symantec™

Confidence in a connected world.

vb 2008 OTTAWA

Clean Data Profiling

Bartek Uscilowski

Julie Weber

October 2008

Introduction & outline

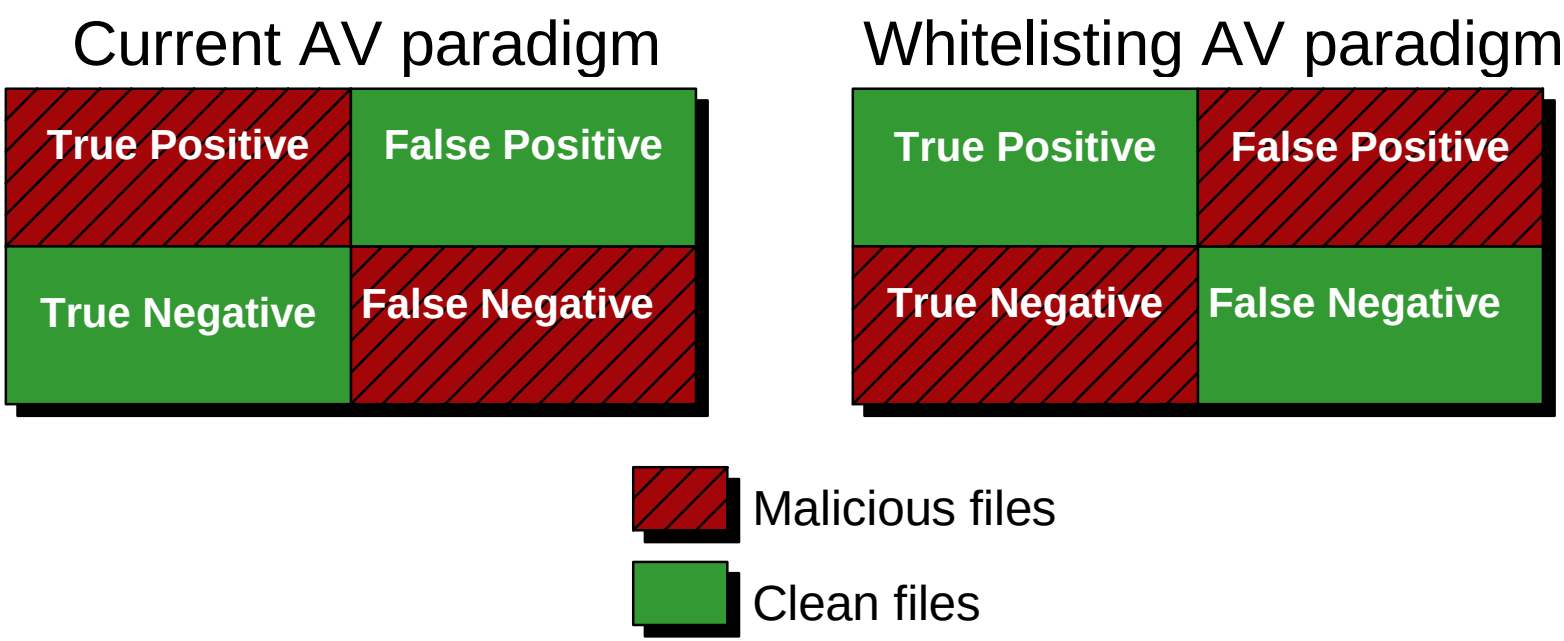
- The paradigm change
- Clean Data sets
- Metadata
  - Lower-level metadata
  - Higher-level metadata
- Benefits of profiling
- Summary

Outline

- The paradigm change
- Clean Data sets
- Metadata
  - Lower-level metadata
  - Higher-level metadata
- Benefits of profiling
- Summary

The paradigm change

- Exponential growth of malware
- Current AV paradigm possibly insufficient
- Increasing importance of clean data



Outline

- The paradigm change
- Clean Data sets
- Metadata
  - Lower-level metadata
  - Higher-level metadata
- Benefits of profiling
- Summary

Clean Data Sets

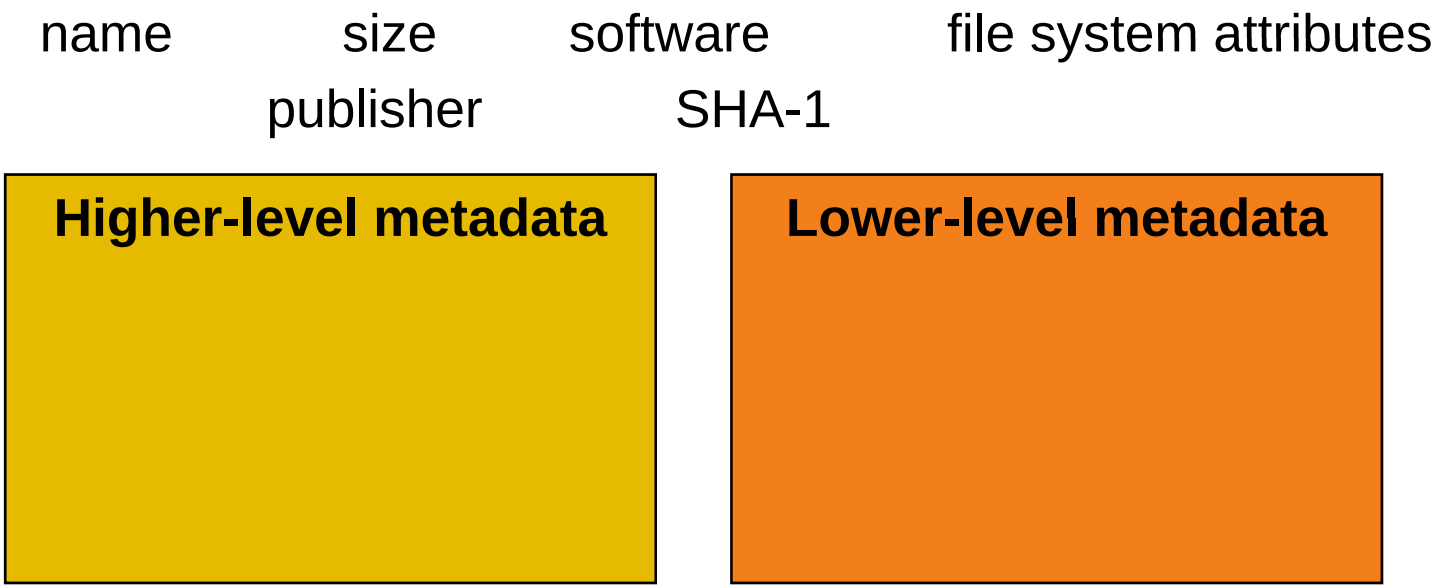
- Set of data derived from legitimate software
- Targeted
  - Specific purpose (FP ?)
  - Clean data is associated with software that is known and identified in advance
- Non targeted
  - Degree of randomness
  - May have specific purpose
  - Web crawler, user images…

Outline

- The paradigm change
- Clean Data sets
- Metadata
  - Lower-level metadata
  - Higher-level metadata
- Benefits of profiling
- Summary

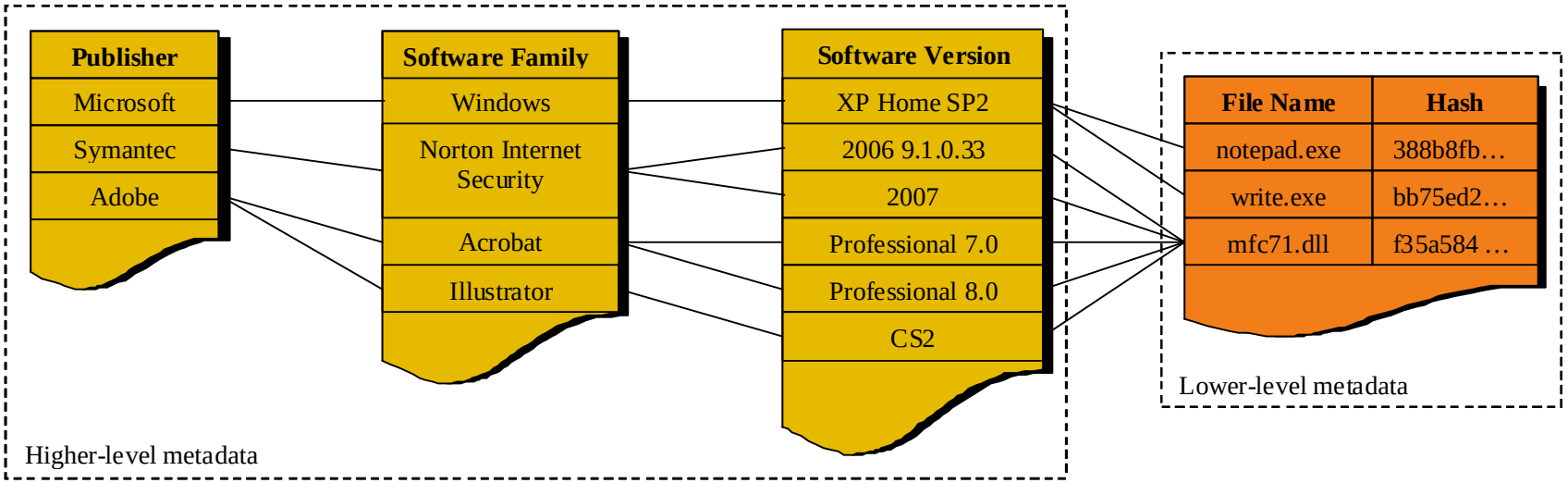
Metadata

- Describes the features of the set
- Information relative to a file
  - Example:



Metadata

- Relationship between metadata levels



- Files vs associated software & publisher
- Lower-level – extracted from files
- Higher-level – annotated or profiled

Outline

- The paradigm change
- Clean Data sets
- Metadata
  - Lower-level metadata
  - Higher-level metadata
- Benefits of profiling
- Summary

Lower-level metadata

- Hashes
  - → MD5, SHA-1, SHA-2 family
  - → cryptographic libraries
- Size, attributes
- MIME and Magic
  - → Type of file
  - → *GNU file* tool

Example for mfc71.dll

```
MD5:    f35A584E947A5B401FEB0FE01DB4A0D7
SHA-1:  664dc99e78261a43d876311931694b6ef87cc8b9
SHA-256: 22d1fab8518459265dbd54b9bbc4247c9d4d45bb99b3b4baec589973d
```

```
C:\>file -i C:\windows\notepad.exe
C:\windows\notepad.exe: application/x-dosexec

C:\>file C:\windows\notepad.exe
C:\windows\notepad.exe: PE executable for MS Windows (GUI) Intel 80386 32-bit
```

Lower-level metadata

- Architecture
  - → For PE files: 16, 32, 64 bits
- PE Version information
  - → Version info in executable file
  - → Sigcheck
- Digital signatures
  - → Signers chain, dates
  - → Sigcheck, Signtool

```
sigcheck v1.53 - sigcheck
Copyright (c) 2004-2008 Mark Russinovich
Sysinternals - www.sysinternals.com

c:\windows\notepad.exe:
        Signers:
            Microsoft Windows Publisher
            Microsoft Windows Verification Intermediate PCA
            Microsoft Root Authority
        Signing date:   10:03 04/08/2004
        Publisher:      Microsoft Corporation
        Description:    Notepad
        Product:        Microsoft® Windows® Operating System
        Version:        5.1.2600.2180
        File version:   5.1.2600.2180 (xpsp_sp2_rtm.040803-2158)
        Original Name:  NOTEPAD.EXE
        Internal Name:  Notepad
        Copyright:      © Microsoft Corporation. All rights reserved.
        Comments:       n/a
```

```
C:\>file C:\windows\notepad.exe
C:\windows\notepad.exe: PE executable for MS Windows (GUI) Intel 80386 32-bit
```

Lower-level metadata

```
C:\>signtool verify /a /v /tw 3600 c:\WINDOWS\notepad.exe
Verifying: c:\WINDOWS\notepad.exe
File is signed in embedded mode.
Timestamp Verified by:
    Issued to: Microsoft Root Authority
    Issued by: Microsoft Root Authority
    Expires:   31/12/2020 08:00:00
    SHA1 hash: A43489159A520F0D93D032CCAF37E7FE20A8B419
        Issued to: Microsoft Root Authority
        Issued by: Microsoft Root Authority
        Expires:   10/05/2021 08:00:00
        SHA1 hash: 3EA99A60058275E0ED83B892A909449F8C33B245
            Issued to: Microsoft Windows Verification Intermediate PCA
            Issued by: Microsoft Root Authority
            Expires:   29/12/2010 08:00:00
            SHA1 hash: 1C324CA9520DD0690882292DD85E2671CAE9E
                Issued to: Microsoft Windows Component Publisher
                Issued by: Microsoft Windows Verification Intermediate PCA
                Expires:   10/06/2009 23:07:51
                SHA1 hash: 612FC4EC7012F375A249EE9DE2D8E1AA363
Number of files successfully Verified: 1
Number of warnings: 0
Number of errors: 0
```

Lower-level metadata

- Caveats
  - PE Version information
  - Detached signatures
    - Where are .cat files?
    - How to use .cat files?
  - File type based on filename extension

Symantec
Symantec Corp
Symantec Corp.
Symantec Corporation
Symantec, Inc.
Symantec Ltd.
Symantec New Zealand Limited
Symantec Corporationn
Symantec Corporation http://www.symantec.com

Outline

- The paradigm change
- Clean Data sets
- Clean data generation
- Metadata
  - Lower-level metadata
  - Higher-level metadata
- Benefits of profiling
- Summary

Higher-level metadata

- Nature
  - Publisher
  - Software name
  - Software version
  - Locale
  - Source of software/file
- Sourcing
  - Targeted datasets
    - → known in advance
    - → tracked
  - Non-targeted datasets
    - → Profiling techniques

Symantec
Norton Internet Security
2006 9.1.0.33
German
Read-only physical media (e.g. CD/DVD)

Higher-level metadata (profiling)

1. Cross-reference

- Pros
  - Accurate
  - Consistent

- Cons
  - Requires annotated set
  - Quality of the reference set

Higher-level metadata (profiling)

2. PE Version information

- Pros
  - Easily automated
  - Extracted directly from files
  - Could be interpolated on non-PE files
- Cons
  - Extracted only from PE files
  - Not accurate. A file may be shipped with various applications
  - Inconsistent / missing

Higher-level metadata (profiling)

3. W

- P
- C

Web

??0x80070422:???????,??????????????? ...
DLL - MFCDLL Shared Library - Retail Version - **f35a584e947a5b401feb0fe01db4a0d7** O40 -
Explorer.EXE - Microsoft Corporation - C:\Program Files\Common ...
bbs.winos.cn/viewthread.php?tid=35163 - 48k - Cached - Similar pages - Note this

MFC71.dll - File Information [Is mine authentic?]
File Hash: **F35A584E947A5B401FEB0FE01DB4A0D7**. File HashType: MD5. File Size: 1.01
MB (1060864 Bytes). File Modified Time: 8/4/2006 4:00:00 PM. Date Added: ...
www.programchecker.com/file/8466.aspx - 76k - Cached - Similar pages - Note this

AOFroobs.com :: View topic - Problem With Loader
27 Aug 2007 ... MFC71.dll:**f35a584e947a5b401feb0fe01db4a0d7**
msvcp71.dll:561fa2abb31dfa8fab762145f81667c2
msvcr71.dll:86f1895ae8c5e8b17d99ece768a70732 ...
www.aofroobs.com/forum/viewtopic.php?t=5309 - 75k - Cached - Similar pages - Note this

Spyware Browser Antispyware, adware and malware removal.
f35a584e947a5b401feb0fe01db4a0d7, MFCDLL Shared Library - Retail Version. MFC71.DLL
Microsoft Corporation. Classified as: Good application / MFC ...
www.antispyware.tv/reports/-/231/?cmd=1 - 185k - Cached - Similar pages - Note this

Higher-level metadata (profiling)

4. Source i.e. URL

- Pros
  - Easy for files gathered by crawling the internet
- Cons
  - Does not establish all metadata, mostly either publisher or software
  - Difficult for automation but not impossible

FTP source:

ftp://gamefiles.bflyonder.co.uk/bflyonder-games/unrealtournament/modifications/defencealliance/patches/win32/dabeta1.exe

This would indicate that the file is a patch for the game Unreal Tournament published by Epic GamesFTP

FTP source:

ftp://ftp.dell.com/sas-non-raid/BR123319.exe

This would indicate that the file is a Dell driver

Outline

- The paradigm change
- Clean Data sets
- Clean data generation
- Metadata
  - Lower-level metadata
  - Higher-level metadata
- Benefits of profiling
- Summary

Benefits of profiling

- Analysis & fitness for business (Quality Assurance)
- Making purpose built sets
- Improvement / extension of existing data sets
- Easy access to samples (hash lookup)
- Detailed information for engineers and customers

Outline

- The paradigm change
- Clean Data sets
- Clean data generation
- Metadata
  - Lower-level metadata
  - Higher-level metadata
- Benefits of profiling
- Summary

Summary

- Profiling :
- What is the metadata?
- How to obtain the metadata?
  - Lower level is in the file
  - Higher level:
    - known for targeted sets
    - 4 techniques for non-targeted sets:
      - Cross Reference
      - PE Version Information
      - Web Search
      - Source

Summary

- Clean Data Sets: what are they for?
  - Currently: FP prevention
  - Future: potential paradigm change?

Questions

Finally

Clean data is becoming more and more important

Metadata is vital

Several profiling techniques exist

Thank You!

Bartlomiej_uscilowski@symantec.com

Julie_weber@symantec.com