%)	FN (%)	Note
Random Forest	8.6	1.4	better accuracy but large model
Bayes Net	1.2	24.2	Low FP, High FN
J48	9.2	1.8	Better FP, but large model

paloalto NETWORKS

Summary

- A set of predictive features that can effectively detect malicious PDF documents
- Features cover three aspects of malicious PDF documents: embedded code, PDF functionalities and PDF structure
- Evaluation on over 25,000 labeled samples and over 150,000 real world PDF documents shows high detection rate and low false positive rate
- Compare different machine learning models to study the trade-off between performance and accuracy and to better tune the filter

paloalto NETWORKS

Thanks
Q&A

paloalto NETWORKS