FFRI Dataset 2018 の紹介

アジェンダ

- データセット提供の目的と現状
- FFRI Dataset 2013～2017 (動的解析ログ)
- データセットに関するアンケート結果
- FFRI Dataset 2018
- 意見募集

データセット提供の目的と現状

目的

- 研究分野における FFRI の知名度向上と人材交流・共同研究

現状

- 人材交流、共同研究への発展は少ない
- FFRI Dataset を使用した研究論文が減少傾向
- ここ数年、マルウェアの挙動に着目した研究が下火、機械学習を用いた研究が盛ん

ToDo

- ニーズに応えるデータセットの提供（MWS Cup 2017 課題3アンケート）
- 自らデータセットを使って研究する

FFRI Dataset 2013～2017 (動的解析ログ)

- FFRI が収集したマルウェアの動的解析ログ

  - 2013年: 約2,600検体分
  - 2014年: 約3,000検体分
  - 2015年: 約3,000検体分
  - 2016年: 約8,000検体分
  - 2017年: 約6,200検体分

具体的なデータ項目

項目（大見出し）	内容
info	解析の開始、終了時刻、id等（idは1から順に採番）
signatures	ユーザー定義シグニチャとの照合結果（今回は使用無）
static	検体のファイル情報（インポートAPI、セクション構造等）
dropped	検体が実行時に生成したファイル
behavior	検体実行時のAPIログ（PID、TID、API名、引数、返り値、動作概要等）
target	解析対象検体のファイル情報（ハッシュ値等）
debug	検体解析時のCuckoo Sandboxのデバッグログ
strings	検体中に含まれる文字列情報
network	検体が実行時に行った通信の概要情報

具体的なデータ項目(behavior)


```
"processtree": [
    {
        "children": [],
        "process_name": "lsass.exe",
        "command_line": "C:¥¥Windows¥¥system32¥¥lsass.exe",
        "track": false,
        "pid": 492,
        "ppid": 376,
        "first_seen": 1493811724.752586
    },
    {
        "children": [
            {
                "children": [
                    {
                        "children": [],
                        "process_name": "mshta.exe",
```

具体的なデータ項目(behavior)


```
"summary": {
    "regkey_deleted": [
        "HKEY_CURRENT_USER¥¥Software¥¥Microsoft¥¥Windows¥¥C…
        "HKEY_LOCAL_MACHINE¥¥SOFTWARE¥¥Microsoft¥¥Windows¥¥…
        "HKEY_LOCAL_MACHINE¥¥SOFTWARE¥¥Microsoft¥¥Windows¥¥…
        "HKEY_CURRENT_USER¥¥Software¥¥Microsoft¥¥Windows¥¥C…
    ],
    "file_opened": [
        "C:¥¥Users¥¥Smith¥¥AppData¥¥Local¥¥Microsoft¥¥Windo…
        "c:¥¥tmp8t2cjf¥¥lib¥¥core¥¥",
        "c:¥¥Users¥¥Public¥¥documents¥¥",
        "c:¥¥tmp8t2cjf¥¥modules¥¥packages¥¥pub.py",
        "C:¥¥Users¥¥Smith¥¥AppData¥¥Roaming¥¥Microsoft¥¥Win…
        "C:¥¥",
        "C:¥¥Users¥¥Smith¥¥AppData¥¥Roaming¥¥Microsoft¥¥Win…
        "c:¥¥PerfLogs¥¥",
        "c:¥¥Users¥¥Smith¥¥Desktop¥¥gndikchadqgjpza.ppt",
```

具体的なデータ項目(behavior)


```
"calls": [
    {
        "tid": 3544,
        "status": 1,
        "time": 1493811728.057385,
        "return_value": 32775,
        "category": "system",
        "stacktrace": [],
        "arguments": {
            "mode": 32769
        },
        "flags": {
            "mode": "SEM_FAILCRITICALERRORS|SEM_NOOPENFILEE…
        },
        "api": "SetErrorMode"
    },
    {
```

データセットに関するアンケート結果

- MWS Cup 2017 の問題として、マルウェア対策研究用データセットに関するアンケートを実施

- 主な回答（太字は複数のチームから挙がった意見）
  - **良性（非マルウェア）のデータセット**
  - **人間に優しい、容易性のあるデータセット、訓練用データセット**
  - **機械学習のツールに適用しやすいデータセット**
  - **仮想環境ではなく実機環境で取得した動的解析ログ**
  - 実際のインシデントレスポンスやフォレンジックで見られるマルウェアのログに近いもの
  - 非 PE マルウェア（ファイルレス、Android, IoT 関連）
  - 通信データ（pcap）

参考意見（人海戦術 Whiteチーム）

必要なのは新規性，明瞭性，容易性の3つ

新規性：新鮮なデータであること

- 【課題】日々刻刻と変化するセキュリティの状況に対応するには，1年前，2年前のデータセットだけでなく，常に最新のデータが望ましい
- 【解決策】データ収集基盤の構築。研究者・企業へのデータ提供の呼びかけ

明瞭性：取り扱い方が明らかなデータであること

- 【課題】データセットによっては提供の意義（どのようにして使ってほしいか）や適用例などが不明瞭である場合がある
- 【解決策】具体的な適用例を提示する（資料にしてデータセットと一緒に配布）
  データセットの使い方講座（解説編，実践編，応用編）を開く

容易性：利用者がデータを研究に活かすための「つなぎ」が用意されていること

- 【課題】データセットの量や質が高い場合でも，実際に研究する際にそのデータセットのフォーマットに適したツールを開発しなくてはならない場合がある
- 【解決策】データ加工ツールの提供
  例：データセットを機械学習(weka etc...)に投げる前の1次加工用プログラム、pcapデータからjsやhtml，セッション情報を取得するプログラムを提供"

FFRI Dataset 2018 の変更点

- **データを動的解析ログから表層解析ログへ**
  - 動的解析ログよりわかりやすいため
  - より新しい大量のデータを提供できるため
  - 良性データ（非マルウェア）のデータも提供できるため
- **良性データも提供**
  - 研究成果を検知率だけでなく誤検知率も評価可能にするため

FFRI Dataset 2018 の概要

- マルウェアおよび良性ファイルのハッシュ値、表層解析データ
- **データ量：マルウェアデータ約29万件、良性データ約21万件(予定)**
- データ形式：CSV, text
- データソース情報
  - マルウェア
    - 2017年に収集した新しい検体
  - 良性ファイル
    - 2008～2017年に収集
      - Windows や Microsoft アプリに含まれるファイル
      - プリインストールされている 3rd パーティソフトウェア
      - Vector で公開されているフリーウェアなど

FFRI Dataset 2018 のデータ項目

- 収集日（マルウェアのみ）
- 各種ハッシュ値
  - MD5, SHA-1, SHA-256
  - ssdeep, imphash, impfuzzy, peHash
- 表層情報
  - アーキテクチャ(32bit/64bit)
  - DLLか否か
  - パッキング有無
  - Anti-Debug 有無
  - プログラム種別(GUI/CUI)
  - PEiD シグネチャ名
  - ファイル種別
  - ヘッダのダンプ

データ取得ツール

- ssdeep
  - ssdeep 2.7 https://ssdeep-project.github.io/ssdeep/
- imphash, ヘッダのダンプ
  - pefile https://github.com/erocarrera/pefile
- impfuzzy
  - impfuzzy https://github.com/JPCERTCC/impfuzzy
- peHash
  - https://github.com/knowmalware/pehash
- アーキテクチャ(32bit/64bit)、DLLか否か、パッキング有無、Anti-Debug有無、プログラム種別(GUI/CUI)、PEiDシグネチャ
  - https://github.com/K-atc/PEiD
- ファイル種別
  - TrID http://mark0.net/soft-trid-e.html

データの例（表層解析データ）

- "md5", "sha1", "sha256", "ssdeep", "imphash", "impfuzzy", "Totalhash", "AnyMaster", "AnyMaster_v1.0.1", "EndGame", "Crits", "peHashNG", "PE", "GUI Program", "Console Program", "DLL", "Packed", "Anti-Debug", "mutex“, “contains base64”, "AntiDebug“,“PEiD”, "TrID“
- dbf246f600c39dd725bf85aca3c25936,d2a4e71ad4820cd85f296ed4b59a30eb8a2d24b8,000011415202841c1eaf64759762a33511461b353f3fd33bc73f0e46fb80c7f7,48:6++Z5YVOeJVkrm1pwbEX7PFUE7aaO0yB+BDq9J5S1XU:6eJVkrmgbCbFUaaayB+FqX5S1k,0ed11c14a2759c69bfa93dd64e1f1654,6:oZ/OWdiVgAaTRLEmDzoVguWqnVWXuRcgtC+5CEp:oZGZGRLhAOteSg75V,8e69e971b3b7895bb9c2e48ee30bae5f9fecb678,4b965799afa33f0d9807d94312e6fe78d0d46f12,a144cb5ad7ec88a8a56f1b55a301e39219083e54,80f7567a918200327b711d92c0b86242,292a243d80b1b1efb89e66d9e555586f88d759e3,62d739591e6aeeecb561c17ee3fec744ebe8de70de6347fc55913c2aa377fe75,32 bit,yes,no,yes,no,no,no,yes,"","",58.9% (.EXE) Win64 Executable (generic) (27625/18/4)¥n14.0% (.DLL) Win32 Dynamic Link Library (generic) (6578/25/2)¥n9.6% (.EXE) Win32 Executable (generic) (4508/7/1)¥n4.4% (.EXE) Win16/32 Executable Delphi generic (2072/23)¥n4.3% (.EXE) OS/2 Executable (generic) (2029/13)

データの例（ヘッダのダンプ）

```
----------Parsing Warnings----------

Byte 0x00 makes up 68.9341% of the file's contents. This may indicate truncation / malformation.

----------DOS_HEADER----------

[IMAGE_DOS_HEADER]
0x0       0x0   e_magic:                    0x5A4D
0x2       0x2   e_cblp:                     0x90
0x4       0x4   e_cp:                       0x3
0x6       0x6   e_crlc:                     0x0
0x8       0x8   e_cparhdr:                  0x4
0xA       0xA   e_minalloc:                 0x0
0xC       0xC   e_maxalloc:                 0xFFFF
0xE       0xE   e_ss:                       0x0
0x10      0x10  e_sp:                       0xB8
0x12      0x12  e_csum:                     0x0
0x14      0x14  e_ip:                       0x0
0x16      0x16  e_cs:                       0x0
0x18      0x18  e_lfarlc:                   0x40
0x1A      0x1A  e_ovno:                     0x0
0x1C      0x1C  e_res:
0x24      0x24  e_oemid:                    0x0
0x26      0x26  e_oeminfo:                  0x0
0x28      0x28  e_res2:
0x3C      0x3C  e_lfanew:                   0x80

----------NT_HEADERS----------

[IMAGE_NT_HEADERS]
0x80      0x0   Signature:                  0x4550

----------FILE_HEADER----------

[IMAGE_FILE_HEADER]
0x84      0x0   Machine:                    0x14C
0x86      0x2   NumberOfSections:           0x6
0x88      0x4   TimeDateStamp:              0x40BDE4FE [Wed Jun  2 14:32:30 2004 UTC]
0x8C      0x8   PointerToSymbolTable:       0x0
0x90      0xC   NumberOfSymbols:           0x0
0x94      0x10  SizeOfOptionalHeader:       0xE0
0x96      0x12  Characteristics:            0x210E
Flags: IMAGE_FILE_32BIT_MACHINE, IMAGE_FILE_DLL,
IMAGE_FILE_EXECUTABLE_IMAGE, IMAGE_FILE_LINE_NUMS_STRIPPED,
IMAGE_FILE_LOCAL_SYMS_STRIPPED


----------OPTIONAL_HEADER----------

[IMAGE_OPTIONAL_HEADER]
0x98      0x0   Magic:                    0x10B
0x9A      0x2   MajorLinkerVersion:        0x2
0x9B      0x3   MinorLinkerVersion:        0x37
0x9C      0x4   SizeOfCode:                0x600
0xA0      0x8   SizeOfInitializedData:     0xC00
0xA4      0xC   SizeOfUninitializedData:   0x0
0xA8      0x10  AddressOfEntryPoint:       0x1190
0xAC      0x14  BaseOfCode:                0x1000
0xB0      0x18  BaseOfData:                0x3000
0xB4      0x1C  ImageBase:                 0x10000000
0xB8      0x20  SectionAlignment:          0x1000
0xBC      0x24  FileAlignment:             0x200
0xC0      0x28  MajorOperatingSystemVersion:   0x1
0xC2      0x2A  MinorOperatingSystemVersion:   0x0
0xC4      0x2C  MajorImageVersion:         0x0
0xC6      0x2E  MinorImageVersion:         0x0
0xC8      0x30  MajorSubsystemVersion:     0x4
...
```

募集（データセットへの意見・要望、共同研究）

- **FFRI Dataset に関するご意見、要望**
  - データを取得したい検体のハッシュや種類
  - 取得したいデータ項目とデータ取得ツールの共有（GitHubなど）
  - その他、どのようにしたら研究に使いやすいか

- **共同研究**
  - 最新の検体データを利用可能（詳細応相談）

- **マルウェア対策技術の研究開発に興味のあるリサーチエンジニア、データサイエンティスト**

MWS の Slack にてお気軽にご連絡ください

ありがとうございました