FFRI Dataset 2023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 - 2022	2023
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 2023の特徴

FFRI Dataset 2023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 2023は”ML-ready”なデータセット

FFRI Dataset 2023

スペック

FFRI Dataset 2023の概要

	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33GB	
展開後サイズ	約125GB	約44GB
期間（収集日）	2022/01/01- 2022/12/31	

FFRI Dataset 2023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 2023の特徴

FFRI Dataset 2022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 2023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	b30548d（※1）	https://github.com/JPCERTCC/impfuzzy
LIEF	0.12.3	https://pypi.org/project/lief/
TrID	2.24	https://mark0.net/soft-trid-e.html
strings	2.38	https://www.gnu.org/software/binutils/
pypeid	0.1.2	https://github.com/FFRI/pypeid
pefile	593d094（※1）	https://pypi.org/project/pefile/
Manalyze	e951f34（※1）	https://github.com/JusticeRage/Manalyze
Detect-It-Easy	3.07	https://github.com/horsicq/DIE-engine

※1 これはshort commit hash

FFRI Dataset 2023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 https://github.com/FFRI/ffridataset-scripts/releases/tag/v2023.1	FEXRD https://github.com/FFRI/FEXRD/releases/tag/v2023.1
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 2023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 2022との差異

収集期間の日付を除いてスキーマの変更はなし

	ライブラリ・ツール	変更点
スキーマ変更なし	TrID	定義ファイルの更新
	impfuzzy	アップデート
	LIEF	アップデート
	strings	アップデート
	pypeid	アップデート
	pefile	アップデート
	Manalyze	アップデート
	Detect-It-Easy	アップデート

FFRI Datasetを使用した論文紹介

利用事例

FFRI Dataset 2013~2017を使用した論文

ここに載せた論文はごく一部です

	Automatically generating malware analysis reports using sandbox logs	Malware function classification using apis in initial behavior
使用データセット	2013~2015	2014
概要	Sandboxのログとベンダーのレポートからhuman readableなレポートを生成	呼ばれたAPIからマルウェアの機能を推定
Cite	Sun, Bo, et al. "Automatically generating malware analysis reports using sandbox logs." IEICE TRANSACTIONS on Information and Systems 101.11 (2018): 2622-2632.	Kawaguchi, Naoto, and Kazumasa Omote. "Malware function classification using apis in initial behavior." 2015 10th Asia Joint Conference on Information Security. IEEE, 2015.

FFRI Dataset 2018~2022を使用した論文

ここに載せた論文はごく一部です

	Evaluation of printable character-based malicious PE file-detection method	Robust detection model for portable execution malware
使用データセット	2019~2021	2018
概要	文字列を用いてマルウェアとクリーンウェアを分類。時系列的な影響も調査	次元削減により Adversarial Attack に対しRobustな分類器を作成
Cite	Mimura, Mamoru. "Evaluation of printable character-based malicious PE file-detection method." Internet of Things 19 (2022): 10052	Zheng, Wanjia, and Kazumasa Omote. "Robust detection model for portable execution malware." ICC 2021-IEEE International Conference on Communications. IEEE, 2021.

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)