Vol.2015-CSEC-70 No.6
Vol.2015-SPT-14 No.6
2015/7/2

マルウェア対策のための研究用データセット ～MWS Datasets 2015～

神薗 雅紀[†1]　秋山 満昭[†2]　笠間 貴弘[†3]
村上 純一[†4]　畑田 充弘[†5]　寺田 真敏[†6]

マルウェアの脅威に対し様々なアプローチで研究が行われているが，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，必要となる情報を収集して客観的な評価と研究成果の共有を容易にするためのデータセット（MWS Datasets 2015）を作成した．本稿では，MWS Datasets 2015 を構成する BOS 2015，D3M 2015，FFRI Dataset 2015，NICTER Darknet Dataset 2015，NCD in MWS Cup 2014 および継続的に提供される CCC DATAset および PRACTICE Dataset の概要を報告する．

Datasets for Anti-Malware Research ～MWS Datasets 2015～

Masaki KAMIZONO[†1]　Mitsuaki AKIYAMA[†2]　Takahiro KASAMA[†3]
Junichi MURAKAMI[†4]　Mitsuhiro HATADA[†5]　Masato TERADA[†6]

Many security researches 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 dataset for security research. For such a problem, anti-Malware engineering WorkShop (MWS) collected data related with malware threats and made the datasets (MWS Dataset 2014) for studies to evaluate the proposals and share the research achievements. In this paper, we introduce an overview of MWS Datasets 2014 comprised of BOS 2015, D3M 2015, FFRI Dataset 2015, NICTER Darknet Dataset 2015 and NCD in MWS Cup 2014. CCC DATAset and PRACTICE Dataset are additionally contained in the datasets.

1. はじめに

高度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，“共通の研究素材がないこと”および“研究素材の収集の困難さ”が近年のマルウェア対策研究を推進する上での阻害要因となっている．

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などのことを指し，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作成し，個々の技術の有効性や妥当性を評価してきた．このため，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に，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．

もう一つの阻害要因は，研究素材そのものが収集困難になってきていることである．攻撃者は検知回避手法や解析妨害手法を用いてサイバー攻撃やマルウェアを用いた攻撃を行い，またそれが年々高度化しているためである．例えば，ドライブバイダウンロード攻撃を行う Web 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ず，その結果として定性的にも定量的にも研究素材としての収集が難しくなっている．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，研究用データ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．

†1 プライスウォーターハウスクーパース株式会社
PricewaterhouseCoopers Co., Ltd.
†2 日本電信電話株式会社，NTT セキュアプラットフォーム研究所
NTT Secure Platform Laboratories, Nippon Telegraph and Telephon Communication Corporation
†3 国立研究開発法人 情報通信研究機構
National Institute of Information and Communications Technology
†4 株式会社 FFRI
FFRI, Inc.
†5 エヌ・ティ・ティ・コミュニケーションズ株式会社
NTT Communications Corporation
†6 株式会社日立製作所
Hitachi, Ltd
a) masaki.kamizono@jp.pwc.com

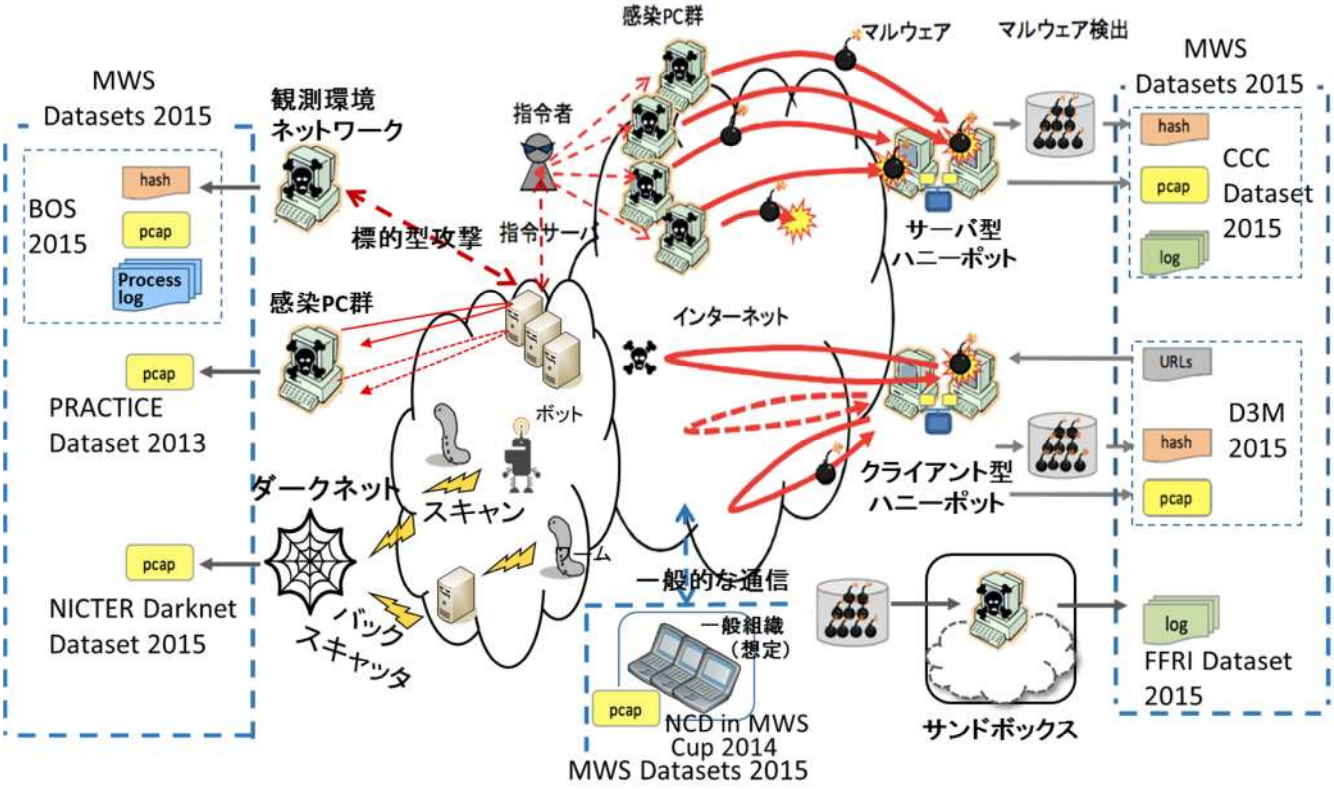

図 2　MWS Datasets 2015 の概要

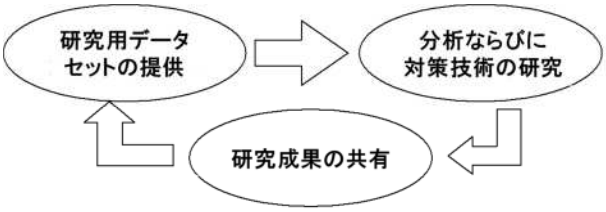

図 1 マルウェア対策研究サイクル

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 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，研究活動の推進を行ってきた．具体的な活動として，当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．今後，より研究を発展させるため，研究用データセット自身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 の活動の一環である研究用データセットについて報告する．2015 年は下記のデータセットから構成される MWS Datasets 2015（図 2）を作成し，MWS2015[1]を開催する.

（1） BOS 2015
標的型攻撃観測データ

（2） D3M 2015
Web 感染型マルウェアの観測データ

（3） NICTER Darknet Dataset 2015
ダークネットパケットデータ

（4） FFRI Dataset 2015
マルウェアの動的解析データ

（5） 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

（6） CCC DATAset
待受型ハニーポットで収集したデータ

（7） PRACTICE Dataset
マルウェアの長期観測データ
（2015 年はデータセットの内容更新は無く，2013 年およびそれ以前のデータセットが提供される）

2 章にて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し， 3 章で MWS Datasets2015 の各データセットの概要を述べる．なお，CCC DATAset および PRACTICE Dataset に関しては文献[2][3][4][5]を参照して欲しい。

2. 関連研究

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 [6]
ネットワーク運用に関わる通信ログのデータセット
・MAWILab [7]
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
・PREDICT Dataset [8]
ネットワーク運用およびセキュリティ装置に関わる通信ログのデータセット
・MALICIA Dataset [9]
ドライブバイダウンロード攻撃行う悪性 Web サイトから入手したマルウェア検体のデータセット
・Honeynet Project hpfeeds/hpfriends [10]
各種ハニーポット，サンドボックスの解析ログを集めたデータセット
・Contagio Malware Dump [11]
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 [12]
マルウェアファミリ毎に分類された Android マルウェア検体

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問題点について考察する.

(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 PREDICT は米国の政府（国土安全保障省，DHS）や米国の大学が主体，iSecLab [13]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)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，つまり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[14]）は 1998 年から 2000 年に作成されたものである).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.

(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 PREDICT 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 2015

本章では，MWS Datasets2015 の各データセットの概要を述べる.

3.1 BOS 2015

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[15].

BOS 2015 では BOS 2014 も含め，計 6 つの攻撃事例が提供される．本節では，BOS 2014 で提供されている 2 つの攻撃事例を示す.

Vol.2015-CSEC-70 No.6
Vol.2015-SPT-14 No.6
2015/7/2

(1) 目的

これまで，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ものであった．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，これら機能のいずれを使ったのか，どの順番で使ったのかなど，攻撃者の行動という視点で把握や解析することはなかった．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 2015 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を特徴付けできる研究用データセットとなっている．

(2) 観測環境

動的活動観測環境は，実インターネット上の攻撃者が試みる組織内ネットワークへの侵害活動を観測するシステムで，システムそのものが組織内ネットワークを模擬している(図 3)．クライアントは，電子メールに添付された検体を実行する PC 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，インターネットとの接続性を持つことができる．

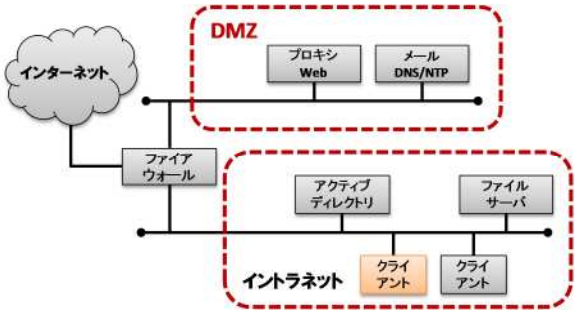

図 3 動的活動観測環境の概要図

(3) データセット構成

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．

(a) 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．

(b) 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であり，攻撃者の行動に関する解析が可能である．

(c) 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．

(4) 観測事例

BOS に含まれる 2 つの事例 Case11、Case21 について述べる．Case11 のマルウェア検体は，exe ファイルであり，Microsoft Word ファイルのアイコンで偽装されていた．実行後の観測事象を表 1 に示す. Case21 のマルウェア検体は，exe ファイルであり，フォルダアイコンで偽装されていた．実行後の観測事象を表 2 に示す．なお，攻撃者の行動視点に注目するために，操作者を併記した時系列イベントの形で観測事象を記載する．

表 1：マルウェア検体 Case11 での観測事象

時刻	操作者	観測事象
11:06	O	C:¥data 直下にて，マルウェア検体(exe ファイル)をダブルクリック実行
11:06	M	C:¥windows¥FlashHelpx64.exe をドロップ
11:06	M	自動起動を目的としたレジストリ改変
11:06	M	**.**.160.125 との接続を確立
11:15	A	スクリーンキャプチャの取得
11:15	A	端末基本情報の取得
11:15	A	スクリーンキャプチャの取得
11:40	A	スクリーンキャプチャの取得
11:40	A	C:¥RECYCLER¥に a.exe をアップロード
11:41	A	cmd.exe からコマンド「a /stext aaa.txt」経由で a.exe を実行
11:41	A	コマンド「del a.*」で a.exe を削除
11:41	M	**.**.160.125 との接続を解除
11:41	M	プロセスが終了

表 2 マルウェア検体 Case21 での観測事象

時刻	操作者	観測事象
15:06	O	デスクトップ上でマルウェア検体(exe ファイル)をダブルクリック実行
15:06	M	www.google**.com/windowsxp/Snews.asp に対して HTTP POST 要求を送信
15:06	M	HTTP POST 応答「HTTP/1.0 200 OK」を受信，以降継続
15:07	M	検体のプロセス lplus.exe が起動した cmd.exe で，コマンド「net start」，「tasklist」，「systeminfo」，「netstat -an」などを実行
16:29	A	検体のプロセス lplus.exe が起動した cmd.exe で，「arp -a」を実行
16:38	A	検体のプロセス lplus.exe が起動した cmd.exe で，「at ¥¥I160***」を実行
17:06	A	検体のプロセス lplus.exe が起動した cmd.exe で，「net group "domain computers" /domain」を実行
17:19	A	検体のプロセス lplus.exe が起動した cmd.exe で，「ping 10.*.***.1」を実行 (10.*.***.1 は共有フォルダを提供するファイルサーバ)
17:49	A	C:¥WINDOWS¥Debug¥Rar.exe で ¥¥10.*.***.1¥public¥mail ¥testMail にアクセス
17:49	A	C:¥WINDOWS¥Debug ¥Rar.exe で ¥¥10.*.***.1 ¥public¥012 営業本部¥顧客先アドレス**.zip にアクセス
17:55	A	検体のプロセス lplus.exe が起動した cmd.exe で，「ftp -s:c¥windows¥debug¥ftpo.txt」を実行

[操作者] O：観測者，M：マルウェア検体，A：攻撃者

3.2 D3M 2015

D3M（Drive-by-Download Data by Marionette）2015 は NTT セキュアプラットフォーム研究所の高対話型の Web クライアント型ハニーポット（Marionette [16][17]）で収集したドライブバイダウンロード攻撃に関連するデータである．ドライブバイダウンロード攻撃は，マルウェアの主要な感染経路の一つで Web ブラウザおよびそのプラグインの脆弱性を利用して制御を奪い，マルウェアを強制的にダウンロードおよびインストールさせる攻撃である．

D3M はドライブバイダウンロード攻撃に関する，攻撃通信データ，マルウェア検体，およびその通信データを収録した Web 感染型マルウェアの観測データ群から構成している．Marionette は脆弱性に対する攻撃を受けるが，ダウンロードされたマルウェアの実行は許可しない．このため取得したマルウェアを 24 時間以内に動的解析システム（BotnetWatcher [18]）にて解析することでデータセットに必要な情報を収集する．なお、ハニーポットで観測されたデータの一部を D3M 2015 として提供している。

D3M 2015 は感染手法の検知ならびに解析技術の研究のための“攻撃通信データ”，マルウェアの解析技術のための“マルウェア検体（ハッシュ値）”，および“マルウェア通信データ”から構成される．データセットの取得環境を図 4 に示す．

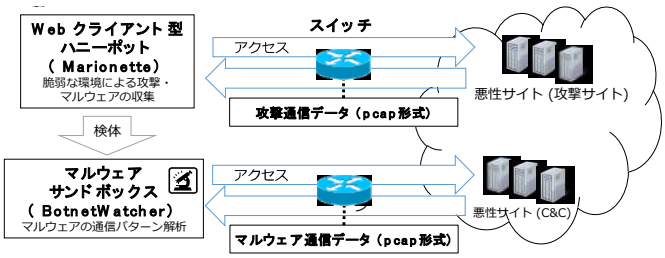

図 4 D3M の取得環境

なお，過去のデータとの傾向を比較分析することができるよう，D3M 2010 - 2014 も提供している．それぞれのデータの概要は次の通りである．

(1) 攻撃通信データ

Web クライアント型ハニーポットの通信を tcpdump でパケットキャプチャした PCAP 形式のファイルである．ハニーポットの OS は Windows XP，Web ブラウザは Internet Explorer, プラグインは Adobe Reader, Flash Player, WinZip, QuickTime，Java をあらかじめインストールしてあり，何れも脆弱性を含むバージョンでありセキュリティパッチは未適用である．ハニーポットはインターネット接続されており，パケットキャプチャは上流のネットワーク装置で行っている．

データ収集日は 2014 年 4 月 11 日, 2014 年 5 月 2 日, 2014 年 8 月 2 日，2014 年 12 月 1 日，2015 年 2 月 8 日，および 2014 年 2 月 19 日であり, 日毎に 1 ファイル, 計 6 ファイルである．

攻撃通信データは，あるブラックリストに登録されている URL を巡回し，攻撃を検知した URL を再度巡回した際の通信である．このため，攻撃コードが含まれる可能性が高く，かつ雑音（正常な Web サイトとの通信など）が少ない通信データになっている．巡回時に Web ブラウザに入力した URL を参考情報として付与している．なお, 入力 URL から派生してアクセスされる URL（リダイレクト，スクリプト読み込み, 画像読み込み）は通信データに含まれるが，前述の URL リストには含まれない．

(2) マルウェア検体

Web クライアント型ハニーポットで収集した Web 感染型マルウェア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．項番(1)で収集した攻撃通信データに含まれる検体である．

(3) マルウェア通信データ

項番(1)で収集した検体を 24 時間以内に動的解析システムで解析した際の通信のフルキャプチャデータである．動的解析システムはインターネットに接続した環境でマルウェアを 10 分間動作させており, ボットなどの遠隔制御されるマルウェアの動的解析が可能である．なお，外部ホストやネットワークに対する攻撃は動的解析システム内の仮想インターネット環境に転送することで，解析時の安全性を確保している．

3.3 NICTER Darknet Dataset 2015

NICTER Darknet Dataset 2015 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 [19][20] 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を提供する．NICTER Darknet 2015 では NICTER Darknet Dataset 2013 - 2014 も提供する．

(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）など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．

(2) ダークネットトラフィックデータ

NICTER Darknet 2015 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 オクテットの値を提供な値に置換している．

観測期間は 2011 年 4 月 1 日から 2015 年 3 月 31 日を基本とし，2015 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 2014 年 4 月 1 日から 2015 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 5 に示す．図 5 を見ると，パケット数およびユニークホスト数ともに増加傾向にあることがわかり，昨年度までも観測されていた DRDoS 攻撃のための探索活動や，最近活発になっている組込み機器を狙った Telnet（23/TCP）宛てのスキャンなど攻撃活動が活発化していることがわかる．

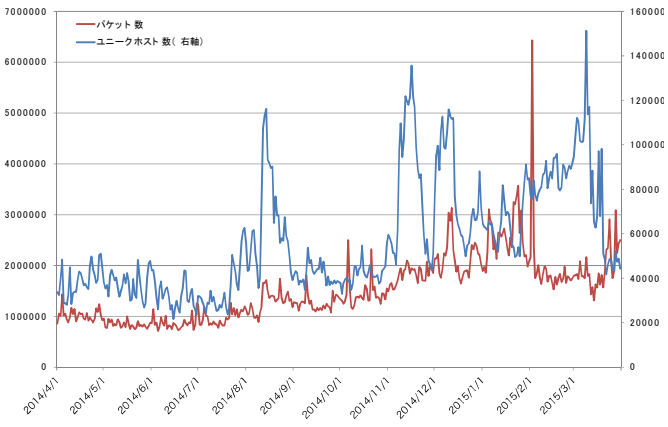

図 5 パケット数およびユニークホスト数の推移

(3) NONSTOP

NICTER Datknet Dataset の提供には，NICTER で開発した NONSTOP（NICTER Open Network Security Test Out Platform）[21]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など）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である．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インストール可能としている．

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などを実施している．

3.4 FFRI Dataset 2015

FFRI Dataset 2015 は株式会社 FFRI が独自に収集した計 3,000 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 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．

(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．

2015 年 1 月から 2015 年 4 月の期間に 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，2015 年 5 月時点で，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ていたものを選定した．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のマルウェアの感染トレンドを反映していると考えられる．

当該マルウェア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の結果であり，当該マルウェア自体は含まない．また，FFRI Dataset 2015 には FFRI Dataset 2013 - 2014 の全データが含まれている．

(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 [22]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，ゲスト環境の復元など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 120 秒間実行した．ゲスト OS は Windows 8.1(x64)である．

また Cuckoo Sandbox は VirusTotal [23]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, 当該時点での各アンチウィルス製品での検知状況を取得することができる．本データセットの解析ログは, 解析を実施した 2014 年 5 月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．

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
yara	yara [24]の有する標準ルールセットとの照合結果
signatures	ユーザ定義シグネチャとの照合結果（未使用）
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

3.5 NCD in MWS Cup 2014

NCD in MWS Cup 2014（Normal Communication Dataset in MWS Cup 2014）は，MWS 実行委員および組織委員にて作成したデータセットである．

既存の MWS Datasets はマルウェアに起因する様々な攻撃を網羅しており，必然的に悪性なデータセットとなっている．主に悪性なものを検知する手法の評価として利用されている。一方，検知手法の False Positive を評価する場合のデータ（以降、ノーマル通信データ）は，研究者が独自にデータを作成し評価していた．False Positive を評価する上で，ノーマル通信データは、悪性なものではないインターネット環境に接続した一般的な組織の通信が望ましい．しかし，実際の組織の通信データを提供することはプライバシーの問題等，様々な阻害要因に阻まれ実現することは難しい状況にある. そこで, NCD in MWS Cup 2014 は MWS の活動の一環として MWS Cup 2014 参加者に協力してもらい，参加者の同意のもと，競技中の通信データを収集することで作成したノーマル通信データとなっている．MWS Cup 2014 会場に単一の HUB に接続する形で複数の Wireless LAN Access Point を設置し, Cup 参加者に AP を利用してもらい, HUB 経由でノーマル通信データに使用するデータを取得している．

4. MWS Datasets 利用状況

MWS Datasets を利用し，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おり, 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．

表 4 MWS Datasets を用いた MWS での論文発表数
（MWS2008 - 2014 まで）

MWS Datasets	'08	'09	'10	'11	'12	'13	'14
CCC (マルウェア検体)	5	7	6	5	7	3	3
CCC (攻撃通信データ)	9	14	5	6	2	0	－
CCC (攻撃元データ)	8	6	5	4	0	0	－
MARS	－	－	1	1	0	－	－
D3M	－	－	4	3	3	9	14
IIJ MITF	－	－	－	1	－	－	－
FFRI	－	－	－	－	－	5	2
PRACTICE	－	－	－	－	－	3	1
NICTER Darknet	－	－	－	－	－	6	2
データセット説明	0	1	1	1	0	1	0
合計	22	28	22	20	13	25	21
():学生発表件数	(8)	(15)	(10)	(9)	(9)	(10)	(10)

一部，複数のデータセットを利用した論文あり．“ - ”は提供なし

CCC DATAset 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，Web 感染型マルウェアを含むデータセットである D3M や FFRI Dataset, 昨年から急増している DR-DoS 攻撃を含む NICTER Darknet Dataset などを用いた研究が増加傾向にある．実際のサイバー攻撃における攻撃やマルウェアのトレンドの変化に伴い，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 としては，このような攻撃手法やマルウェアのトレンドの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[25].

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、その研究成果の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 2015 についてその概要を述べた．これら研究用データセット自体が研究者間で共通言語として役割を担うことや，研究用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，研究用データセット自身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の構築および提供など, 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致します.

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 2015（MWS2015）http://www.iwsec.org/mws/2015/
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,CSS2009(MWS2009) (2009.10)
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～, CSS2010(MWS2010) (2010.10)
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～,CSS2011(MWS2011) (2011.10)
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～,CSS2003(MWS2013) (2013.10)
6) CAIDA Data - Overview of Datasets, Monitors, and Reports,http://www.caida.org/data/overview/
7) MAWILab, http://www.fukuda-lab.org/mawilab/
8) PREDICT Dataset Catalog, https://www.predict.org/Default.aspx?tabid=104
9) MALICIA Project, http://malicia-project.com/dataset.html
10) hpfriends, http://hpfeeds.honeycloud.net/¥#/home
11) Contagio Malware Dump, http://contagiodump.blogspot.jp
12) Android Malware Genome Project, http://www.malgenomeproject.org
13) International Secure Systems Lab, http://www.iseclab.org
14) DARPA Intrusion Detection Data Sets, http://www.ll.mit.edu/mission/communications/cyber/CSTcorpora/ideval/data/
15) 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, CSS2014,2014 年 10 月
16) Mitsuaki Akiyama, Kazufumi Aoki, Yuhei Kawakoya, Makoto Iwamura, and Mitsutaka Itoh:Design and Implementation of High Interaction Client Honeypot for Drive-by-download Attacks,IEICE Transaction on Communication, Vol.E93-B No.5 pp.1131-1139 (2010.05)
17) Mitsuaki Akiyama, Takeshi Yagi, Youki Kadobayashi, Takeo Hariu, and Suguru Yamaguchi,“Client Honeypot Multiplication with High Performance and Precise Detection,” IEICE Trans., vol.E98–D, no.4, pp. 775–787, 2015.
18) Kazufumi Aoki, Takeshi Yagi, Makoto Iwamura, and Mitsutaka Itoh:Controlling malware HTTP communication in dynamic analysis system using search engine,The 3rd International Workshop on Cyberspace Safety and Security (CSS2011)
19) 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.
20) 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.
21)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
22) Cuckoo Sandbox: Automated Malware Analysis,http://www.cuckoosandbox.org/
23) VirusTotal - Free Online Virus, Malware and URL Scanner,https://www.virustotal.com/ja/
24) yara-project - A malware identification and classification tool, https://code.google.com/p/yara-project/
25) 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