Vol.2018-CSEC-82 No.38
Vol.2018-SPT-29 No.38
2018/7/26

マルウェア対策のための研究用データセット ～MWS Datasets 2018～

高田 雄太[1,a)]　寺田 真敏[2]　松木 隆宏[3]　笠間 貴弘[4]　荒木 粧子[5]　畑田 充弘[6]

概要：マルウェア感染の脅威に対して，多くの研究者がマルウェアの検知，解析，対策に関する様々な手法を提案している．しかしながら，近年の脅威は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”の収集と共有が困難な状況が続いている．このような状況に対して我々は，研究に必要となる情報を収集して研究成果の客観的な評価と共有を容易にするためのデータセット MWS Datasets 2018 を作成した．本稿では，MWS Datasets 2018 として新しく提供する BOS 2018，FFRI Dataset 2018，NICTER Dataset 2018，Soliton Dataset 2018 および継続的に提供する CCC DATAset，D3M，NCD in MWS Cup 2014，PRACTICE Dataset 2013，PRACTICE (AmpPot) Dataset 2015 の概要を報告する．

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 2018～

Yuta Takata[1,a)]　Masato Terada[2]　Takahiro Matsuki[3]　Takahiro Kasama[4]　Shoko Araki[5]
Mitsuhiro Hatada[6]

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

1 日本電信電話株式会社 セキュアプラットフォーム研究所
NTT Secure Platform Laboratories
2 株式会社日立製作所
Hitachi, Ltd.
3 株式会社 FFRI
FFRI, Inc.
4 国立研究開発法人 情報通信研究機構
National Institute of Information and Communications Technology
5 株式会社ソリトンシステムズ
Soliton Systems K.K.
6 エヌ・ティ・ティ・コミュニケーションズ株式会社
NTT Communications Corporation
a) yuta.takata.xy@hco.ntt.co.jp

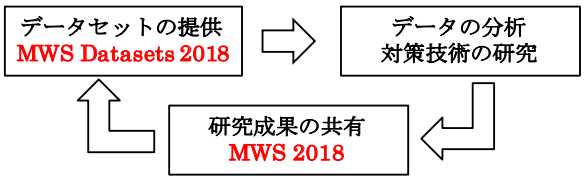

図 1 マルウェア対策研究のサイクル

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 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] (2018 年は MWS2018 [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 2018 (図 2) について報告する．2018 年は下記のデータセットから構成される．

(1) **BOS 2018** — 標的型攻撃の観測データ (§3.1)
(2) **FFRI Dataset 2018** — サンドボックスにおけるマルウェアの動的解析データ (§3.2)
(3) **NICTER Dataset 2018** — ダークネットにおけるパケットデータ (§3.3)
(4) **Soliton Dataset 2018** — 物理環境におけるマルウェアの動的解析データ (§3.4)
(5) **MWS Cup Dataset** — セキュリティ競技 MWS Cup 参加チームが収集したデータ (§3.5)

なお，CCC DATAset，D3M Dataset，NCD in MWS Cup 2014，PRACTICE Dataset 2013, 2015 はデータセットの内容に更新がないため，2015 年およびそれ以前のデータセットを継続的に提供する．これらデータセットの詳細は，文献 [3], [4], [5], [6], [7], [8], [9] 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は次の通りである．これら以外にも研究用データセットは存在するが，データセット作成が 10 年以上前のものや，データセット提供を終了しているものが多い．

- **CAIDA Data** [10] — ネットワーク運用に関わる通信ログのデータセット
- **MAWILab** [11] — サンプリングで保存された通信リポジトリにラベル付けしたデータセット
- **IMPACT Dataset** [12] — ネットワークデータ装置やセキュリティ装置，通信ログ等から得られるセキュリティ脅威に関するデータセット
- **MALICIA Dataset** [13] — 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から収集したマルウェア検体のデータセット
- **Malware-Traffic-Analysis.net** [14] — マルウェア感染およびエクスプロイトキットに関する通信データ
- **Contagio Malware Dump** [15] — 各種ファイルフォーマットの正規ファイルおよび悪性ファイル
- **Android Malware Genome Project Dataset** [16] — マルウェアファミリ毎に分類された Android マルウェア検体
- **ACODE dataset** [17] — Google Play とサードパーティマーケットから収集した Android アプリ 20 万個の説明文に関するデータセット
- **EMBER** [18] — PE ファイルのハッシュ値とその特徴量，およびそれを識別するベンチマークモデル

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 [19] 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

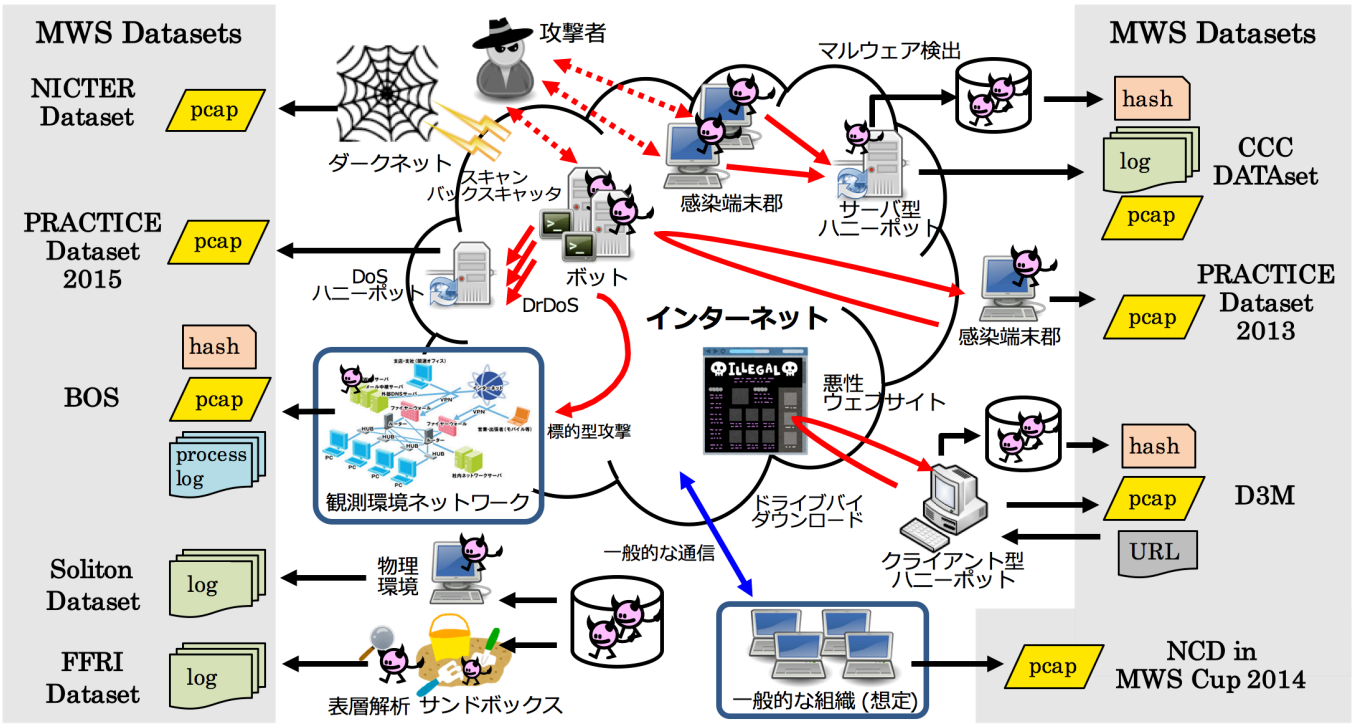

図 2 MWS Datasets 2018 の概要

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 [20] は 1998 年から 2000 年までに作成された IDS のトラヒックデータセットであり，Kyoto_data [21] は 2006 年から 2009 年までに様々な種類のハニーポットから収集されたトラヒックデータである．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 2018

本章では，MWS Datasets 2018 において新しく更新のあったデータセットの概要を述べる．

3.1 BOS 2018

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

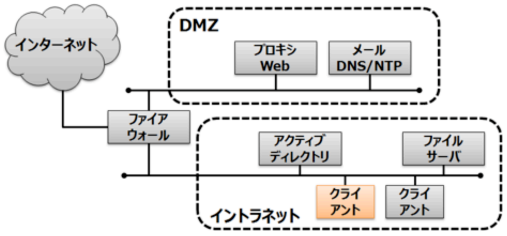

図 3 動的活動観測環境の概要

表 1 BOS 2018 の観測事例

No.	観測期間		マルウェア検体名	進行度
	開始	終了		
g01	2017/07/27	2017/07/30	LNK_DLOADER.AUSBXT	6
g02	2017/08/03	2017/08/06	TROJ_DYER.BMC	4
g03	2017/08/17	2017/08/20	BKDR_FARFLI.SMB	2
g04	2017/08/30	2017/09/02	TROJ_SHELLDOWN.ZKEH-A	6
g05	2017/09/04	2017/09/07	TSPY_KONNI.A	2
g06	2017/11/02	2017/11/09	TROJ_DEDEX.GQA	6
g07	2017/11/30	2017/12/07	VAN_DROPPER_UMXX	7
g08	2017/11/30	2017/12/01	VAN_DROPPER_UMXX	7
g09	2017/12/19	2017/12/26	TROJ_GEN.R011C0WL817	1
g10	2017/12/19	2017/12/26	TROJ_GEN.R011C0WL817	5
g11	2018/01/12	2017/01/19	W2KM_SHELLEX.BYZ	6
g12	2017/01/16	2017/01/23	W2KM_SHELLEX.BYZ	5
g13	2018/01/19	2018/01/26	BKDR_PLEAD.SMZTDK-A	8
g14	2018/01/23	2018/01/31	BKDR_PLEAD.SMZTDK-A	8

築支援」で得られた成果の一部である [22].

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．

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 2018 は，表 1 に示す観測事例に関連する下記 3 種類のデータ，1. マルウェア検体，2. 通信観測データ，3. プロセス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．

[*1] STIX は，MITRE Corporation の商標である．

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

また，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．なお，過去の BOS 2014–2016 は，2013 年以降の観測事例に関連する上記 3 種類のデータを含んでおり，これらは BOS 2018 に含まれる．

3.2 FFRI Dataset 2018

従来の FFRI Dataset 2013–2017 は，動的解析ログのデータセットであったが，新しく提供する FFRI Dataset 2018 は，マルウェアと良性ファイルの表層解析ログのデータセットである．表層解析ログは，マルウェア解析の初心者にも理解しやすいため，研究に取り組みやすいと考えら

[*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 2018 表層解析データ項目一覧

No.	項目名	概要
1	date	収集日 (マルウェアのみ)
2	MD5	ハッシュ値
3	SHA-1	ハッシュ値
4	SHA-256	ハッシュ値
5	ssdeep	ファジーハッシュ値
6	imphash	インポートテーブルから算出したハッシュ値
7	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
8	Totalhash	peHash の実装の一種
9	AnyMaster	peHash の実装の一種
10	AnyMaster_v1.0.1	peHash の実装の一種
11	EndGame	peHash の実装の一種
12	Crits	peHash の実装の一種
13	peHashNG	peHash の実装の一種
14	Platform	32bit または 64bit
15	GUI Program	GUI プログラムか否か
16	Console Program	Console プログラムか否か
17	DLL	DLL か否か
18	Packed	パッキングの有無
19	Anti-Debug	Anti-Debug の有無
20	mutex	mutex の有無
21	Contain base64	Base64 文字列の有無
22	Anti-DebugMethod	AntiDebug 手法 (デリミタは “|”)
23	PEiD	マッチした PEiD シグネチャ名 (デリミタは “|”)
24	TrID	ファイル種別推定結果

れる．また，従来の動的解析ログよりも多くのデータを提供するとともに，良性ファイルのデータも提供することで，検出率と誤検出率の評価が可能となる．したがって，FFRI Dataset 2018 は，機械学習を利用したマルウェアの検出技術の研究等に活用されることを想定している．なお，FFRI Dataset 2018 には，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 [6], [7], [8], [9] を参照されたい．

3.2.1　データ量およびデータソース

FFRI Dataset 2018 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は 29 万件，良性ファイルのデータは 21 万件である．マルウェアは，2017 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，2008 年から 2018 年までの間に収集したファイルである．種類は Windows OS や Microsoft Office 等に含まれるファイル，メーカー製 PC にプリインストールされているサードパーティソフトウェアのファイル，Vector [23] で公開されているフリーウェア等が含まれる．

3.2.2　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，CSV とテキストの 2 種類の形式に分かれている．

CSV ファイルは，マルウェアデータ malware.csv と良性ファイルデータ cleanware.csv の 2 つで構成されており，それぞれ表 3 に示す 24 項目のデータが含まれる．ただし，No.1 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.2 から No.13 までは，一般的なハッシュ値および ssdeep [24] 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 [25], impfuzzy [26], および peHash [27] の値が含まれている．No.14 から No.23 までは，PEiD [28] による表層解析の結果であり，No.24 は TrID [29] によるファイル種別の推定結果である．

テキストファイルは，フォルダ malware と cleanware に含まれており，pefile [30] の `pe.dump_info()` で取得したファイルヘッダのダンプ情報が各検体ごとに記録されている．ファイル名は検体の SHA-256 となっている．

3.3　NICTER Dataset 2018

NICTER Dataset 2018 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の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 アドレス空間のことを指す．未使用，すなわち PC やサーバ等が接続されていないため，一般的なインターネットの利用において，ダークネット宛てにトラフィックが送られることは無いはずだが，実際には常時大量のトラフィックがダークネット宛てに送信されている．これらダークネットに届くトラフィックの多くは，ネットワークを経由して感染を広げるタイプのマルウェアによるスキャンやマルウェア同士が P2P ネットワークを確立するためのランデブーパケット，送信元 IP アドレスを詐称した DDoS 攻撃を受けている被害サーバからの応答 (バックスキャッタ) 等，何らかの攻撃活動に起因したトラフィックである．したがって，ダークネットに届くトラフィックを大規模に観測・分析することにより，インターネット上における攻撃活動の傾向把握が可能となる．

NICTER Dataset 2018 では，NICTER [31] 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．当該ダークネットは，ある連続した /20 ネットワーク (4,096 IP アドレス) であり，観測対象の秘匿のため，宛先 IP アドレスにおける第 1 および第 2 オクテットの値をランダム値に置換している．観測期間は 2011 年 4 月 1 日から 2018 年 3 月 31 日までを基本とするが，2018 年 4 月 1 日以降のデータについても随時提供する．

3.3.2　スパムメールデータ

NICTER Dataset 2018 では，上述したダークネットトラフィックデータに加えて，NICT のメールサーバに届いたスパムメールのうちダブルバウンスメールと呼ばれるタ

イプのメールデータを提供する．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．このようなメールは，通常のメール利用で発生することは稀であり，主に送信元メールアドレスを詐称して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が当てはまる．したがって，ダブルバウンス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 [32]．

3.3.3 NONSTOP

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 [33] と呼ばれる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 は，PaaS の形態となっており，利用を希望する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アクセスすることで各種データセットを利用できる．セキュリティに関連するデータという機微な情報を扱うため，情報流出を防ぐため，NONSTOP 外部とのデータ送受信に複数のフィルタによる検査機能や転送ファイルの保存機能等が実装されている．

3.4 Soliton Dataset 2018

Soliton Dataset 2018 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” 導入環境でマルウェアを実行したセキュリティログの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，マルウェアを動作させた際の Mark II ログをデータセットとして提供する．

3.4.2 データ量およびデータソース

2017 年 1 月から 2018 年 2 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た WannaCry や Petya 等の検体を中心にファイルタイプにこだわらず検体を収集した．このうち後述のマルウェア実行環境でプロセスとして起動した検体のログ 117 件を提供対象とした．なお，具体的な侵害活動に至らなかったものの，プロセスとして起動した検体のログも含まれている．

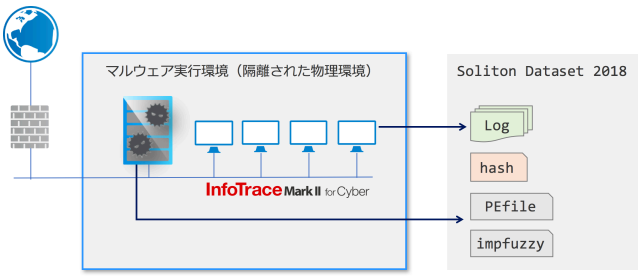

図 4 Soliton Dataset 2018 マルウェア実行環境

3.4.3 マルウェア実行環境

仮想環境上では動作しないマルウェアが増加している．そこで，より有益なマルウェア動的解析ログを取得するため，物理環境上でのマルウェア実行と Mark II によるログ取得を行う独自のマルウェア実行環境 (図 4) を構築した．本マルウェア実行環境は，Windows 7 Pro ベースの環境であり，マルウェア実行時に w32tm.exe による時刻同期結果および systeminfo.exe による出力結果を取得する．さらに本マルウェア実行環境では，物理端末のディスクロールバックをデバイスマッパーで行い，マルウェアを活性化させるためダイアログのボタンを自動的にクリックする等の工夫を施している．なお，1 検体 30 分の実行を基本としたが，必要に応じて実行時間を増減した．また，本マルウェア実行環境は隔離されており，原則 1 検体を 1 端末上で実行しているが，ネットワークを通じて他端末に感染を試みる (横展開する) マルウェアについては，横展開が可能な端末設定を行い，感染元と感染先両方でのログを取得した．

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 ログおよびマルウェア実行環境情報である．PE ファイルに関しては，pefile [30] および impfuzzy [26] の出力結果も提供する．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 Dataset は，セキュリティ競技 “MWS Cup” を通じて参加チームが収集したデータセットである．MWS Datasets 2018 には，MWS2017 で開催された MWS Cup 2017 の参加チームが収集したデータセットである Alkanet データセットおよび Jinkai Dataset 2017 が含まれる．

表 4 MWS2008–2017 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
CCC (マルウェア検体)	5	7	6	5	7	3	3	0	1	2
CCC (攻撃通信データ)	9	14	5	6	2	–	–	2	1	2
CCC (攻撃元データ)	8	6	5	4	–	–	–	1	1	0
MARS	–	–	1	1	–	–	–	–	–	–
D3M	–	–	4	3	3	9	14	9	2	7
IIJ MITF	–	–	–	1	–	–	–	–	–	–
FFRI	–	–	–	–	–	5	2	4	3	2
PRACTICE	–	–	–	–	–	3	1	0	1	0
NICTER	–	–	–	–	–	6	2	3	0	2
BOS	–	–	–	–	–	–	1	4	2	3
NCD	–	–	–	–	–	–	–	0	0	3
データセット説明	0	1	1	1	0	1	0	0	1	0
合計	22	28	22	20	13	27	23	23	11	17
学生発表件数	8	15	10	9	9	10	10	14	8	12

チーム “たこ焼き Lab” が提供するデータセット “Alkanet データセット” は，システムコールトレーサ Alkanet によるマルウェアの動的解析ログである．発行されたシステムコールのログと解析のサマリー結果が含まれている．

チーム “人海戦術 White” が提供するデータセット “Jinkai Dataset 2017” は，Drive-by Download 攻撃に関連するトラフィックのデータセットである．独自の Web クライアント型ハニーポットで 2017 年 6 月以降に観測した様々な攻撃キャンペーンと Exploit Kit に関するトラフィックが含まれている．具体的には，概要を説明するテキストファイル，ハニーポットに投入した URL，URL 巡回時におけるハニーポットのスクリーンショット，そしてトラフィックデータである pcap ファイルで構成されている．

4. MWS Datasets 利用状況

MWS Datasets を利用した研究成果を共有する場である「マルウェア対策研究人材育成ワークショップ (MWS)」では，多くの研究成果が発表されている．過去の MWS Datasets と MWS で発表された研究における利用内訳を表 4 に示す．

CCCDATAset は従来のネットワーク感染型マルウェアのデータセットであり，さらに提供情報量も少なくなっているため，当該データセットを利用した研究は減少傾向にある．一方で，FFRI Dataset や NICTER Dataset，BOS 等を用いた研究は，継続して活用されている傾向にある．また，ウェブ感染型マルウェアを含む D3M の件数は一時期減少したものの，依然として利用件数は多かった．実際のサイバー攻撃における攻撃やマルウェアの傾向変化に伴い，研究対象も徐々に変化していることが定量的にわかる結果となっている．MWS は，このような攻撃手法やマルウェアの傾向変化を網羅できるデータセットを継続的に提供し続けることができる活動へと発展していく必要がある．

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 [35]．

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 2018 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 2018 (MWS2018)”, https://www.iwsec.org/mws/2018/
[3] 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.
[4] 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.
[5] 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.
[6] 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.
[7] 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.
[8] 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.
[9] 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.
[10] “CAIDA Data - Overview of Datasets, Monitors, and Reports,” http://www.caida.org/data/overview/
[11] “MAWILab,” http://www.fukuda-lab.org/mawilab/
[12] “The Information Marketplace for Policy and Analysis of Cyber-risk & Trust (IMPACT),” https://www.impactcybertrust.org/
[13] “MALICIA Project,” http://malicia-project.com/dataset.html
[14] “Malware-Traffic-Analysis.net,” http://www.malware-traffic-analysis.net/
[15] “Contagio Malware Dump,” http://contagiodump.blogspot.jp
[16] “Android Malware Genome Project,” http://www.malgenomeproject.org
[17] “The ACODE dataset,” http://nsl.cs.waseda.ac.jp/projects/acode/
[18] 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
[19] “International Secure Systems Lab,” http://www.iseclab.org
[20] “DARPA Intrusion Detection Data Sets,” http://www.ll.mit.edu/mission/communications/cyber/CSTcorpora/ideval/data/
[21] “Traffic Data from Kyoto University’s Honeypots,” http://www.takakura.com/Kyoto_data/
[22] 寺田 真敏，佐藤 隆行, 青木 翔，亀川 慧，清水 努, 萩原 健太，“研究用データセット「動的活動観測 2017」”, 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) 2017, 2017 年 10 月.
[23] “Vector: ソフトライブラリ&PC ショップ”, https://www.vector.co.jp/
[24] “ssdeep - Fuzzy hashing program”, https://ssdeep-project.github.io/ssdeep/
[25] “FireEye - Tracking Malware with Import Hashing”, https://www.fireeye.com/blog/threat-research/2014/01/tracking-malware-import-hashing.html
[26] “Fuzzy Hash calculated from import API of PE files”, https://github.com/JPCERTCC/impfuzzy
[27] “Compilation of peHash implementations.”, https://github.com/knowmalware/pehash
[28] “Yet another implementation of PEiD with yara”, https://github.com/K-atc/PEiD
[29] “TrID – File Identifier”, http://mark0.net/soft-trid-e.html
[30] “pefile”, https://github.com/erocarrera/pefile
[31] 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.
[32] 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.
[33] 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.
[34] L. Kramer, J. Krupp, D. Makita, T. Nishizoe, T. Koide, K. Yoshioka, C. Rossow, “AmpPot: Monitoring and Defending Amplification DDoS Attacks,” In Proceedings of the 18th International Symposium on Research in Attacks, Intrusions and Defenses (RAID), November, 2015.
[35] “研究用データセット MWS Datasets を用いた研究活動について,” http://www.iwsec.org/mws/2014/about.html\#relatedActivities