Vol.2019-CSEC-86 No.8
Vol.2019-SPT-34 No.8
2019/7/23

マルウェア対策のための研究用データセット ～MWS Datasets 2019～

荒木 粧子[1,a)]　笠間 貴弘[2]　押場 博光[3]　千葉 大紀[4]　畑田 充弘[5]　寺田 真敏[6,7]

概要：
多くの研究者によってマルウェアの検知，解析をはじめとした様々な対策手法が提案されているが，その一方で，マルウェアを用いた攻撃の多様化や高度化により，研究を進める上で基礎となる“研究素材”の継続的な収集と維持が難しくなってきている．本報告では，研究成果の客観的な評価と活用を研究課題として取り組んでいるデータセット MWS Datasets 2019 について概説する．

キーワード：データセット，マルウェア，MWS Datasets，BOS，CCC DATAset，D3M Dataset，FFRI Dataset，NICTER Dataset，PRACTICE Dataset，Soliton Dataset

Datasets for Anti-Malware Research ～MWS Datasets 2019～

1. はじめに

高度化および複雑化が進むサイバー攻撃は世界的な脅威となっており，各組織による対策はもちろん，国家や多国間連携による対策が急務となっている．特に，マルウェアに起因するサイバー攻撃は様々な社会問題を引き起こすことから，マルウェア対策やそこから派生する様々な対策研究が盛んに行われている．しかしながら，“共通の研究素材がないこと”および“研究素材の収集が困難であること”が近年のマルウェア対策研究を推進する上で阻害要因となっている．

一つ目の阻害要因である共通の研究素材とは，研究開発した技術の評価に用いるマルウェア，マルウェアによるスキャンや不正送金等に関わる一連の攻撃通信データ，マルウェア感染後の通信データ，標的型攻撃などを指し，可能な限り網羅的，かつ攻撃の進化に合わせて適切に選択されたデータが望ましい．従来では研究素材となるデータは，主に研究者自らが収集環境を構築して収集し，個々の技術の有効性や妥当性評価に使用してきた．すなわち，同じ研究テーマに取り組んだとしても研究素材が異なるため，研究結果を相互に比較し適切に評価することが困難であった．

二つ目の阻害要因は，研究素材の収集自体が困難になってきていることである．検知回避手法や解析妨害手法を用いた攻撃や，それらが年々高度化していることが，研究素材の収集を困難にさせる．例えば，ドライブバイダウンロード攻撃を仕掛ける悪性ウェブサイトは解析および検知を回避する様々な機能を有しており，情報を収集する環境によっては期待した情報を取得できない．また，ボットの C&C サーバとの通信を収集する場合においても，近年の C&C サーバは短期間で活動を停止するため，期待した通信データを継続的に収集することは困難である．さらに標的型攻撃においては，攻撃者の標的組織となり，侵入された後に組織内でどのように振る舞うか等の挙動を逐次記録，保全する必要がある．これらを研究者自らが収集することは非常に困難である．

複雑化の一途をたどるサイバー攻撃に対峙していくため，我々はマルウェア対策研究コミュニティである anti Malware engineering WorkShop (MWS) を組織した．MWS

[1] 株式会社ソリトンシステムズ
Soliton Systems K.K.
[2] 国立研究開発法人情報通信研究機構
[3] 株式会社 FFRI
[4] NTT セキュアプラットフォーム研究所
[5] 日本電信電話株式会社
[6] 東京電機大学
[7] 株式会社日立製作所
[a)] shoko.araki@soliton.co.jp

図 1 マルウェア対策研究のサイクル

は図 1 に示すとおり「研究用データセットの提供」,「分析ならびに対策技術の研究」,「研究成果の共有」というマルウェア対策研究のサイクルを継続的に回すことで，マルウェア対策研究活動を推進してきた．具体的な活動として，本コミュニティ内で研究用データセットを共有することで研究を促進し，また研究成果を共有する場として「マルウェア対策研究人材育成ワークショップ (MWS)」を 2008 年から毎年開催してきた [1](2019 年は MWS2019 [2] を開催する予定)．さらなる研究発展のため，研究用データセットの作成そのものが研究対象分野として立ち上がり，より活発に研究サイクルが回るよう後押しする活動を展開していきたいと考えている．

本稿では，MWS の活動の一環で作成した研究用データセット MWS Datasets 2019 (図 2) について報告する．2019 年は下記のデータセットから構成される．

(1) **BOS 2019** — 標的型攻撃の観測データ (§3.1)
(2) **FFRI Dataset 2019** — マルウェアと良性ファイルの表層解析データ (§3.2)
(3) **NICTER Dataset 2019** — ダークネットにおけるパケットデータ (§3.3)
(4) **Soliton Dataset 2019** — マルウェアの動的解析データ (§3.4)
(5) **MWS Cup Dataset** — MWS Datasets を活用した競技 MWS Cup [3] の参加チームが作成したデータ (§3.5)

なお，CCC DATAset，D3M Dataset，PRACTICE Dataset 2013, 2015 はデータセットの内容に更新がないが，MWS Dataset 2019 に含めて継続的に提供する．これらデータセットの詳細は，文献 [4], [5], [6], [7], [8], [9], [10], [11] を参照して欲しい．

2. 関連研究

本章では関連研究として他のデータセットや研究コミュニティを紹介する．

2.1 研究用データセット

非商用のうち，代表的なセキュリティに関連する研究用データセットの例を表 1 に示す．ここでは，2015 年に国内で発生した公的機関へのサイバー攻撃，2016 年に流布した IoT 機器を狙うマルウェア Mirai によるサイバー攻撃などの動向を踏まえ，5 年以内に更新された研究用データセットとそれ以前とで大別した．数年以上更新されていない研究用データセットの多くは，脅威状況の変化など時間軸の影響を受けている．その一方で，産業用制御システムを対象とした研究用データセットなど，これまで取り扱ってこなかった分野でのデータセット作成が広がりつつある．

表 1 非商用の研究用データセット

区分	データセット名
2014 年以前に作成，現在は更新なし	DARPA Intrusion Detection [12] MALICIA Project [13] Android Malware Genome Project [14] CTU-13 DATASET [15]
2015 年以降に作成，あるいは，2015 年以降も更新	Kyoto 2016 Dataset [16] CAIDA Data [17] IMPACT [18] MAWILab [19] Malware-Traffic-Analysis.net [20] Contagio Malware Dump [21] EMBER [22] Microsoft Malware Predoction [23] CSE-CIC-IDS2018 [24] MaxNet [25] MODBUS ICS DATASET [26]

2.2 研究用データセットの課題

本節では広くマルウェア対策研究を推進するにあたり，研究用データセットの活用を促進させる上での問題点について考察する．

2.2.1 データセット入手の容易性

多くのデータセット共有コミュニティにおいて，データセットを入手するためにはコミュニティへの加入が必要であり，加入の際に契約締結もしくは審査が行われる．政府がスポンサーとなっているコミュニティや地域性の高いコミュニティが多く，例えば IMPACT は米国の政府（国土安全保障省，DHS）や米国の大学が主体となり，iSecLab [27] は欧州の大学やセキュリティ研究所および企業が主体となっている．このようなコミュニティに対して，日本の学術機関や企業が単独で加入しデータセットを入手するためには，多大なコミュニケーションコストを必要とする．一方，MWS は日本の学術機関や企業を中心とするため，MWS コミュニティへの参加は容易であり，かつ参加継続も容易に行えるよう配慮している．今後はコミュニティ間で連携を計ることにより，相互に研究用データセットの共有を行うことが MWS に求められる．

2.2.2 データセットの継続性

通信形態やプラットフォームの変化に伴い，サイバー攻撃やマルウェア感染手法は日々進化するため，研究用データセットには数年にわたる継続性が求められる．しかし，研究用データセットに継続性がない場合，すなわちデータセットの更新がなく最新の傾向を反映できていない場合，研究用途としての活用は難しい．例えば，DARPA Intrusion Detection Data Sets は 1998 年から 2000 年までに作成され

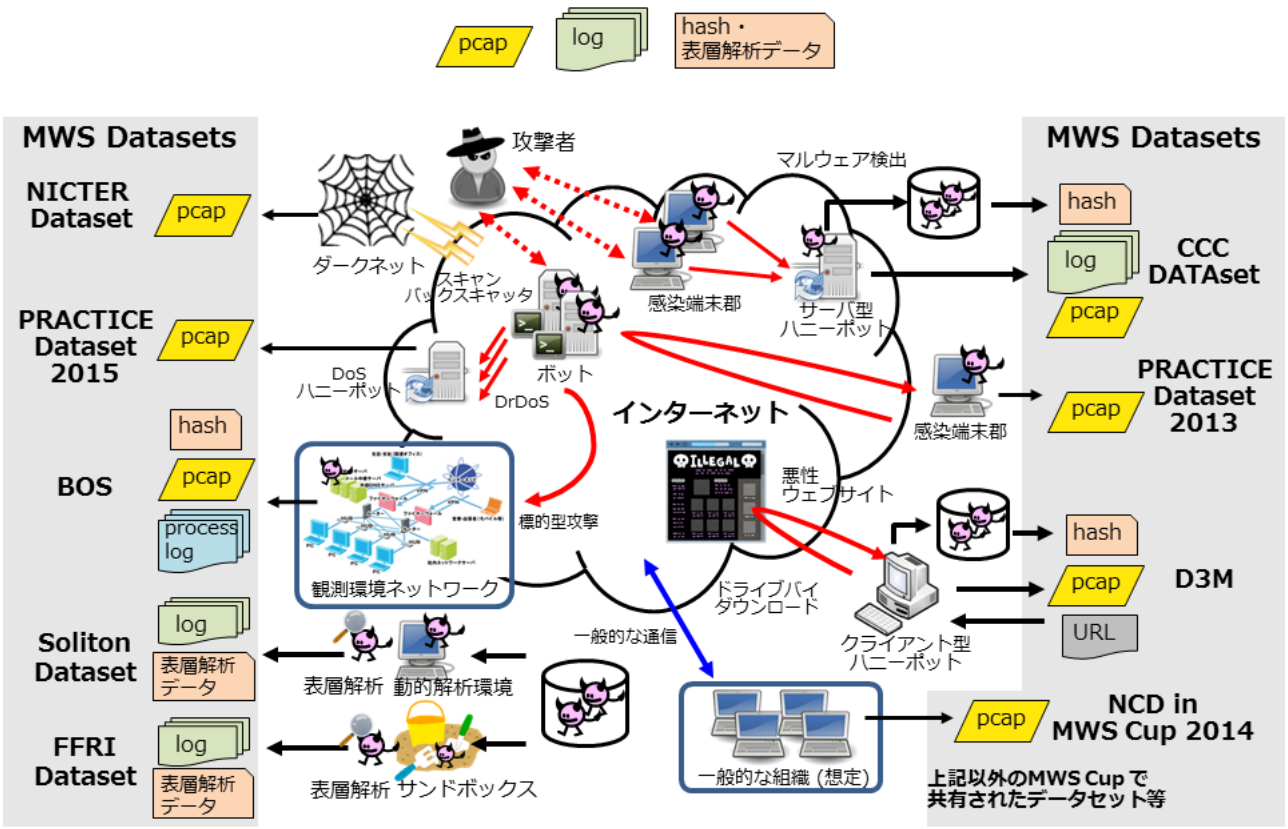

図 2 MWS Datasets 2019 の概要

た IDS のトラヒックデータセット，CTU-13 DATASET は 2011 年に収集されたボットネット通信のデータセットであるが，いずれも更新されていない．また，Android Malware Genome Project Dataset および MALICIA Dataset はリソースの制限や担当者の所属変更によりそれぞれ 2015 年，2016 年に提供を停止している．データセットの継続性を担保するためには，収集環境の整備とデータ作成者へのインセンティブが必要である．MWS でも同様に，個々のデータセット提供者の収集環境に依存してデータセットの更新や共有の停止が発生することがあるため，コミュニティとしてデータセットの継続性を担保するための仕組みを検討および運用する必要がある．

2.2.3 データセットの網羅性

多種多様なサイバー攻撃に対して多角的かつ全域的な分析を実施するためには，データセットの種類および観測点の網羅性が求められる．CAIDA Data や IMPACT Dataset は様々な組織で収集した数十種類のデータセットを提供することでデータセットの種類と観測点の網羅性を向上させている．MWS はマルウェアに着目し，感染前活動，感染時，感染後の各データセットを提供しており，昨今のサイバー攻撃を広く網羅していると言える．観測点の網羅性については，さらにデータセット提供者やデータセット取得環境を増やすことで向上させたい．また，一部のデータセットに関しては，研究に必要十分なデータ容量を提供できていないものも存在するため，これらについても今後検討する必要がある．

3. MWS Datasets 2019

3.1 BOS_2019

動的活動観測 BOS (Behavior Observable System) データセットは，組織内ネットワークへの侵害活動を想定した研究用データセットであり，総務省「サイバー攻撃解析・防御モデル実践演習の実証実験」，国立研究開発法人 情報通信研究機構「実践的サイバー防御演習シナリオ・環境構築支援」で得られた成果の一部である [28].

3.1.1 データセット提供背景

マルウェア検体の静的／動的解析では，マルウェアの挙動に着目した解析であり，攻撃者の行動という視点で把握や解析することは少なかった．多くの場合，攻撃者の行動＝マルウェアの挙動という想定の下，静的／動的解析によって対応してきた．しかし，組織内ネットワークへの侵害活動においては，攻撃者の存在を意識する必要がある．そこで，BOS では，マルウェアの挙動に加えて，どのような操作をしたのか，どのようなファイルにアクセスしたのかなど攻撃者の行動と組み合わせていくことで，攻撃者行動視点で脅威の特徴付けを試みる研究用データセットとなっている．

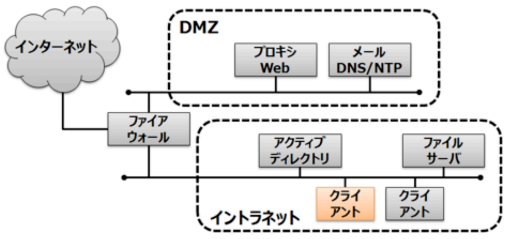

図 3 動的活動観測環境の概要

図 4 BOS_2019 の観測事例 (抜粋)

No.	観測期間		マルウェア検体名	進行度
	開始	終了		
h01	2018/09/03	2018/09/09	BC0WHT18	5
h02	2018/09/19	2018/09/26	ANELDLDR.ZKFH-A	5
h03	2018/10/29	2018/11/05	SelfDel.glou	5
h05	2018/11/07	2018/11/15	SNOWBALL.ZLFK-A	5
h07	2018/11/16	2018/11/23	SMOKELOAD.AB	2
h09	2018/12/07	2018/12/10	PLEAD.SMZTEG	1
h11	2019/01/17	2019/01/24	DLOADR.AUSUON	5
h13	2019/02/01	2019/02/01	DLOADR.AUSUON	5
h15	2019/02/22	2019/03/01	MSExcel.DdeExec.b	5

表 2 動的活動観測における進行度

進行度	区分		概要
1	通信発生なし		検体実行不可，もしくは マルウェアではない
2			検体実行するも，通信発生なし
3	通信発生 あり	C2 サーバ と攻撃通信 成立せず	C2 サーバの名前解決不可
4			C2 サーバへの SYN パケット 送信のみ
5			C2 サーバとの通信不可 (HTTP ステータスコード 403, 404, 503 等)
6		C2 サーバ と攻撃通信 成立	攻撃 (活動/操作) を観測できず
7			攻撃 (活動/操作) を観測できた
8			攻撃 (活動/操作) を継続的に 観測できた

3.1.2 動的活動観測環境

動的活動観測環境は，小規模な遠隔拠点の情報システムを模擬するハニーポットである (図 3)．本環境は，組織内ネットワークのパソコンにおいてマルウェア感染が発生した以降を対象に，実インターネット上の攻撃者が組織内ネットワークで試みるサイバー攻撃活動を観測するシステムとなっている．クライアントは，標的型攻撃メールに添付されたマルウェア検体を実行するパソコンであり，プロキシ経由／プロキシ経由なしのいずれかの形態で，実インターネットへのアクセスが可能である．

3.1.3 データセット構成

BOS_2016 以降，表 2 に示す進行度という動的活動観測における侵害活動の進み具合の区分を設け，標的型攻撃の段階に応じた研究用データセットとなるよう工夫を施している．BOS_2019 は，表 4 に示す観測事例に関連する観測データを含んでいる．

(1) **マルウェア検体** 動的活動観測に使用したマルウェア検体のハッシュ値を STIX (Structured Threat Information eXpression)[*1] 形式で記載した XML ファイルである．

(2) **通信観測データ** マルウェア検体を実行した際の通信のフルキャプチャデータ，ファイアウォールログ，プロキシサーバログ等である．

(3) **プロセス観測データ** マルウェア検体を実行したクライアントでのプロセスの稼働状況を記録したデータや Windows[*2] イベントログである．

(4) **不正接続先観測データ** 動的活動観測で対象とした不正接続先の稼動状況に関する 2018 年分のデータである．動的活動観測のデータセット対象範囲を広げる試みとして BOS_2019 に含めた．

3.2 FFRI Dataset 2019

3.2.1 データセット概要

FFRI Dataset 2019 は FFRI Dataset 2018 に引き続き，マルウェアと良性ファイルの表層解析ログからなるデータセットである．用途として主に機械学習によるマルウェアの検知の研究を想定している．

表層解析ログを採用した理由は，動的解析ログに比して，実行時のコンテキストが存在しないため，初学者にも理解しやすく，研究に取り組みやすいと考えるからである．ここで実行時のコンテキストと呼称しているのは，マルウェアの挙動に影響を与えるすべての外的要因である．例えば，マルウェアを動作させたとき，その C2 サーバーとの通信が成功する場合と通信が失敗する場合では，そのマルウェアの挙動は大きく変化する．他にも解析環境が仮想環境であるか否か，解析環境の IP 帯，インストールされているソフトウェア等が例として挙げられる．

3.2.2 データ量およびデータソース

FFRI Dataset 2019 の生成元となるファイルは，株式会社 FFRI が収集したマルウェアおよび，良性の PE ファイルである．マルウェアのデータ，良性ファイルのデータはともにそれぞれ 25 万件である．マルウェアは，2018 年以降に収集した新しい検体である．種類は特に限定しておらず，ランサムウェアや標的型攻撃に使用されたマルウェア等さまざまな検体が含まれる．良性ファイルは，AV-TEST (https://www.av-test.org/en/) の FLARE サービスにより入手した PE ファイルである．種類は Windows OS のシステムファイル等が含まれる．

[*1] STIX は，MITRE Corporation の商標である．

[*2] Windows は，Microsoft Corporation の米国およびその他の国における登録商法または商標である．

表 3 FFRI Dataset 2019 表層解析データ項目一覧

No.	項目名	概要
1	id	id （検体の SHA-256）
2	file_size	ファイルサイズ
3	label	マルウェアは 1，クリーンウェアは 0
4	date	収集日 (マルウェアのみ)
5	hashes	ハッシュ値．5.1 から 5.13 を含む
5.1	md5	ハッシュ値
5.2	sha1	ハッシュ値
5.3	sha256	ハッシュ値
5.4	ssdeep	ファジーハッシュ値
5.5	imphash	インポートテーブルから算出したハッシュ値
5.6	impfuzzy	インポートテーブルに ssdeep を適用したファジーハッシュ値
5.7	tlsh	ファジーハッシュ値
5.8	totalhash	peHash の実装の一種
5.9	anymaster	peHash の実装の一種
5.10	anymaster_v1.0.1	peHash の実装の一種
5.11	endgame	peHash の実装の一種
5.12	crits	peHash の実装の一種
5.13	pehashng	peHash の実装の一種
6	peid	シグネチャのスキャン結果．6.1 から 6.10 を含む
6.1	Platform	32bit または 64bit
6.2	GUI Program	GUI プログラムか否か
6.3	Console Program	Console プログラムか否か
6.4	DLL	DLL か否か
6.5	Packed	パッキングの有無
6.6	Anti-Debug	Anti-Debug の有無
6.7	mutex	mutex の有無
6.8	Contain base64	Base64 文字列の有無
6.9	AntiDebug	AntiDebug 手法
6.10	PEiD	マッチした PEiD シグネチャ名
7	lief	ファイル表層情報
8	TrID	ファイル種別推定結果
9	strings	文字列出力結果

3.2.3 データフォーマットおよびデータ項目

データフォーマットは，jsonl ファイル形式である．

jsonl ファイルは，マルウェアデータ malware.jsonl と良性ファイルデータ cleanware.jsonl の 2 つで構成されている．jsonl ファイルは，1 行が 1 検体のデータの json となっている．検体のデータにはそれぞれ表 3 に示す項目が含まれる．ただし，No.4 の収集日はマルウェアデータにのみ含まれ，良性ファイルデータでは空欄になっている．No.5 の hashes キーの下に No.5.1 から No.5.13 が含まれている．これらには，一般的なハッシュ値および ssdeep [29] によるファジーハッシュ値，そしてマルウェアの識別・分類を目的に考案されたハッシュアルゴリズム imphash [30], impfuzzy [31], TLSH [32], および peHash [33] の値が含まれている．No.6 の peid キーの下の No.6.1 から No.6.10 までは，PEiD [34] による表層解析の結果であり，No.7 は lief [35] によるファイル表層解析の結果である． また，No.8 は TrID [36] によるファイル種別の推定結果である．No.9 は strings コマンドによる出力結果である．

なお，FFRI Dataset 2019 には，FFRI Dataset 2018 及び FFRI Dataset 2013–2017 の動的解析ログも含まれているが，その詳細は過去のデータセット解説論文 [7], [8], [9], [10], [11] を参照されたい．

3.3 NICTER Dataset 2019

NICTER Dataset 2019 は，国立研究開発法人情報通信研究機構 (NICT) で収集したダークネットトラフィックデータとスパムメールデータからなるデータセットである．

3.3.1 ダークネットトラフィックデータ

ダークネットとは，インターネット上で到達可能かつ未使用の IP アドレスのことを指す．通常のインターネット利用において，ダークネット宛てに対してトラフィックが送信されることは無いはずだが，実際にはネットワーク経由で感染を広げるマルウェアからのスキャントラフィック等のインターネット上における何らかの攻撃活動に起因したトラフィックが常時送信されている．こうしたダークネットに届くトラフィックを大規模に観測・分析することで，インターネット上における大局的な攻撃活動 (例えば，ワームタイプのマルウェアのパンデミックなど) を把握する手法をダークネット観測と呼ぶ．NICTER Dataset 2019 では，NICTER [37] プロジェクトで観測したダークネットトラフィックデータの一部を提供する．提供するデータは，ある連続した /20 ネットワーク (4,096 IPv4 アドレス) のダークネットで観測されたトラフィックであり，観測期間は 2011 年 4 月 1 日から 2019 年 3 月 31 日までを基本とするが，2019 年 4 月 1 日以降のデータについても随時提供される．8 年間以上の連続した観測データが提供されるため，攻撃活動の時間的な変化なども分析することが可能となっている．

3.3.2 スパムメールデータ

NICTER Dataset 2019 では，ダークネットトラフィックデータに加えて，スパムメールのデータセットとしてダブルバウンスメールを提供する．

ダブルバウンスメールはエラーメールの一種であり，主に送信元および宛先メールアドレスが共に存在しないメールによって発生するエラーメールである．実際にそのようなメールの大半は，送信元メールアドレスを詐称し，かつ，ランダムな宛先メールアドレスに対して送信するスパムメールである．したがって，ダブルバウンスメール (に含まれるオリジナルのメール) の分析を行うことにより，メールを経由した攻撃活動の把握が可能となる [38].

3.3.3 遠隔解析環境 NONSTOP

上記のダークネットトラフィックデータおよびスパムメールデータは，NONSTOP [39] と呼ぶ NICT が開発したサイバーセキュリティ情報分析用プラットフォーム上で提供される．NONSTOP では，各ユーザに対して専用の仮想マシン環境を提供し，ユーザはその環境にリモート

アクセスすることで各種データセットにアクセスできる．NONSTOP では，セキュリティに関連する機微なデータを扱うため，情報漏洩を防ぐために，NONSTOP 外とのデータ送受信時に適用される複数のフィルタ機能や，ロギング機能などが実装されている．

3.4 Soliton Dataset 2019

Soliton Dataset 2019 は，株式会社ソリトンシステムズのエンドポイントセキュリティソリューション “InfoTrace Mark II for Cyber”[40] 導入環境でマルウェアを実行したセキュリティログを中心としたデータセットである．

3.4.1 データセット提供背景

サイバー攻撃の高度化・巧妙化が進み，事前に侵入を阻止することが難しくなった今日，いち早く侵害に気付き，被害を極小化することが求められている．インターネットの出入り口 (境界) を監視するだけでは侵害の事実を明らかにすることが難しいことから，エンドポイント (端末) における脅威監視・検知が注目されている．しかしながら，攻撃の痕跡を削除あるいは改ざんする攻撃も増加しており，侵害の痕跡が残るエンドポイントのログ確保は，フォレンジック現場でますます重要となっている．

InfoTrace Mark II for Cyber (以下 Mark II) は上述した現場のニーズに応える EDR (Endpoint Detection and Response) として開発されたエンタープライズ向けセキュリティ製品である．Mark II は，サイバー攻撃対策だけではなく内部不正対策としても利用できるログ取得を目指していることから，マルウェアとは関係のない OS の挙動やユーザによる操作も記録される．こうしたログは，実際のフォレンジック現場で目にするデータに近いものとしてマルウェア対策研究に役立つと考えた．Soliton Dataset 2018 では，Mark II が導入された物理環境でマルウェアを動作させたログを提供したが，様々な角度からマルウェア対策研究を行うためには，複数種類の動的解析データが提供されていることが望ましい．そこで Soliton Dataset 2019 では，動的解析システム Cuckoo Sandbox[41] 上に Windows 7 Pro ベースで Mark II を導入したゲスト環境を構築し，Cuckoo ログと Mark II ログの両方をデータセットとして提供することとした．またデータセットに含まれる検体数は多いほうが良く，かつ，出来るだけ実際の攻撃に近い状態で取得した情報が望ましい．このため、この 1 年で話題になったマルウェアを実行する環境と，エクスプロイトキットを観測して得られたマルウェアを実行する環境の 2 種類の環境を利用した (図 5) .

3.4.2 データ量およびデータソース

2018 年 1 月から 2019 年 3 月までに話題になったマルウェアで，セキュリティベンダから解析結果が公開されているファミリを中心に，Windows 7 32bit 環境で動作するものをファイルタイプにこだわらず収集し，後述の実行環境で動作させたログ 485 件を提供対象とした（メイン環境）．また，類似の構成を持つ別環境においてエクスプロイトキット経由で得られたマルウェアを動作させて取得したログ 3 件についても提供対象とした（エクスプロイトキット観測環境およびマルウェア動作ログ取得環境）．

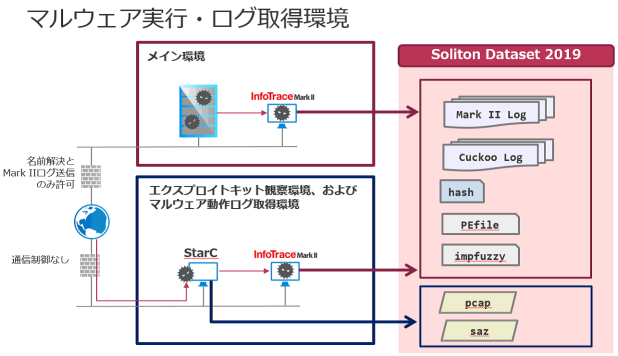

図 5 Soliton Dataset 2019 マルウェア実行・ログ取得環境

3.4.3 マルウェア実行環境

取得したマルウェア 485 検体を実行させたメイン環境は，名前解決と Mark II の管理用通信のみ許可し，その他のインターネットへのアクセスができない隔離環境とし，1 検体 6 分を基本として実行した．

一方，エクスプロイトキット観測環境は，StarC[42] を用いて構築された，インターネットへのアクセスが可能な環境である．ゲスト環境起動から 20 秒後に fiddler[43] と tshark[44] を VM 内で起動し，30 秒後に Web ブラウザ (Internet Explorer) に与えた悪性 url へのアクセス行った．エクスプロイトキット観測環境で得られた検体を，Cuckoo 上で Mark II を導入した Windows 7 のゲスト環境で実行し，動作ログを取得した．この環境での実行も 6 分を基本とした．このエクスプロイトキット観測環境で検体を取得した際のトラフィックデータとその後検体を実行した際の動作ログをデータセットに含めた．なお、いずれの環境の Mark II ログ/Cuckoo ログとも，マルウェアが起動できたかどうかにかかわらずデータセットに含めた．

3.4.4 データフォーマットおよびデータ項目

提供するデータは，マルウェア実行時の Mark II ログ，Cuckoo ログおよびマルウェア実行環境情報等である．PE ファイルに関しては，PEfile [45] および impfuzzy [31] の出力結果も提供する．また，エクスプロイトキット観測環境で取得したトラフィックデータ (pcap および saz) も提供対象とする．Mark II ログのデータフォーマットは，Key=Value 形式であるが，これを JSON 形式に変換するツール (mk2log) およびプロセスツリーを描画するツール (mk2tree) もあわせて提供する．

3.5 MWS Cup Dataset

MWS Cup 2018 参加 16 チームにより，事前課題とし

て各チームが作成したツールやデータセットとその概説資料からなるデータセットである．各チームが蓄積したきた技術を競うのみならず，その成果を MWS コミュニティに還流することを目的に，MWS Cup 2015 から成果物を MWS Datasets の一部として共有してきた．ここでは，MWS Cup 2018 事前課題の採点結果上位 3 チームが作成したツール，データセットを紹介する。

チーム「取鳥」が作成したオーケストレーションツールは，インシデント対応時に SOC から通知されたメールをもとに，被疑端末の特定と隔離、当該管理者への通知，チケット管理システムへの登録等を自動的に行い，実用性が高く評価され，詳細は論文としても MWS で報告 [46] している．チーム「m1z0r3」は，悪用されそうなドメイン名の取得状況や状態を調査する domain_scanner を作成した．様々な類似ドメイン名の生成手法に適用しており，その網羅性が高く評価された．Chrome Web Store 及び非公式サイトで公開されている Web ブラウザ拡張機能 1,458 件から抽出した特徴量を，データセットとして作成したチーム「UN 頼み」は新規性を高く評価されたものの悪性・良性のラベル付けやデータセットを用いた研究に期待が寄せられた．

4. MWS Datasets 利用状況

MWS Datasets を利用した多くの研究成果が発表されている．ここでは過去の MWS で発表された研究成果についてデータセットの利用内訳を表 4 に示す．データセットの更新が行われていない CCCDATAset，PRACTICE，NCD などの利用は減少傾向にある．一方で，FFRI Dataset や NICTER Dataset，BOS 等の利用は継続して行われている傾向にあるほか，2018 年より提供を開始した Soliton Dataset を利用もあった．サイバー攻撃の傾向変化に伴い研究対象も変化していることや，毎年一定数の学生の利用があることから教材としての有用性も定量的にわかる結果となっている．なお，MWS Datasets を利用した研究発表は MWS だけに留まらず，多数の国際会議や論文誌等への掲載を確認している [47].

5. おわりに

切磋琢磨を通して，新たなサイバー攻撃に対応可能な研究人材の育成に寄与する MWS コミュニティは，マルウェア対策研究に必要となる研究用データセットを継続的に作成および提供し，その研究成果を共有するフレームワークを推進している．本稿では最新のデータセットである MWS Datasets 2019 の概要を述べた．本データセットが研究者間で共通言語としての役割を担うことや，本データセットを用いて研究開発した技術等の共有により人材育成を含む本研究分野の発展に寄与すること，データセット作成そのものが研究対象分野として立ち上がり，研究活動をさらに発展させていくことが期待できる．

今後は最新の脅威を見据えた研究用データセットの拡充ならびにデータセットの利用環境構築および提供等，包括的なフレームワークを検討するとともに，評価用として利用可能なよりよい研究用標準データの作成に向けて検討していきたい．

謝辞 本研究にあたって，有益な助言とデータセット作成の協力を頂いた研究者コミュニティ，ならびに総務省実証実験プロジェクトおよび CCC 運営連絡会の関係者各位に深く感謝いたします．

参考文献

[1] “マルウェア対策研究人材育成ワークショップ”，https://www.iwsec.org/mws/
[2] “マルウェア対策研究人材育成ワークショップ 2019 (MWS2019)”，https://www.iwsec.org/mws/2019/
[3] “MWS Cup”，https://www.iwsec.org/mws/mwscup.html
[4] 畑田 充弘，中津留 勇，寺田 真敏，篠田 陽一，“マルウェア対策のための研究用データセットとワークショップを通じた研究成果の共有”, 情報処理学会シンポジウムシリーズ, Vol.2009, No.11, CSS2009 (MWS2009), pp.1–8, 2009 年 10 月.
[5] 畑田 充弘，中津留 勇，秋山 満昭，三輪 信介，“マルウェア対策のための研究用データセット ～MWS 2010 Datasets～”, 情報処理学会, マルウェア対策研究人材育成ワークショップ 2010 (MWS2010), 2010 年 10 月.
[6] 畑田 充弘，中津留 勇，秋山 満昭，“マルウェア対策のための研究用データセット ～MWS 2011 Datasets～”, 情報処理学会, マルウェア対策研究人材育成ワークショップ 2011 (MWS2011), 2011 年 10 月.
[7] 神薗 雅紀，畑田 充弘，寺田 真敏，秋山 満昭，笠間 貴弘，村上 純一，“マルウェア対策のための研究用データセット ～MWS datasets 2013～”, 情報処理学会, マルウェア対策研究人材育成ワークショップ 2013 (MWS2013), 2013 年 10 月.
[8] 秋山 満昭，神薗 雅紀，松木隆宏，畑田 充弘，“マルウェア対策のための研究用データセット ～MWS datasets 2014～”, 情報処理学会, CSEC-66(19), pp.125–131, 2014 年 7 月.
[9] 神薗 雅紀，秋山 満昭，笠間 貴弘，村上 純一，畑田 充弘，寺田 真敏，“マルウェア対策のための研究用データセット ～MWS datasets 2015～”, 情報処理学会, CSEC-70(6), pp.37–44, 2015 年 7 月.
[10] 高田 雄太，寺田 真敏，村上 純一，笠間 貴弘，吉岡 克成，畑田 充弘，“マルウェア対策のための研究用データセット ～MWS datasets 2016～”, 情報処理学会, CSEC-74(17), pp.1–8, 2016 年 7 月.
[11] 高田 雄太，寺田 真敏，松木 隆宏，笠間 貴弘，荒木 粧子，畑田 充弘，“マルウェア対策のための研究用データセット ～MWS datasets 2018～”, 研究報告コンピュータセキュリティ（CSEC）, 2018-CSEC-82(38), pp.1–8, 2018 年 7 月.
[12] “DARPA Intrusion Detection Data Sets,” http://www.ll.mit.edu/r-d/datasets
[13] “MALICIA Project,” http://malicia-project.com/dataset.html
[14] “Android Malware Genome Project,” http://www.malgenomeproject.org
[15] “CTU-13 DATASET,” https://mcfp.weebly.com/

表 4 MWS2008–2018 における MWS Datasets を用いた論文の発表件数 (一部の論文は複数データセットを利用．“–” は提供なし．)

MWS Datasets	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
CCC	22	27	16	15	9	3	3	3	3	4	1
MARS	–	–	1	1	–	–	–	–	–	–	–
D3M	–	–	4	3	3	9	14	9	2	7	2
IIJ MITF	–	–	–	1	–	–	–	–	–	–	–
FFRI	–	–	–	–	–	5	2	4	3	2	5
PRACTICE	–	–	–	–	–	3	1	0	1	0	0
NICTER	–	–	–	–	–	6	2	3	0	2	0
BOS	–	–	–	–	–	–	1	4	2	3	4
NCD	–	–	–	–	–	–	–	0	0	3	0
Soliton	–	–	–	–	–	–	–	–	–	–	1
データセット説明	0	1	1	1	0	1	0	0	0	0	0
合計	22	28	22	21	12	27	23	23	11	21	13
内，学生発表件数	8	15	10	9	9	10	10	14	8	12	9

the-ctu-13-dataset-a-labeled-dataset-with-botnet-normal-and-background-traffic.html
[16] “NIDS 評価用データセット：Kyoto 2016 Dataset の作成,” http://id.nii.ac.jp/1001/00183519/
[17] “CAIDA Data - Overview of Datasets, Monitors, and Reports,” http://www.caida.org/data/overview/
[18] “The Information Marketplace for Policy and Analysis of Cyber-risk & Trust (IMPACT),” https://www.impactcybertrust.org/
[19] “MAWILab,” http://www.fukuda-lab.org/mawilab/
[20] “Malware-Traffic-Analysis.net,” http://www.malware-traffic-analysis.net/
[21] “Contagio Malware Dump,” http://contagiodump.blogspot.com/
[22] H. S. Anderson and P. Roth, “EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models,” ArXiv e-prints, http://adsabs.harvard.edu/abs/2018arXiv180404637A
[23] “Microsoft Malware Prediction,” https://www.kaggle.com/c/microsoft-malware-prediction
[24] “A Realistic Cyber Defense Dataset (CSE-CIC-IDS2018),” https://registry.opendata.aws/cse-cic-ids2018/
[25] “MaxNet,” http://public.avast.com/research/maxnet.html
[26] “CYBER-SECURITY MODBUS ICS DATASET,” https://ieee-dataport.org/documents/cyber-security-modbus-ics-dataset
[27] “International Secure Systems Lab,” http://www.iseclab.org
[28] 寺田 真敏，佐藤 隆行, 青木 翔，亀川 慧，清水 努, 津田 侑，“研究用データセット「動的活動観測 2018」”, 情報処理学会, コンピュータセキュリティシンポジウム (CSS) 2018, 2018 年 10 月.
[29] “ssdeep - Fuzzy hashing program”, https://ssdeep-project.github.io/ssdeep/
[30] “FireEye - Tracking Malware with Import Hashing”, https://www.fireeye.com/blog/threat-research/2014/01/tracking-malware-import-hashing.html
[31] “Fuzzy Hash calculated from import API of PE files”, https://github.com/JPCERTCC/impfuzzy
[32] “TLSH - Trend Micro Locality Sensitive Hash”, https://github.com/trendmicro/tlsh
[33] “Compilation of peHash implementations.”, https://github.com/knowmalware/pehash
[34] “Yet another implementation of PEiD with yara”, https://github.com/K-atc/PEiD
[35] “LIEF - Library to Instrument Executable Formats”, https://lief.quarkslab.com/
[36] “TrID – File Identifier”, http://mark0.net/soft-trid-e.html
[37] D. Inoue, M. Eto, K. Yoshioka, S. Baba, K.Suzuki, J. Nakazato, K. Ohtaka, K. Nakao, “nicter: An Incident Analysis System Toward Binding Network Monitoring with Malware Analysis,” Proceedings of the WOMBAT Workshop on Information Security Threats Data Collection and Sharing (WISTDCS 2008), pp. 58–66, 2008.
[38] 笠間 貴弘，神宮 真人，清水 雄介，井上 大介，“ダブルバウンスメールを活用した悪性メール対策の有効性,” 情報処理学会 マルウェア対策研究人材育成ワークショップ (MWS) 2016, 2016 年 10 月.
[39] 竹久達也，井上大介，衛藤将史，吉岡克成，笠間貴弘，中里純二，中尾康二，“サイバーセキュリティ情報遠隔分析基盤 NONSTOP”, 電子情報通信学会 情報通信システムセキュリティ研究会 (ICSS), pp.85–90, 2013 年 6 月.
[40] “InfoTrace Mark II for Cyber,” https://www.soliton.co.jp/mark2/
[41] “Cuckoo Sandbox,” https://cuckoosandbox.org/
[42] “Starc,” https://github.com/nao-sec/starc
[43] “fiddler,” https://www.telerik.com/fiddler
[44] “tshark,” https://www.wireshark.org/docs/man-pages/tshark.html
[45] “PEfile”, https://github.com/erocarrera/pefile
[46] 大森 幹之，東野 正幸，川戸 聡也，宮田 直輝，高橋 健一，川村 尚生，“On Automation and Orchestration of an Initial Computer Security Incident Response Using Centralized Incident Tracking System,” 情報処理学会 マルウェア対策研究人材育成ワークショップ (MWS) 2018, 2018 年 10 月.
[47] “研究用データセット MWS Datasets を用いた研究活動について,” https://www.iwsec.org/mws/achievements.html