Vol.2016-CSEC-74 No.17
Vol.2016-SPT-19 No.17
2016/7/14

マルウェア対策のための研究用データセット ～MWS Datasets 2016～

高田 雄太[1,a)]　寺田 真敏[2]　村上 純一[3]　笠間 貴弘[4]　吉岡 克成[5]　畑田 充弘[6]

概要：マルウェアの脅威に対してマルウェアの検知，解析，対策と様々なアプローチで研究が行われている．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 2016 を作成した．本稿では，MWS Datasets 2016 を構成する BOS 2016，FFRI Dataset 2016，NICTER Darknet Dataset 2016，PRACTICE (AmpPot) Dataset 2015 および継続的に提供する CCC DATAset，D3M，NCD in MWS Cup 2014，PRACTICE Dataset 2013 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 Dataset，CCC DATAset，D3M，FFRI Dataset，NICTER Darknet Dataset，PRACTICE Dataset

Datasets for Anti-Malware Research ～MWS Datasets 2016～

Yuta Takata[1,a)]　Masato Terada[2]　Junichi Murakami[3]　Takahiro Kasama[4]
Katsunari Yoshioka[5]　Mitsuhiro Hatada[6]

Abstract: Many security researches, such as malware detection, analysis, and prevention, have continued to take countermeasures against malware threats. However, diversification and evolution of the recent attack make it increasingly difficult to collection and sharing of “research materials” that form the basis of security research. To overcome this problem, we collected data related with malware threats and made the datasets called MWS Datasets 2016 for evaluation of the proposals and sharing of the research achievements. In this paper, we introduce an overview of MWS Datasets 2016 comprised of BOS 2016, FFRI Dataset 2016, NICTER Darknet Dataset 2016 and PRACTICE (AmpPot) Dataset 2015. We also summarize CCC DATAset, D3M, NCD in MWS Cup 2014, and PRACTICE Dataset 2013 that are continuously provided and additionally contained in MWS Datasets 2016.

Keywords: Dataset, Malware, MWS Datasets, BOS Dataset, CCC DATAset, D3M, FFRI Dataset, NICTER Darknet Dataset, PRACTICE Dataset

1. はじめに

高度化かつ複雑化したサイバー攻撃が世界的な問題となっており，各組織が個別に対策することはもちろんのこと，国家レベルや国家間での対策が急務となっている．特

[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] 横浜国立大学
Yokohama National University
[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)] takata.yuta@lab.ntt.co.jp

に，マルウェアに起因した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．

一つ目の阻害要因におけ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感染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，標的型攻撃などで組織内に侵入された際のマルウェアの挙動等を指し，可能な限り網羅的に，かつ攻撃の進化に合わせて適切に選択されたものが望ましい．従来では研究素材となるこのようなデータは，主に研究者が収集環境を構築して自ら収集し，個々の技術の有効性や妥当性を評価してきた．すなわち，同じ研究テーマに取り組んだ場合であっても研究素材が異なるため，その研究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，またそれらが年々高度化していることが，収集を困難にさせる主な原因である．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様々な解析および検知を回避する機能を有しており，情報を収集する環境によっては期待した情報を取得することができないため，その結果として定性的にも定量的にも研究素材の収集が困難になっている．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が困難である．さらに標的型攻撃においては，例えば攻撃者が RAT 等を利用して標的組織内でどのように振る舞うかが主な焦点となるが，これらの情報を収集するには攻撃者の標的組織となり，かつ侵入された際の挙動を保全しておく必要がある．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．なお，研究素材を収集することが困難となってきている傾向は，マルウェアを含むサイバー攻撃による脅威全般に当てはまると言える．

このように進化を続け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 に示す通り「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，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 (2016 年も MWS2016 [1] を開催

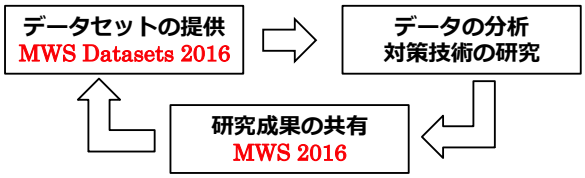

図 1 マルウェア対策研究のサイクル

する予定である．)．さらなる研究の発展のため，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 2016 (図 2) について報告する．2016 年は下記のデータセットから構成される．

(1) **BOS 2016** — 標的型攻撃の観測データ
(2) **FFRI Dataset 2016** — マルウェアの動的解析データ
(3) **NICTER Darknet Dataset 2016** — ダークネットパケットデータ
(4) **PRACTICE (AmpPot) Dataset 2015** — DRDoS 攻撃の観測データ
(5) **CCC DATAset** — 待受型ハニーポットで収集したデータ
(6) **D3M** — Web 感染型マルウェアの観測データ
(7) **NCD in MWS Cup 2014** — インターネット回線に接続された組織の一般的な通信を想定した悪性ではないデータ
(8) **PRACTICE Dataset 2013** — マルウェアの長期観測データ

なお，CCC DATAset，D3M，NCD in MWS Cup 2014，PRACTICE Dataset 2013 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 [2], [3], [4], [5], [6], [7] 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** [8] — ネットワーク運用に関わる通信ログのデータセット
- **MAWILab** [9] — 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** [10] — ネットワークデータ装置

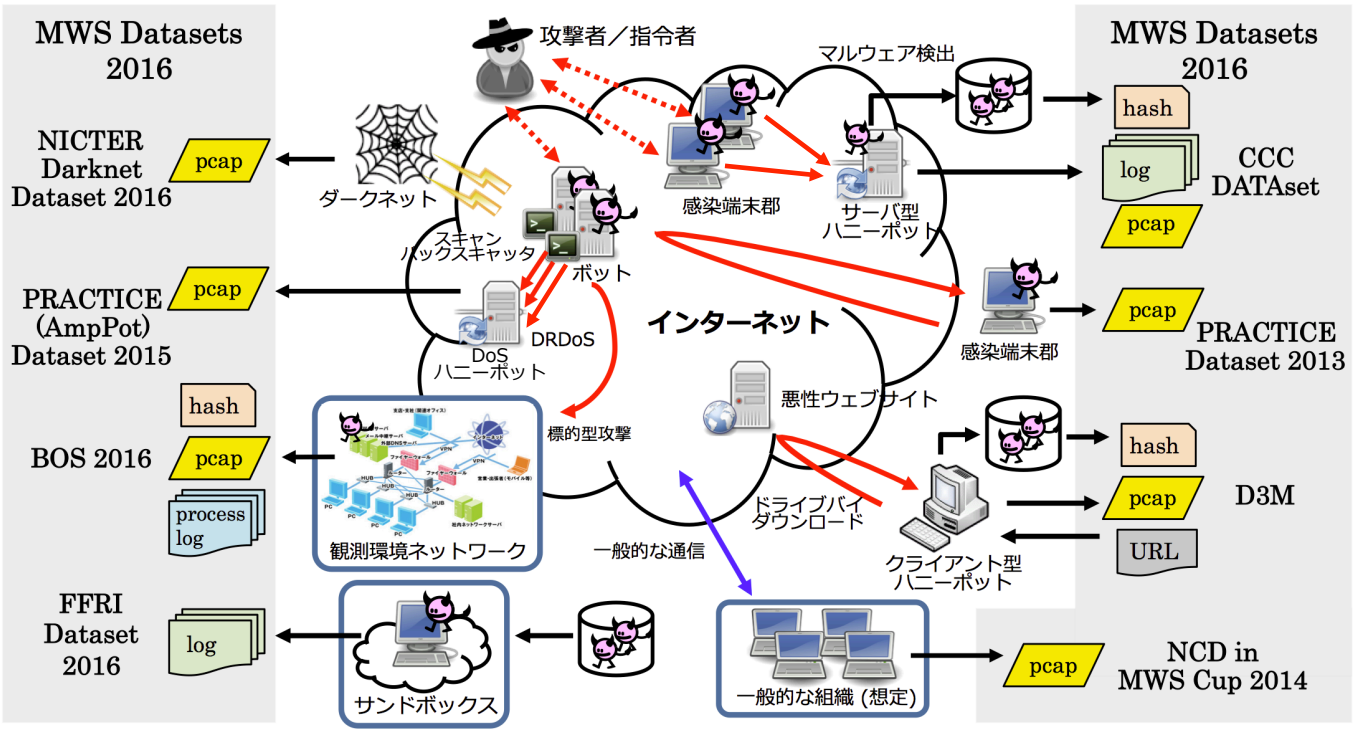

図 2 MWS Datasets 2016 の概要

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** [11] — 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** [12] — 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** [13] — 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** [14] — 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** [15] — 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

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において，そのデータセット入手の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 [16] 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ともな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 [17] 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 [18] 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 2016

本章では，MWS Datasets 2016 の各データセットの概要を述べる．

3.1 BOS 2016

動的活動観測 Behavior Observable System (BOS) 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，総務省実証事業「サイバー攻撃解析・防御モデル実践演習の実証実験の請負」にて得られた成果の一部である [19], [20].

3.1.1 目的

これまでのマルウェア検体の静的解析および動的解析は，マルウェアの挙動に着目した解析が多かった．例えば，指令サーバ接続，情報窃取，バックドアなどの機能の存在や挙動把握に重点が置かれ，これら機能のどれを使ったのか，どのような順番で使ったのか等，攻撃者の行動という観点での解析は少なかった．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，静的解析および動的解析が実施されてきた．

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 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等，攻撃者の行動情報を組み合わせることで，攻撃者の行動という観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 観測環境

動的活動観測環境は，組織内ネットワーク自身を模擬した観測環境を構築している (図 3)．この環境は，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，プロキシ経由またはプロキシ経由なしの形態で，実インターネットへのアクセスが可能である．

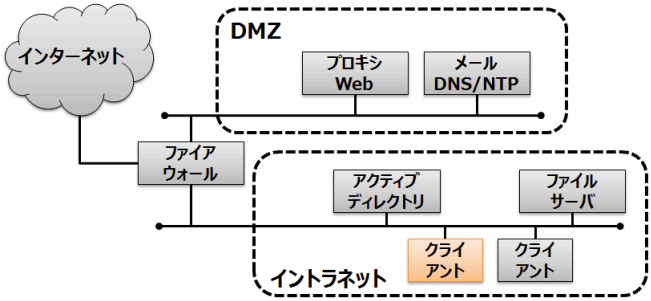

図 3 動的活動観測環境の概要

表 1 BOS 2014 – BOS 2016 観測事例の一覧

BOS	事例	観測期間	検知名
2014	c11	2013/11/14 – 11/27	BKDR_POISON.BWB
	c21	2014/03/19 – 03/26	BKDR_ZACOM.AD
2015	d18	2014/10/06 – 11/07	BKDR_EMDIVI.I
	d19	2014/10/06 – 11/07	BKDR_EMDIVI.F
	d33	2014/12/08 – 12/22	BKDR_PLUGX.DUKLR
	d37	2014/01/23 – 02/02	BKDR_EMDIVI.AB
2016	e04	2015/08/06 – 08/20	BKDR_EMDIVI.SMB

3.1.3 データセット構成

BOS 2016 は，BOS 2014–2015 を含め，計 7 つの観測事例を含んでいる (表 1)．また，BOS 2016 から攻撃者の活動観測に至らなかった事例も研究用データセットとして活用できると考え同梱している．

BOS が提供するデータセットは，マルウェア検体，通信観測データ，プロセス観測データの 3 つである．

(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テキスト形式で記載したファイルである．

(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キャプチャデータであり，攻撃者の行動に関する解析が可能である．

(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保全したデータであり，攻撃者の行動に関する解析が可能である．

3.1.4 観測事例

観測事例として，2014 年 9 月中旬頃に流布した医療費通知の偽装メールに起因する事例 `d18` を紹介する．医療費通知の偽装メールは，健康保険組合などからの医療費通知メールを偽装し，ユーザのパソコンを遠隔操作可能な不正プログラム (検知名: BKDR_EMDIVI.I) に感染させようとする攻撃であった．

医療費通知メールの添付ファイルには，文書アイコン偽

表 2　事例 d18 におけるマルウェア検体の観測事象

日付	時刻	観測事象
10/06	15:43	検体 (医療費通知のお知らせ.exe) を実行し，ファイル 2 つ (leassaq.exe, kptl.doc) が生成され，指令サーバとの接続を確立.
	22:42 – 23:32	指令サーバとの接続確立より 7 時間後，反応あり．攻撃者がプロセス終了処理，しかし，プロセス名を間違え，正しく終了せず．1 時間後，正しい名前で終了処理を実施.
10/07		攻撃者によるコマンド操作でのプロセス終了のみ.
10/09	15:14 – 15:48	1 回目の攻撃発生．攻撃者は，実施端末だけでなく，他端末のシステム構成情報やディレクトリ情報を確認．また，実施端末に設置していたおとりファイルを窃取.
10/10		攻撃者によるコマンド操作でのプロセス終了のみ.
10/16	20:19 – 21:50	2 回目の攻撃発生．1 回目の攻撃と同様に，構成情報・ディレクトリ確認や，ファイル窃取を実施．また，端末に不正ファイルをダウンロードし，Active Directory に接続を行ってユーザ情報などの構成情報をファイル化し窃取.
10/17	10:36 – 11:02	3 回目の攻撃発生．Active Directory の構成情報やドメイン参加者を確認したほか，1 回目と同様に構成情報の確認やおとりファイルの窃取を実施.
10/18		攻撃者によるコマンド操作なし (指令サーバに一定回数以上接続を行ったら自身でプロセスを終了する仕組み).

装された実行形式の不正プログラムが含まれていた．動的活動観測環境では，不正プログラムへのパソコン感染後，約 7 時間後に攻撃者が観測環境を訪れ侵害活動を開始し，活動を停止するまでの 12 日間，攻撃者による計 3 回，3 時間の活動を通して，システム構成やディレクトリ情報の確認，感染パソコンなどからファイルの窃取等の行動を確認した (表 2).

3.2 FFRI Dataset 2016

FFRI Dataset 2016 は，株式会社 FFRI が独自に収集した計 8,243 件のマルウェアを動的解析することで得られたマルウェアの解析ログ群である．FFRI Dataset ではマルウェアの端末内での挙動に着目する．データセットの仕様については次の通りである．

3.2.1 マルウェア

マルウェアはすべて PE (Portable Executable) 形式，かつ Windows プラットフォーム上で実行可能なファイルである．2016 年 1 月から 2016 年 3 月までの期間に Web クローリング等によって広く世界中から収集された比較的新しいマルウェアであり，10 社以上のアンチウィルス製品にてマルウェアと判定されたものを選定した．収集された全数からランダムサンプリングを行っており，その内訳は収集時点におけるインターネット上のマルウェア感染のトレンドを反映していると考えられる．マルウェア検体は，当該検体を利用した評価により研究成果の現実的な有効性を確認することを目的として選定されている．なお，データセットはこれらマルウェアの動的解析結果を収録しており，当該マルウェア自体は含まない．また，FFRI Dataset 2016 には FFRI Dataset 2013–2015 も含まれている．

表 3　解析ログに含まれるデータ項目

項目	概要
`info`	解析の開始，終了時刻等
`signatures`	定義シグネチャとの照合結果
`virustotal`	VirusTotal に登録されている各アンチウィルス検出結果
`static`	マルウェアファイルの静的情報 (セクション構造，インポート API 等)
`dropped`	マルウェアが実行時に生成したファイルに関する情報
`behavior`	マルウェアが実行時に呼び出した API，引数，返り値等の情報
`processtree`	マルウェアが実行時に起動したプロセスの階層情報
`summary`	マルウェアが実行時にアクセスしたファイル，レジストリキー等の情報
`target`	解析対象となったマルウェアファイルの情報 (ファイルサイズ，ハッシュ値等)
`debug`	動的解析時の Cuckoo Sandbox のデバッグログ
`strings`	マルウェアファイルに含まれる文字列情報
`network`	マルウェアが実行時に発生した通信情報

3.2.2 動的解析

前述のマルウェアをオープンソースのマルウェア解析ツールである Cuckoo Sandbox [21] を用いて動的解析し，解析ログを生成している．

Cuckoo Sandbox は，仮想化された Windows ゲスト内にマルウェアをコピー，実行，実行時挙動の記録，ゲスト環境の復元等の一連の解析動作を自動化するソフトウェアパッケージである．マルウェアの動的解析は，ネットワーク接続を有する専用のマルウェア解析環境上に Cuckoo Sandbox による解析システムを構築し，1 検体あたり 90 秒間実行した．ゲスト OS は Windows 8.1 (x64) と Windows 10 (x64) である．FFRI Dataset 2016 では，同一解析対象ファイルを上記 2 つのゲスト OS 上で実行した際の解析ログを提供する．また，Cuckoo Sandbox は VirusTotal [22] と連携する機能を有しており，解析対象ファイルのハッシュ値に基づいて VirusTotal に問い合わせを行うことで，各アンチウィルス製品での検知状況を取得できる．本データセットの解析ログは，解析を実施した時点での当該検出状況を含んでいる．表 3 に解析ログに含まれる具体的な項目の概要をまとめる．

3.3 NICTER Darknet Dataset 2016

NICTER Darknet Dataset 2016 は，情報通信研究機構で研究開発しているインシデント分析センター NICTER [23], [24] で観測・収集したダークネット宛てのトラフィックデータである．また，NICTER Darknet 2016 には NICTER Darknet Dataset 2013–2015 も含まれている．

3.3.1 ダークネット

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 アドレス空間から構成されたネットワークの総

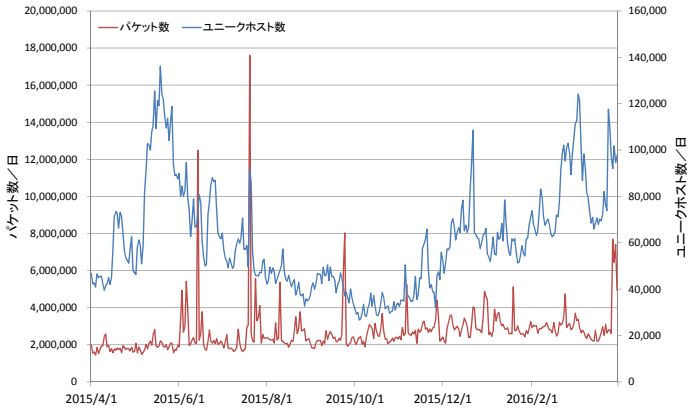

図 4 パケット数およびユニークホスト数の推移

称である．一般的なインターネット利用において，ダークネット宛にトラフィックが発生することは無いが，実際には大量のトラフィックが常時ダークネットで観測されている．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拡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) 等何らかのインターネット上における不正活動に起因したものである．このため，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，不正活動の傾向把握が可能になる．

3.3.2 ダークネットトラフィックデータ

NICTER Darknet 2016 では，NICTER 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．データセットの特徴として，観測されたトラフィックに対して一切応答を返していないため，データセットには外部からダークネット宛ての片方向のトラフィックしか含まれていない．また，観測対象のダークネットを秘匿する目的で，ダークネットトラフィックの宛先 IP アドレスについては，第 1 および第 2 オクテットの値を適当な値に置換している．

観測期間は 2011 年 4 月 1 日から 2016 年 3 月 31 日までを基本とし，2016 年 4 月 1 日以降のトラフィックデータについても後述する NONSTOP 環境を通じてリアルタイムでの提供を行っている．参考までに，提供するデータセットにおける 2015 年 4 月 1 日から 2016 年 3 月 31 日までの日毎の総パケット数とユニークホスト数 (攻撃元ホスト数) の推移を図 4 に示す．図 4 から，パケット数およびユニークホスト数ともに高い水準で推移していることがわかる．

3.3.3 NONSTOP

NICTER Darknet Dataset の提供には，NICTER で開発した NONSTOP (NICTER Open Network Security Test Out Platform) [25] を活用する．NONSTOP は各種サイバーセキュリティ情報 (ダークネットトラフィック，マルウェア検体，スパムメール，マルウェア解析結果等) を遠隔から安全に利活用するためのプラットフォームであり，いわゆる PaaS (Platform as a Service) の形態として開発が進められている．

利用を希望するユーザは，SSH クライアントとあらかじめ発行された認証用 IC カードを利用して NONSTOP へのアクセスを行い，研究内容に応じて提供される仮想マシン内で必要なサイバーセキュリティ情報にアクセスし，分析を行うことになる．そのため，分析用に独自開発したツール等はローカルから仮想マシン内へファイル転送することで仮想マシン内での実行が可能となる．また NONSTOP 内にリポジトリを用意することで，必要な各種ライブラリ等についてもインストールが可能となっている．

一方，提供したサイバーセキュリティ情報のうち外部への転送を禁止している情報の流出を防ぐ目的で，仮想マシンからローカルへのファイル転送に関しては，複数のフィルタ機構による検査，転送ファイルの一定期間の保存等を実施している．

3.4 PRACTICE (AmpPot) Dataset 2015

PRACTICE (AmpPot) Dataset 2015 は，インターネット上のオープンなサーバ (DNS, NTP 等) を踏み台にして通信を増幅させることでサービス妨害を行う分散反射型サービス妨害攻撃 (Distributed Reflection Denial-of-Service Attack; DRDoS 攻撃) を観測したデータセットである．観測には，DRDoS 攻撃観測用ハニーポットである AmpPot [26] が用いられている．当該データセットは総務省委託研究「国際連携によるサイバー攻撃予知・即応技術の研究開発 (H23–H27)」にて得られた成果の一部である．

AmpPot の構成を図 5 に示す．AmpPot は，「サーバプログラム」「アクセスコントローラ」「ハニーポットマネージャ」の三つの要素からなる．サーバプログラムは，DRDoS 攻撃の踏み台となるサービスを提供する．2016 年 6 月現在，QotD, CharGen, DNS, NTP, SNMP, SSDP の 6 種類のサービスが動作している．アクセスコントローラは，ハニーポットが実際の攻撃に加担しないよう通信を制御する．ハニーポットマネージャは，サーバプログラムとアクセスコントローラの制御，および通信ログの出力を担当する．

図 6 に AmpPot により観測される DRDoS 攻撃のうち，悪用される頻度の高い 4 つのサービス CharGen (CHG), DNS, NTP, SSDP の観測結果を示す．図 6 の縦軸は攻撃イベントの件数であり，特に 2014 年以降，攻撃イベント数が急増していることがわかる．なお，ここでは同一の攻撃対象 IP アドレスに対するパケットが 60 秒以上の間隔を空けずに 100 件以上連続して観測された場合にこれを攻撃イベントとしてカウントしている．詳細は割愛するがハッカーグループ Anonymous による OpKillingBay (捕鯨への抗議が目的とされるサイバー攻撃活動) や，サービ

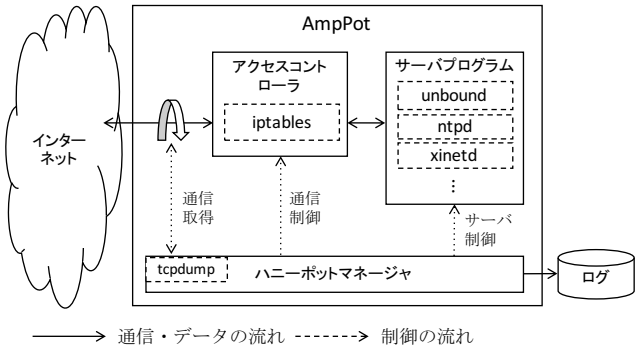

図 5 AmpPot の構成

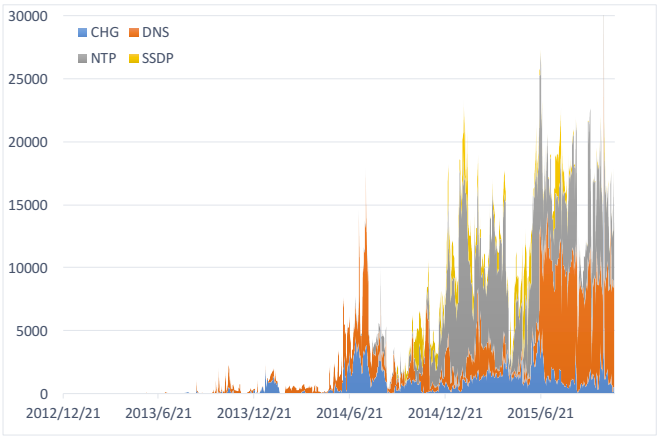

図 6 AmpPot により観測された DRDoS 攻撃件数の推移

ス妨害攻撃により脅迫を行い，金銭 (Bitcoin) を要求する DD4BC による攻撃も観測されている．

PRACTICE (AmpPot) Dataset 2015 は，2015 年 5 月 31 日から 6 月 6 日までの 1 週間に AmpPot 1 台 (日本国内に設置) によって観測されたトラヒックを記録した pcap ファイルであり，その容量は gzip による圧縮後で 2.1 GB となっている．当該ハニーポットでは CharGen, DNS, NTP, SSDP が動作しており，これらサービスを悪用した DRDoS 攻撃や踏み台を探索するためのスキャン通信等が観測されている．なお，データセットにはハニーポットへのリクエストパケットのみが含まれており，上記サービスからのレスポンスは含まれない．

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 Darknet Dataset,

表 4 MWS2008–2015 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	'08	'09	'10	'11	'12	'13	'14	'15
CCC (マルウェア検体)	5	7	6	5	7	3	3	0
CCC (攻撃通信データ)	9	14	5	6	2	–	–	2
CCC (攻撃元データ)	8	6	5	4	–	–	–	1
MARS	–	–	1	1	–	–	–	–
D3M	–	–	4	3	3	9	14	9
IIJ MITF	–	–	–	1	–	–	–	–
FFRI	–	–	–	–	–	5	2	4
PRACTICE	–	–	–	–	–	3	1	0
NICTER	–	–	–	–	–	6	2	3
Darknet								
BOS	–	–	–	–	–	–	1	4
NCD	–	–	–	–	–	–	–	0
データセット説明	0	1	1	1	0	1	0	0
合計	22	28	22	20	13	27	23	23
学生発表件数	8	15	10	9	9	10	10	14

そして新しく提供された標的型攻撃を含む BOS 等を用いた研究は増加傾向にある．また，ウェブ感染型マルウェアを含む D3M の件数は減少したものの，依然として利用件数は最も多かった．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 [27].

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 2016 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 2016 (MWS2016)”, http://www.iwsec.org/mws/2016/
[2] 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.
[3] 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.
[4] 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.
[5] 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.
[6] 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, Vol.114, No.118, CSEC, pp.125–131, 2014 年 7 月.
[7] 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, Vol.115, No.121, CSEC, pp.37–44, 2015 年 7 月.
[8] “CAIDA Data - Overview of Datasets, Monitors, and Reports,” http://www.caida.org/data/overview/
[9] “MAWILab,” http://www.fukuda-lab.org/mawilab/
[10] “The Information Marketplace for Policy and Analysis of Cyber-risk & Trust (IMPACT),” https://www.impactcybertrust.org/
[11] “MALICIA Project,” http://malicia-project.com/dataset.html
[12] “Malware-Traffic-Analysis.net,” http://www.malware-traffic-analysis.net/
[13] “Contagio Malware Dump,” http://contagiodump.blogspot.jp
[14] “Android Malware Genome Project,” http://www.malgenomeproject.org
[15] “The ACODE dataset,” http://nsl.cs.waseda.ac.jp/projects/acode/
[16] “International Secure Systems Lab,” http://www.iseclab.org
[17] “DARPA Intrusion Detection Data Sets,” http://www.ll.mit.edu/mission/communications/cyber/CSTcorpora/ideval/data/
[18] “Traffic Data from Kyoto University’s Honeypots,” http://www.takakura.com/Kyoto_data/
[19] 寺田 真敏, 青木 翔, 楠美 淳弥, 重本 倫宏, 萩原 健太, “研究用データセット「動的活動観測 2014」の検討”, 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 2014 (MWS2014), 2014 年 10 月.
[20] 寺田 真敏, 堀 健太郎, 成島 佳孝, 吉野 龍平, 萩原 健太, “研究用データセット「動的活動観測 2015」の検討”, 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 2015 (MWS2015), 2015 年 10 月.
[21] “Cuckoo Sandbox,” http://www.cuckoosandbox.org/
[22] “VirusTotal,” https://www.virustotal.com/
[23] K. Nakao, K. Yoshioka, D. Inoue, M. Eto, and K. Rikitake, “nicter: An Incident Analysis System using Correlation between Network Monitoring and Malware Analysis,” In Proceedings of the First Joint Workshop on Information Security (JWIS 2006), September 2006.
[24] 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,” In WOMBAT Workshop on Information Security Threats Data Collection and Sharing, pp.58–66, 2008.
[25] 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.
[26] 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.
[27] “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