FFRI Dataset 2022のご紹介

株式会社ＦＦＲＩセキュリティ
https://www.ffri.jp

FFRI Datasetについて

概要と特徴

FFRI Datasetは動的・表層解析データを提供

例年マルウェアの解析データをご提供しています

	2013 - 2017	2018 - 2021	2022
データの種類	動的解析	表層解析	本資料で解説
使用ツール	yarai Analyzer Cuckoo Sandbox 等	LIEF pehash 等	
特徴	動かさないと分からない情報を提供可能	再現性が高い 大量の件数のデータを提供可能	

FFRI Dataset 2022の特徴

FFRI Dataset 2022は表層解析データセットです

特徴	概要
提供の種類	マルウェアとクリーンウェアの両方のデータを提供
再現性と拡張性	一般的なツール・手法を使用 データセットの作成スクリプトを公開
データ量	マルウェア・クリーンウェア合わせて15万件を提供

FFRI Dataset 2022は”ML-ready”なデータセット

FFRI Dataset 2022

スペック

FFRI Dataset 2022の概要

諸事情でクリーンウェアの4~6月中旬が抜けています

	クリーンウェア	マルウェア
データの形式	JSON Lines（1行1検体分のレコード）	
件数	75000件	75000件
展開前サイズ	約26GB	
展開後サイズ	約97GB	約37GB
期間（収集日）	2021/01/01-2021/03/31, 2021/06/20-2021/12/31	2021/01/01-2021/12/31

FFRI Dataset 2022の概要（続き）

	データソース
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
マルウェア	弊社が収集したマルウェアのうち PE形式のものを無作為に抽出

※ AV-TEST | Antivirus & Security Software & AntiMalware Reviews

FFRI Dataset 2022の特徴

FFRI Dataset 2021と要素は同じ

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

FFRI Dataset 2022の特徴（続き）

使用したツール・ライブラリは以下の通り

ツール・ライブラリ	version	URL
ssdeep	3.4	https://pypi.org/project/ssdeep/
TLSH	4.7.2	https://pypi.org/project/py-tlsh/
pehash	0.91	https://github.com/knowmalware/pehash
impfuzzy	0.5	https://pypi.org/project/pyimpfuzzy/
LIEF	0.12.0	https://pypi.org/project/lief/
TrID	2.24	https://mark0.net/soft-trid-e.html
strings	2.34	https://www.gnu.org/software/binutils/
pypeid	0.1.1（※1）	https://github.com/FFRI/pypeid
pefile	2021.9.3	https://pypi.org/project/pefile/
Manalyze	6397357（※2）	https://github.com/JusticeRage/Manalyze
Detect-It-Easy	modified（※3）	https://github.com/horsicq/DIE-engine

※1 ffridataset-scriptsのpypoject.tomlでは0.1.0となっているが、実際は0.1.1と同じものを使用している。
データセットの作成に使用したバイナリをリポジトリで提供している。
※2 これはcommit hash
※3 バージョン3.04を修正したものを使用している。使用したバイナリをリポジトリで提供している。

FFRI Dataset 2022に関連するOSS

OSSによりデータセットの作成・拡張・利用を促進

	作成・拡張	利用
OSS名	ffridataset-scripts Release v2022.1 · FFRI/ffridataset-scripts (github.com)	FEXRD Release v2022.1 · FFRI/FEXRD (github.com)
概要	FFRI Datasetの作成に用いたスクリプト	FFRI Dataset（と同形式）のデータから特徴量を抽出できるライブラリ
想定するユースケース（一例）	FFRI Datasetにない検体から同じ形式のデータを抽出し、Concept DriftやDomain Shiftの研究に用いる	FFRI Dataset 2022を用いた機械学習研究のベースラインに用いる

昨年度との差異

注意点

FFRI Dataset 2021との差異

ライブラリ・ツールのアップデートによる差異がある

	ライブラリ・ツール	変更点
スキーマ変更なし	pypeid	依存関係の更新
	TrID	定義ファイルの更新
	pefile	アップデート
	Manalyze	アップデート
スキーマ変更あり・詳細は次ページ	LIEF	バージョンアップによりレコード追加
	Detect-It-Easy	出力形式の変更
	TLSH	バージョンプレフィックスの付与

スキーマ変更の詳細及び注意点

昨年度以前のデータセットとの比較の際は以下に注意し検討のこと

ライブラリ・ツール	主な変更点・注意点
LIEF	• バージョン0.12.0への更新に伴い、Delay Importsへ対応 • Data Directoriesのtype、TLSのData Directoryに”RESERVED”が追加 • ffridataset-scriptsでは独自ビルドのバイナリは削除し、PyPIからインストールするように変更した
Detect-It-Easy	• 2021のときは{arch: hoge, detects:[…], …}だったのが、{detects: [filetype: foo, …]}のように出力形式が変更 • バージョン出力をUTF-8に対応した修正版ビルドのバイナリをffridataset-scriptsで提供
TLSH	• バージョンプレフィックスとして、「T1」がハッシュ値の先頭につくように変更 • ただし提供しているJSON Schemaではこれまでも考慮していた

おわりに

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」
Slack #dataset @ko.nakagawa

OSSへのコントリビューション
（issue/pull request）も大歓迎です！！

Issues · FFRI/FEXRD (github.com)
Issues · FFRI/ffridataset-scripts (github.com)
Issues · FFRI/pypeid (github.com)