MWS pre-meeting

FFRI Dataset 2019のご紹介

株式会社ＦＦＲＩ
https://www.ffri.jp

2019/06/04

アジェンダ

提供の目的と現状

これまでのFFRI Dataset

- FFRI Dataset 2013 – 2017
- FFRI Dataset 2018

FFRI Dataset 2019について

FFRI Dataset 2018との相違点

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

- 共同研究、人材交流への発展は少ない
- 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 2018 課題3アンケート）

これまでのFFRI Dataset

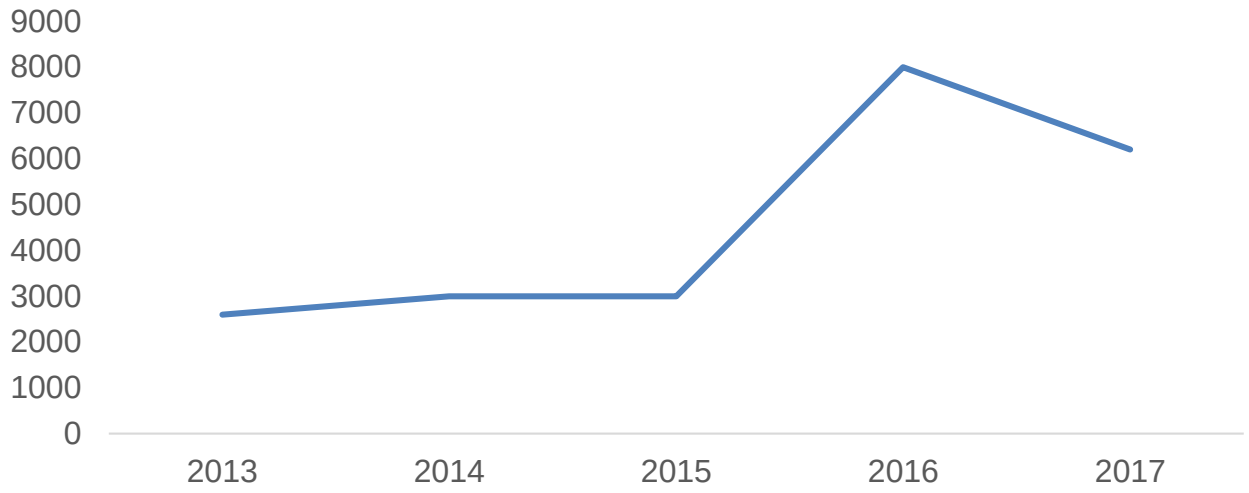
FFRI Dataset 2013-2017

FFRI が収集したマルウェアの動的解析ログ

これまでのFFRI Dataset

FFRI Dataset 2013-2017

累計 約**22,800**検体 平均**4,500**検体/年



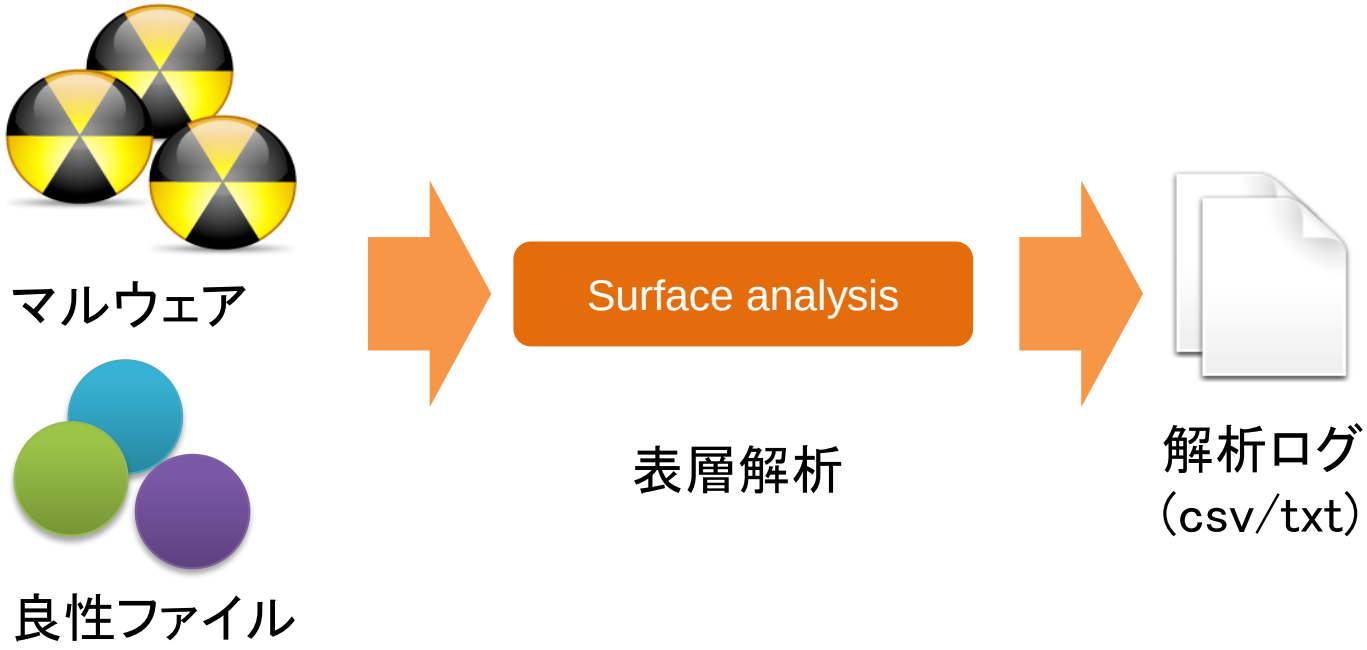
FFRI Dataset 2013-2017

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）
virustotal	VirusTotalから得られる情報
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

これまでのFFRI Dataset

FFRI Dataset 2018

動的解析ログから表層解析ログに変更



FFRI Dataset 2018 の概要

データソース

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

表層情報

- 各種ハッシュ値
- 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

FFRI Dataset 2018

表層解析に変更したことによるメリット

実行時のコンテキストの影響を受けない

- 実行時のコンテキスト: プログラムの動作に影響がある外的要因
  例: C2の稼働状況・解析環境(OS・ソフト・仮想環境) 等

データセットの拡張が容易
分析が比較的容易

FFRI Dataset 2018

表層解析に変更したことによるメリット

良性ファイルについても現実的に提供可能に

解析時間が大幅に削減しデータ数も増加

マルウェア・良性ファイル 計 約 500,000件

FFRI Dataset 2018

反省・指摘

提供形式

- pefileのダンプがtxt形式で機械的にデータを抽出しにくい
- 生成スクリプトを公開したい
- APIを用意してほしい

データの内容

- 文字列データ・ファイルサイズが欲しい

FFRI Dataset 2019

反省・指摘

提供形式

- pefileのダンプがtxt形式で機械的にデータを抽出しにくい
- 生成スクリプトを公開したい
- APIを用意してほしい

FFRI Dataset 2019で対応

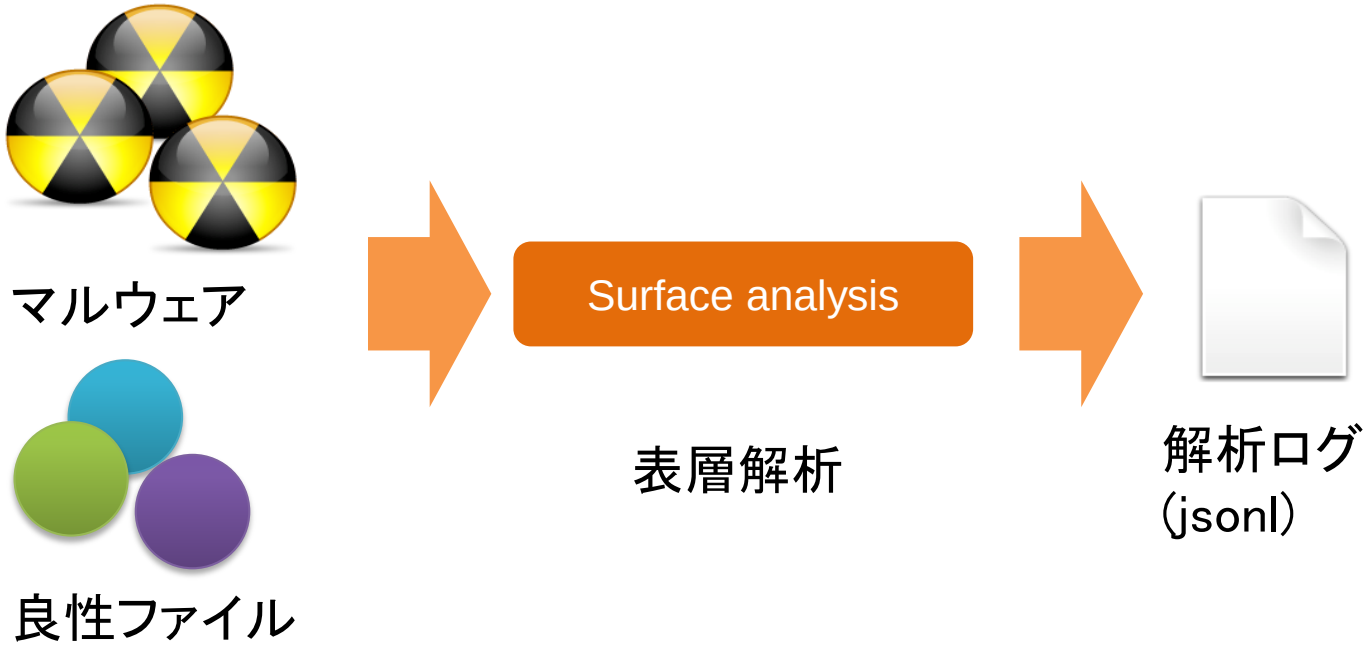
データの内容

- 文字列データ・ファイルサイズが欲しい

FFRI Dataset 2019

昨年に引き続き表層解析情報をご提供

マルウェア・良性ファイル それぞれ25万件 (計50万件)



FFRI Dataset 2019

マルウェア

- 2018-2019年に収集した新しい検体

良性ファイル

- 2018-2019年に AV-TEST社のFLAREから入手した検体
  - OSの構成ファイルや3rdパーティソフトウェアが含まれる

https://www.av-test.org/en/news/endurance-test-do-security-packages-constantly-generate-false-alarms/

FFRI Dataset 2019

項目（大見出し）	内容
id	検体のsha256ハッシュ値
file_size	ファイルサイズ
label	マルウェアか良性ファイルか (マルウェア: 1, 良性ファイル: 0)
date	収集日
hashes	md5, sha1, sha256, ssdeep, impfuzzy, tlsh, anymaster, endgame, crits, pehashngのハッシュ値
peid	PEiDによる表層解析結果
lief	LIEFによる表層解析結果
TrID	TrIDによるファイル種別推定結果
strings	検体中に含まれる文字列情報

FFRI Dataset 2018との相違点

データ数を調整

マルウェア・良性ファイルで均衡データセットに

データレコードの追加

ファイルサイズ/stringsの結果/TLSHを追加

ファイル形式の変更

全部こみこみでJSONL形式に

FFRI Dataset 2018との相違点

生成スクリプトの公開

GitHub上で公開(予定)

表層情報の取得方法の変更

pefileからLIEFへ変更

- JSONへの出力をデフォルトでサポート
- C/C++からも利用可能
- pefileと同等の情報を含む

Library to Instrument Executable Formats

https://lief.quarkslab.com/

FFRI Dataset 2018との相違点

データソース

良性ファイルのソースをFLAREに変更

- FLARE: AV-TEST社の良性ファイルリストサービス
- OSの構成ファイル(kernel32.dll等)から一般ソフトウェアまで

https://www.av-test.org/en/

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

どのようにしたら機械学習を適用しやすいか

欲しいマルウェアのデータ項目

- データ取得ツールを GitHub 等で共有してもらえれば取得可能

データが欲しい検体（ハッシュ、ファイルの種類）

#dataset @oshiba(FFRI)
PR, issue歓迎します！

募集（データセットへの意見・要望、共同研究）

FFRI Dataset を用いた共同研究

毎日新しい検体データを提供可能

- 検体そのものは不可

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

ありがとうございました