FFRI Dataset 2020のご紹介

株式会社ＦＦＲＩ
https://www.ffri.jp

2020/06/08

アジェンダ

ご紹介

スペック

昨年度のとの差異

ご紹介

FFRI Dataset

例年マルウェアのサンプル毎の解析データをご提供

FFRI Dataset 2013 - 2017

- 動的解析ログ (Cuckoo Sandbox等)

FFRI Dataset 2018, 2019

- 表層解析ログ

FFRI Dataset 2020

昨年に続き**表層解析ログをご提供**

6月中旬にリリース予定

FFRI Dataset 2020

表層解析ログの強み

再現性・拡張性が高い

- 同条件で解析すれば同じ結果が得られる
- 解析条件の再現が容易

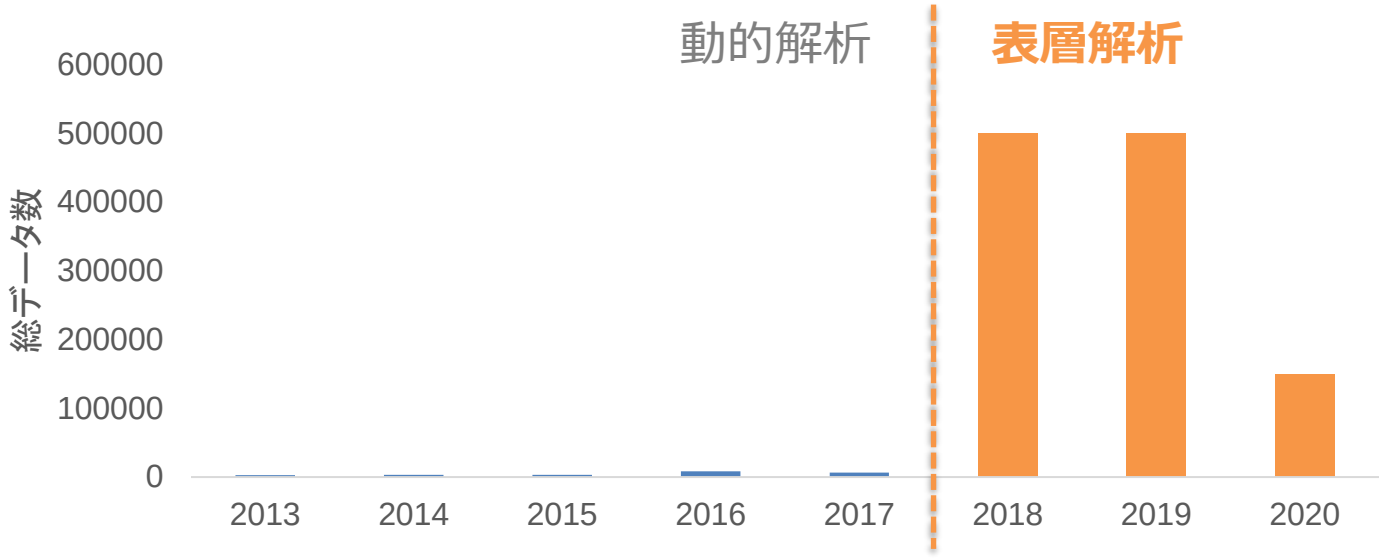
- **取得に利用したスクリプトは公開**
  - https://github.com/ffri/ffridataset-scripts

FFRI Dataset 2020

表層解析ログの強み

より多くのデータをご提供可能

- 1件当たりの解析時間が比較的短い



活用イメージ

主にマルウェア分類の研究に利用されることを想定

表層解析は検知タイミングの中で最速

- 動作前に解析可能かつ基本的には軽量な解析手法
  - 被害が出る前に検知できる (動的解析では被害が生じ得る)
  - (標的環境か否かに影響を受けない)

スペック

FFRI Dataset 2020

データソース

マルウェア

- 2019/01/01–2019/12/31に収集した公知のマルウェアのうち
  PE形式をとっているもの

良性ファイル

- AV-TEST社の良性ファイル提供サービスFLAREにより
  2019/01/01–2019/12/31に提供されたファイルのうち
  PE形式をとっているもの

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

FFRI Dataset 2020

表層解析ログ

良性ファイル/マルウェア 各7.5万件を予定

- 昨年度は各25万件

1行1データのjsonl形式

- 昨年度と同形式

FFRI Dataset 2020

要素	概要
id	検体のSHA256ハッシュ値
file_size	ファイルサイズ
label	ラベル(1=マルウェア, 0=良性ファイル)
date	収集日
hashes	各種ハッシュ値
peid	pypeidの出力
trid	TrIDの出力
strings	stringsの出力
lief	LIEFの出力

https://github.com/FFRI/pypeid

http://mark0.net/softtrid-e.html

https://www.gnu.org/software/binutils/

https://lief.quarkslab.com/

利用ツールバージョン

ffridataset-scripts(v2020.1)

https://github.com/FFRI/ffridataset-scripts/releases/tag/v2020.1

利用ツールバージョン(詳細)

item	バージョン	url
ssdeep	3.4	*https://ssdeepproject.github.io/ssdeep/*
tlsh	4.2.1	*https://github.com/trendmicro/tlsh*
pehash	0.9.1	*https://github.com/knowmalware/pehash*
impfuzzy	0.5	*https://github.com/JPCERTCC/impfuzzy*
pypeid	0.1.0	*https://github.com/FFRI/pypeid*
TrID	4.2.1	*http://mark0.net/softtrid-e.html*
LIEF	0.11.0(forked)	*https://github.com/kohnakagawa/LIEF/releases/tag/0.11.0.ffridataset2020*
strings	2.30	*https://www.gnu.org/software/binutils/*
pefile	2019.4.18	*https://github.com/erocarrera/pefile*

pypeidについてはpythonによる再実装したものを利用(処理は既存のものと同様)
LIEFについてはリリース未反映のためforkしたものを流用

昨年度との差異

昨年度との差異

改善の方針

解析項目・フォーマットの変更は軽微なものに留め，
データセットの全体的な質の向上を目指した

- 利用ライブラリ等によって混入する暗黙的なバイアス除去
- ユーザビリティの向上

昨年度との差異

フォーマットの変更

良性ファイルにも収集日フィールドを入れるように

liefにおいて基本的にはrawデータをご提供

- パースに成功した場合のみ意味データもご提供
- これまではすべて意味抽出しようとしていたため失敗すると即死

昨年度との差異

ユーザビリティの向上

JSON Schemaをご用意

- データセットはこのSchemaでバリデーション
- 各フィールドの意味もdescriptionにてご提供

昨年度との差異

暗黙的なバイアスの除去

暗黙的に除外されていた検体

- 利用しているライブラリのクラッシュ等による除外
  - これはライブラリ・OSSを修正する方向で対応

- **データセット作成時のタイムアウトによる除外**
  - 間接的にサイズの大きなファイルが除外される状態に
  - こういった制限ができるだけかからないようにスクリプトを修正

注意点

昨年度のデータセットと純粋な比較ができない

昨年度混入していたバイアスの除去に努めた

- これは**時間経過によって生じる変化とは無関係**

~~x「2019年にはサイズが大きなマルウェアが増えた」~~

o「サイズの大きなマルウェアが解析対象になった」

程度問題かもしれないが，今年度は特に意識されたい

さいごに

募集

FFRI Datasetに関するご意見・ご要望

「こういう情報も取ってほしい」

「こうして提供してくれれば使いやすい」などなど

Slack:

- #dataset / @oshiba(FFRI)

GitHubでのPR/issueも大歓迎です！

- https://github.com/FFRI/ffridataset-scripts

ありがとうございました