FFRI Dataset 2021のご紹介

株式会社ＦＦＲＩセキュリティ
（東証マザーズ：3692）
https://www.ffri.jp/

2021-06-02

FFRI Datsetについて

FFRI Dataset

例年マルウェアの解析データをご提供

2013 – 2017

- 動的解析ログ(yarai Analyzer, Cuckoo Sandbox等)

2018 - 2020

- **表層解析ログ**

FFRI Dataset 2021

今年度も昨年度に続き、**表層解析ログ**をご提供

既にご提供を開始しています

FFRI Datasetの特徴

1. マルウェアだけでなくクリーンウェアのデータもご提供

2. 解析の再現性及び、データセットの拡張性の高さ

広く利用される手法を利用しスクリプトも公開することで再現性を担保

- 表層解析は特に動的解析と比べて再現性が高い

再現性のある手法によって解析することでデータセットの拡張性が向上

- データセットに含まれていない対象についても同等の条件で解析可能

3. データ件数の多さ

FFRI Dataset 2018以降は例年15万件以上ご提供

機械学習を利用した研究にご活用いただくことを想定

スペック

概要

項目	概要
データ形式	jsonl形式 (1行1検体分のレコード)
件数	150,000件 (マルウェア・クリーンウェア)
展開前サイズ	約20GB
展開後サイズ	約118GB

データソース

種別	ソース	期間(収集日)	検体数
マルウェア	弊社が収集したマルウェアのうちPE形式のものから無差別に抽出	2020/1/1 – 2020/12/31	75000
クリーンウェア	AV-TEST社のクリーンウェア提供サービスFLARE※により提供されたファイルのうちPE形式のものから無差別に抽出	2020/1/1 – 2020/12/31	75000

※ https://www.av-test.org/en/

レコード要素

要素	概要
id	検体のSHA-256ハッシュ値
label	ラベル(1=マルウェア, 0=良性ファイル)
date	収集日
file_size	ファイルサイズ
trid	TrIDの出力
strings	stringsの出力
hashes	各種ハッシュ値
peid	pypeidの出力
lief	LIEFの出力
version	データセットのバージョン
die	DIEの出力
manalyze_plugin_packer	ManalyzeのPackerプラグインの出力

※FFRI Dataset 2021で新規に追加した項目は橙色背景で表示

バリデーション

データセットはjson schemaによってバリデーション済み

利用したjson schemaはGitHub上で公開

- 各フィールドに入り得る値を確認したい場合はこちらを参考にすると良い
- 各フィールドの簡易な説明も記載

ffridataset-scripts/ffridataset_v2021.json at master · FFRI/ffridataset-scripts –
https://github.com/FFRI/ffridataset-scripts/blob/master/schema/ffridataset_v2021.json

利用ツール一覧

item	version	url
ssdeep	3.4	https://github.com/ssdeep-project/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	v2.24	https://mark0.net/soft-trid-e.html
LIEF	Forked	https://github.com/kohnakagawa/LIEF/tree/ae515b687841687367728dff8d178a6f9abfa931
strings	2.34	https://www.gnu.org/software/binutils/
pefile	2019.4.18	https://github.com/erocarrera/pefile
Manalyze	commit hash 04cee36	https://github.com/JusticeRage/Manalyze
Detect-It-Easy (DIE)	3.01	https://github.com/horsicq/DIE-engine

※FFRI Dataset 2021で追加・更新した項目は橙色背景で表示

生成スクリプト

データセット生成に利用したスクリプトは公開中

- ffridataset-scripts(v2021.1)
  - Release v2021.1 · FFRI/ffridataset-scripts (github.com)

お手元の検体によるデータセットの拡張等にご活用いただくことを想定

昨年度との差異

昨年度との差異

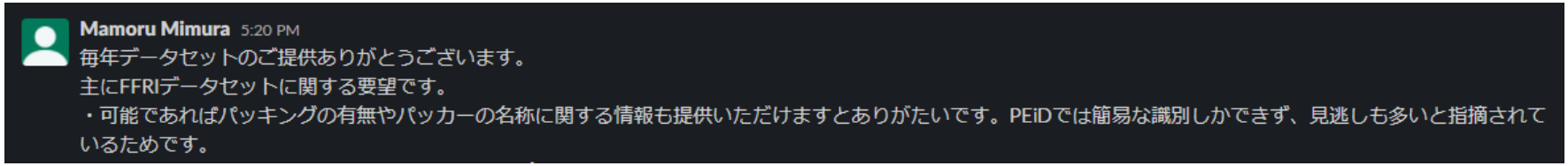
1. パッカー推定・検知ツールによる解析を追加

2. LIEFのバージョンをアップデート

3. 専用特徴抽出ライブラリのご提供

パッカー検知・推定ツールによる解析

#datasetにてパッカーの検知・推定ツールについてご要望を頂いた



検討の結果、**ManalyzeとDIEを導入**

- 検討過程についても公開
  - FFRI/PackerDetectorConsideration: Consideration of packer detection tool for FFRI Dataset scripts
    https://github.com/FFRI/PackerDetectorConsideration
  - FFRI/PackerDetectionToolEvaluation: Evaluation of packer type estimation/detection tools -
    https://github.com/FFRI/PackerDetectionToolEvaluation

LIEFのバージョンアップ

LIEFのバージョンを0.11を改変したものを利用

改善取得可能な情報が増加 (下記一例)

- デジタル署名関連のパース処理の改善
  - 副署名が取得できるようになった等
- リソースセクションのパース処理の改善
  - htmlも抽出可能となった等

例えばマルウェア特徴の経時変化等昨年度のデータセットと比較を行う場合には
こちらの変更の影響を受けていないか、注意深く検討されたい

(PE関連のみ抜粋)Difference of FFRI Dataset between 2021 and 2020 (github.com)

(Full diff)Comparing dev/ffri-dataset...dev/ffridataset_2021 · kohnakagawa/LIEF - https://github.com/kohnakagawa/LIEF/compare/dev/ffri-dataset...dev/ffridataset_2021

専用特徴抽出ライブラリのご提供

FFRI Dataset専用の特徴抽出ライブラリ「FEXRD」を公開

前処理等について詳しくなくても気軽に研究を開始できるように

```
import json
from fexrd import StringsFeatureExtractor

sfe = StringsFeatureExtractor() # instantiae feature extractor class for the "string" element
fin = open("ffridataset_sample.jsonl", "r")
for l in fin.readlines():
    obj = json.loads(l)
    column_names, vector = sfe.get_features(obj["strings"]) # convert to the vector
```

実は昨年より公開しているが、前回はMWS Cupのタイミングでご提供となった

- 開発時期の関係で研究向けには間に合わなかった
- ついでにドキュメントもあまり整備できていなかった

今年はFFRI Dataset 2021に対応し公開済み

- 合わせてドキュメントも拡充

FFRI/FEXRD: Feature Extractor for FFRI Dataset - https://github.com/FFRI/FEXRD

ご意見・ご要望

FFRI Datasetに関するご意見・ご要望はお気軽に！

- 「こういう情報も取ってほしい」
- 「この検体の情報も欲しい」
- 等

Slack:

- #dataset
- @oshiba(FFRI)

もちろんGitHubでのPR/issueの投稿も大歓迎です！

- FFRI/ffridataset-scripts: Make datasets like FFRI Dataset – https://github.com/FFRI/ffridataset-scripts