マルウェア対策のための研究用データセット ～ MWS 2011 Datasets ～

畑田 充弘[†1]　中津留 勇[†2]　秋山 満昭[†3]

[†1] NTT コミュニケーションズ株式会社
〒108-8118 東京都港区芝浦 3-4-1 グランパークタワー17F
[†2] 一般社団法人 JPCERT コーディネーションセンター
〒101-0054 東京都千代田区神田錦町 3-17 廣瀬ビル 11F
[†3] NTT 情報流通プラットフォーム研究所
〒180-8585 東京都武蔵野市緑町 3-9-11

E-mail: [†1] m.hatada@ntt.com, [†2] office@jpcert.or.jp, [†3] akiyama.mitsuaki@lab.ntt.co.jp

あらまし マルウェアによる脅威が複雑化する中，様々な対策研究が盛んに行われている．客観的な評価と研究成果の共有を容易にするため，サイバークリーンセンターで収集しているデータをもとに研究用データセット(CCC DATAset 2008/2009/2010)を利用したマルウェア対策研究人材育成ワークショップ(MWS 2008/2009/2010)を開催してきた．本稿では，MWS 2011 で利用する研究用データセット(MWS 2011 Datasets)を構成する CCC DATAset 2011，Web 感染型マルウェアのデータセット(D3M 2011)の概要を報告する．

Datasets for Anti-Malware Research ～ MWS 2010 Datasets ～

Mitsuhiro Hatada[†1]　You Nakatsuru[†2]　Mitsuaki Akiyama[†3]

[†1] NTT Communications Corporation
Gran Park Tower 17F, 3-4-1 Shibaura, Minato-ku, Tokyo 108-8118, Japan
[†2] Japan Computer Emergency Response Team Coordination Center
3-17 Kandanishikicho, Chiyoda-ku, Tokyo 101-0054, Japan
[†3] NTT Information Sharing Platform Laboratories
Midori-Cho 3-9-11, Musashino, Tokyo 180-8585, Japan

E-mail: [†1] m.hatada@ntt.com, [†2] office@jpcert.or.jp, [†3] akiyama.mitsuaki@lab.ntt.co.jp

Abstract There has been a lot of researches on countermeasures against the complicated threats by malware. anti-Malware engineering WorkShop (MWS) were held annually (2008 - 2010) in order to evaluate the proposals objectively and share the research achievements by using CCC DATAset 2008 - 2010. This paper presents an overview of MWS 2011 Datasets for MWS 2011: CCC DATAset 2011 and D3M 2011.

1. はじめに

サイバー攻撃への対応が国家及び企業レベルで求められる中，依然としてマルウェアは進化を続け，脅威は複雑化しており，万全な対策を講じることは不可能といっても過言ではない状況が続いている．マルウェア対策という分野だけを見ても，様々な研究が盛んに行われているが，研究を行う上で様々な課題があり，その一つとして「共通の教材がないこと」が挙げられる．ここでの教材とは，提案手法の評価に用いるマルウェアのサンプルや，感染前後の通信データなどのことであり，マルウェアの進化に合わせて適切に選択されたものであることが望ましい．教材となるこのような研究用データは，研究者らが独自にハニーポットを設置して収集し，各々の解析手法や対策手法の妥当性を評価してきた．そのため，同じテーマに取り組む研究者同士であっても，研究成果を単純に比較することが難しい．新たに研究を始めようとしても，所属組織のセキュリティポリシーによっては「研究用データを収集すること自体が難しくなっていること」も大きな課題である．

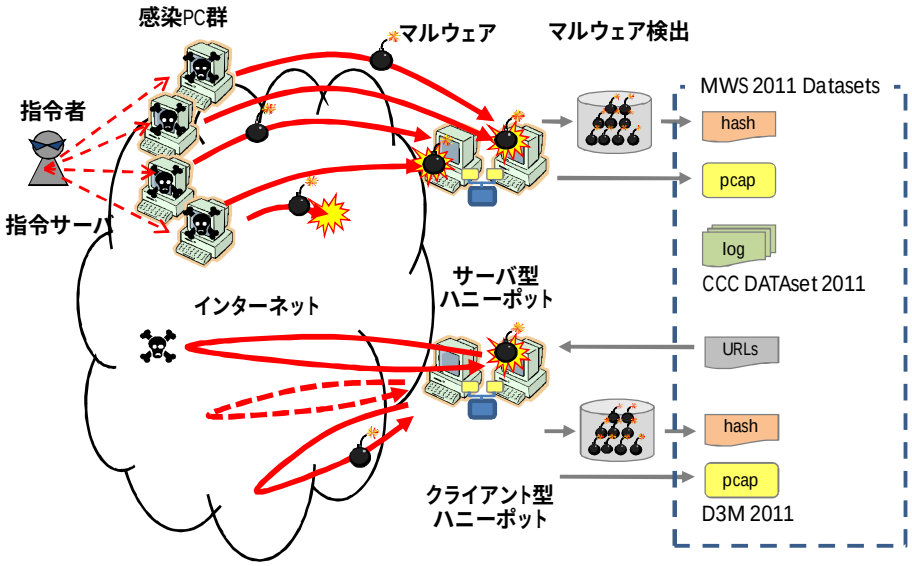

図 1 MWS 2011 Datasets の概要

現在でも侵入検知システムの評価に用いられる DARPA Intrusion Detection Evaluation Data Sets[1]は，2000 年のデータセットが最新であり，the 2009 Inter-Service Academy Cyber Defense Exercise datasets[2]は，サイバー防御演習時のデータセットでありマルウェアによる攻撃を想定したものではない．一方で，大規模セキュリティ関連データの収集と分析をもとに，より良いデータとナレッジの共有を図る BADGERS2011[3]や，コンピュータ・ネットワークの運用データをレポジトリとして蓄積し，インフラ防護と脅威評価に活用する PREDICT[4]などのプロジェクトもある．

著者らは，サイバークリーンセンター(CCC) [5]での収集データを活用した研究用データセット：CCC DATAset 2010 とともに，研究者コミュニティーが収集した動的解析データ MARS や，Web 感染型マルウェアの観測データ D3M を含む MWS 2010 Datasets[6]を研究者に提供して，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 2010(MWS2010)」を開催した．MWS は 2008 年から年次開催しており，各データセットを利用した発表件数は
表 1 の通りである．データセット全般を総括する発表 1 件も含めて，毎年 20 件以上の研究発表が行われ，約半数は学生による発表となっている．研究発表と合わせて，パネルディスカッションや研究用データセットを用いた解析コンテスト MWS Cup 2009/2010 を通して，大学，研究機関，企業の垣根を越えた活発な議論を行ってきた．今後もマルウェア対策のための研究用データセットそのものをテーマとする研究活動や前述の海外における取り組み等との連携により，サイバー攻撃への対策を進めていく必要がある．

表 1 データセット毎の発表件数

データセット		2008 年	2009 年	2010 年
CCC DATASet	マルウェア検体	5	7	6
	攻撃通信データ	9	14	5
	攻撃元データ	8	6	5
MARS				1
D3M				4
総括			1	1
合計		22(8)	28(15)	22(10)

本年開催する MWS2011[7]で利用する研究用データセット (MWS 2011 Datasets) は，図 1 に示すように，CCC DATAset 2011 と Web 感染型マルウェアのデータセットである D3M 2011 (Drive-by-Download Data by Marionette 2011) から構成され，以下，各章で概要を述べる．

2. CCC DATAset 2011

マルウェアの解析技術の研究のための「マルウェア検体」，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」，ボットの活動傾向把握技術の研究のための「攻撃元データ」の

三つから構成される．以下，概要を述べる　．

2.1.　マルウェア検体

ハニーポットで収集したマルウェア検体のハッシュ値（MD5，SHA1）50 個をテキスト形式で記載したファイルであり，以下の観点で選定している．

(1) 解析結果を照合できる検体：10 検体

(2) 未知検体：40 検体

(1)は特徴的な機能を有し，技術的に目を通しておきたい検体であり，事前に静的解析が完了している．そのため，解析精度の評価に活用することを考慮した要件に対応する検体である．具体的には，国内外で流行している検体と同様の機能（FTP アカウント窃取，ドメイン名自動生成，スクリプトボットなど）を持つ検体や，独自かつ高度な通信および本体の難読化機能を有する検体である．(2)は 2011 年 1 月に収集した未知検体のうち，収集日が偏らないよう任意で選定した検体であり，相当数の検体の自動解析や自動分類を考慮した要件に対応する検体である．なお，対象となるマルウェア検体は，以降の攻撃通信データ及び攻撃元データに一部含まれる検体である．

2.2.　攻撃通信データ

ハニーポットの通信を tcpdump でパケットキャプチャした libpcap 形式のファイルである．ハニーポットは，ホスト OS 上の 2 台（honey001，honey002）のゲスト OS がそれぞれインターネット接続されており，パケットキャプチャはホスト OS 上で行っている．ゲスト OS は，2 台とも Windows XP SP1 であり，ゲスト OS は定期的にクリーンな状態にリセットされる．データ収集日は 2010 年 8 月 18 日から 8 月 31 日と 2011 年 1 月 18 日から 1 月 31 日，総パケット数が 23,009,309 パケット，約 3.8GB のデータサイズである．

2.3.　攻撃元データ

2010 年 5 月 1 日から 2011 年 1 月 31 日までの 9 ヶ月間にハニーポットで記録したマルウェア取得時のログで，表 2 に示す項目を 1 レコードとして記録した csv 形式のファイルである．Windows2000 が稼働するハニーポットも一部含み，国内の複数の ISP にそれぞれ接続された 72 台のハニーポットで記録された約 22MB のデータである．攻撃元データの基本情報を表 3 に示す．

マルウェア検体のダウンロードを開始した時刻がマルウェア検体の取得時刻であり，ゲスト OS の Windows 上でのファイル作成日時となる．送信元 IP アドレスまたは宛先 IP アドレスにおいて，ハニーポットの IP アドレスは各ハニーポットに対応する ID（honey001 等）に置換されて記載されている．ウイルス名称は収集日の翌日午前 3 時の最新パターンファイルを適用したウイルススキャナ（トレンドマイクロ社製）により判定された名称であり，マルウェアとして判定されなかったものは UNKNOWN と表記される．このため，パターンファイルのウイルス名称が更新された場合，同一のハッシュ値であっても，異なるウイルス名称が付与される場合がある．

MWS 2011 では，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 CCC DATAset 2008-2010 も参考情報として提供しており，それらの差異を表 4 にまとめる．攻撃通信データの総パケット数は 2010 と 2011 で同程度ではあるものの，収集日は 1 週間と 4 週間と大きく異なる．また，攻撃元データにおいては，2010 と 2011 で収集期間

表 2 攻撃元データのログ項目と例

ログ項目	例（一部を*でマスク）
マルウェア検体の取得時刻	2011-01-14 18:20:01
送信元 IP アドレス	honey016
送信元ポート番号	1029
宛先 IP アドレス	**.179.100
宛先ポート番号	20000
TCP または UDP	TCP
マルウェア検体のハッシュ値（SHA1）	*****6b8124247f988f96725066d3752ef018549
ウイルス名称	Mal_DLDER
ファイル名	C:\WINNT\system32\fewh.exe

表 3 攻撃元データの基本情報

項目	件数
全レコード数	158,734
TCP によるダウンロードレコード数	136,251
UDP によるダウンロードレコード数	22,483
ダウンロードホスト IP アドレス種類数	89,122
マルウェア検体のハッシュ値種類数	12,591
ウイルス名称種類数（UNKNOWN 含まない）	316

表 4 CCC DATAset 2008／2009／2010 の差異比較

項目		2008	2009	2010	2011
マルウェア検体					
	検体数	1	10	50	50
	選定条件	多機能，解読困難	解析結果あり，関連性のある複数検体，特徴的な機能	解析結果あり，特徴的な機能，2010 年 1 月～3 月に収集した未知検体	解析結果あり，特徴的な機能，2011 年 1 月に収集した未知検体
攻撃通信データ					
	ハニーポット	honey001,honey002	honey003, honey004	honey001,honey002	honey001,honey002
	収集日	2008/4/28, 2008/4/29	2009/3/13, 2009/3/14	2010/3/5 ～2010/3/11	2010/8/18 ～2010/8/31, 2011/1/18 ～2011/1/31
	総パケット数	15,901,943	3,511,850	22,486,674	23,009,309
攻撃元データ					
	ﾊﾆｰﾎﾟｯﾄ数	112 台	94 台	92 台	72 台
	ﾊﾆｰﾎﾟｯﾄ ID	なし（ﾀﾞｳﾝﾛｰﾄﾞﾎｽﾄと通信方向のみ）	あり	あり	あり
	収集期間	2007/11/1 ～2008/4/30	2008/5/1 ～2009/4/30	2009/5/1 ～2010/4/30	2010/5/1 ～2011/1/31
	全レコード数	2,942,221	2,470,766	1,162,093	158,734

は 1 年間と 8 ヶ月間という差があるものの，全レコード数は大幅に減少している．これらは CCC のハニーポット運用上，収集期間が短いという差異以外にも，少なくとも当該ハニーポットに対する攻撃が減少していることを意味している．このことは CCC の活動の成果によるボット感染者の減少や，一部のマルウェアで利用されているハニーポットリストに CCC のハニーポットの一部が IP アドレスブロックごと掲載されていることにより攻撃の対象外となっていること，などが考えられる．

3. D3M 2011

D3M 2011 は，NTT 情報流通プラットフォーム研究所の高対話型の Web クライアントハニーポット(Marionette[8])で収集したマルウェア検体，攻撃通信データの 2 つを収録した Web 感染型マルウェアの観測データ群である．

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を許可しない．そのため，CCC DATAset の攻撃通信データとは異なり，感染後のマルウェアの通信挙動は D3M 2011 の攻撃通信データには含まれない．

CCC DATAset はいわゆるサーバ型ハニーポットで収集したデータであり，近年脅威となっている Web ブラウザ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及びインストールさせる Drive-by-download 攻撃を捉えた研究用データセットへの必要性から提供することとなった．

D3M 2011 は，マルウェアの解析技術の研究のための「マルウェア検体」，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」から構成される．以下，それぞれについて概要を述べる ．

3.1. マルウェア検体

Web クライアント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（34 検体分）をテキスト形式で記載したファイルである．2011 年 2 月 8, 14, 16 日に収集した検体であり，攻撃通信データ

には含まれる検体である．

3.2. 攻撃通信データ

Web クライアントハニーポット 10 台の通信を tcpdump でパケットキャプチャした libpcap 形式のファイルである．ハニーポットの OS は Windows XP SP2，ブラウザは Internet Explorer 6.0，プラグインが Adobe Reader，Flash Player，WinZip，QuickTime，JRE であり，何れもセキュリティパッチは未適用である．10 台それぞれが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ネットワークにあるスイッチのミラーポートで行っている．データ収集日は 2011 年 2 月 8，14，16 日であり，日毎に 1 ファイル，計 3 ファイルで約 33MB である．

巡回対象 URL は公開されているブラックリスト（malwaredomainlist.com）に登録されている URL の中から，各データ収集日に攻撃を検知した URL を予め抽出したものを用いており，参考情報として D3M2011 とともに提供している．各収集日においてアクセスした URL は同一とは限らず，また，入力 URL から派生する URL（リダイレクト，スクリプト読み込み，画像読み込みなど）は記載されていない．

CCC DATAset と同様に D3M においても，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 も提供している．

4. おわりに

マルウェア対策研究を進める上での重要な課題である「共通の教材がないこと」と「研究用データを収集すること自体が難しくなっていること」に対して，CCC ならびに研究者コミュニティーから提供された研究用データセットを国内の研究者に提供し，マルウェア対策研究人材育成ワークショップ（MWS）で研究成果を共有する取り組みを行っている．本稿は，MWS2011 に提供している MWS 2011 Datasets の概要を説明し，本データセットを用いた研究の実験データの諸元となることを期待している．

最新の脅威を捉えた研究用データセットの収集・作成・蓄積や利用環境の構築・提供など包括的なフレームワークを検討すべく，情報処理学会コンピュータセキュリティ研究会の配下に MWS 組織委員会 を設立した．今後，年次の MWS 開催とは別に，中長期的な視点でマルウェア対策研究の課題に取り組んでいく．

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サイバークリーンセンターの関係者各位に深く感謝致します．

参考文献

1) MIT Lincoln Laboratory, DARPA Intrusion Detection Evaluation Data Sets, http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/index.html

2) B. Sangster, et al.: Toward Instrumenting Network Warfare Competitions to Generate Labeled Datasets, 18th USENIX Security Symposium CSET'09 (2009.08)

3) BADGERS2011: Building Analysis Datasets and Gathering Experience Returns for Security, http://iseclab.org/badgers2011/ (2011.04)

4) PREDICT: the Protected Repository for the Defense of Infrastructure Against Cyber Threats, https://www.predict.org/

5) サイバークリーンセンター, https://www.ccc.go.jp/

6) 畑田充弘, 他: 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)

7) マルウェア対策研究人材育成ワークショップ 2011, http://www.iwsec.org/mws/2011/

8) Mitsuaki Akiyama, et al: Design and Implementation of High Interaction Client Honeypot for Drive-by-download Attacks, IEICE Transactions on Communication, Vol.E93-B No.5 pp.1131-1139 (2010.05)