特集
マルウェア

7

研究用データセットを用いたマルウェア対策研究人材育成ワークショップ

http://www.iwsec.org/mws/2009/

畑田充弘（NTT コミュニケーションズ（株））

マルウェア対策の課題

マルウェアによる脅威が複雑化する中，さまざまな対策研究が盛んに行われている．一方，研究を行う上でさまざまな課題があり，その 1 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る．教材となるこのような研究用データは，これまで研究者らが独自にハニーポット（コラム参照）を設置して収集し，それぞれの解析手法や対策手法の妥当性を検証するために利用してきた．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，昨今のマルウェアに起因するインシデント事例や所属組織のポリシーによる制約から「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．

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，最新のもので 2000 年に公開されたデータセットである．しかし，2001 年の Code Red や Nimda，2003 年の Slammer などインターネットで猛威を奮ったワームの出現，2004 年頃から現在に至るボットによる脅威などへ大きく変化した攻撃手法が含まれていない．近年のものでは the 2009 Inter-Service Academy Cyber Defense Exercise datasets[2] というサイバー防御演習時のデータセットが公開されているが，マルウェアによる攻撃を想定したものではない．

このような課題を抱えている状況において，さらなる進化を続けるマルウェアに対峙していくために，自分に何ができるだろうか．サイバークリーンセンター（以降，CCC）で収集しているデータを有効に活用できないだろうか．そんなことを有志と語る中で，研究用データセット：CCC DATAset 2008[3] を研究者に提供して，研究成果を共有する場・切磋琢磨する環境を作ろうと意気投合して，「マルウェア対策研究人材育成ワークショップ 2008（MWS2008）」を開催することとなった．どれほどの発表件数が集まるか不安を抱いていたが，22 件の発表（うち学生による発表が 8 件）と 2 件のパネルディスカッションを通して，大学や研究機関に限らず産業界も交えた活発な議論を行うことができた．本稿では，2 回目の開催となった MWS2009 の開催模様と MWS2009 で提供した研究用データセット：CCC DATAset 2009[3] の概要を紹介する．

CCC DATAset 2009 の概要

CCC DATAset 2009 は，検体解析技術の研究を想定した「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を想定した「攻撃通信データ」，ボットの活動傾向把握技術の研究を想定した「攻撃元データ」で構成される．それぞれの概要と CCC DATAset 2008 との差異を紹介する．

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 10 個をテキスト形式で記載したファイルである．2009 年のマルウェア検体は次の（1）～（3）の観点で選定している．なお（2）（3）には（1）に記載されたハッシュ値も含まれる．

（1）解析結果を照合できる検体：9 検体

CCC のボットプログラム解析グループによって事前に静的解析をしている検体であり，解析精度の評価に活用する．

（2）関連性のある複数の検体：5 検体

連鎖感染など何らかの関連性のある複数の検体を 2 グループ選定しており，検体間の関連性分析の評価に活用する．

（3）特徴的な機能を有する検体：5 検体

耐解析機能や独自通信機能など特徴的な機能を有する検体であり，検体の特徴分析の評価に活用する．

ログ項目	例（一部を * でマスク）
マルウェア検体の取得時刻	2009-04-01 00:01:58
発信元 IP アドレス	honey035
発信元ポート番号	1034
宛先 IP アドレス	**.215.1.206
宛先ポート番号	80
TCP または UDP	TCP
マルウェア検体のハッシュ値（SHA1）	*****86f2ec74727b14001cfe0b88af718797c91
マルウェア名称	WORM_AUTORUN.CZU
ファイル名	C:\WINDOWS\system32\ptkj.exe

表 -1　攻撃元データのログ項目

項目	件数
全レコード数	2,470,766
TCP によるダウンロードレコード数	2,409,491
UDP によるダウンロードレコード数	61,275
ダウンロードホスト IP アドレス種類数	269,730
マルウェア検体のハッシュ値種類数	67,055
マルウェア名称種類数	1,335

表 -2　攻撃元データの基本情報

●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 pcap 形式のデータである．ハニーポットはホスト OS 上の 2 台のゲスト OS（Windows 2000 と Windows XP SP1）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．また，ゲスト OS はマルウェアに感染していないクリーンな状態に定期的にリセットされる．データ収集日は 2009 年 3 月 13 日（金）と 3 月 14 日（土）の 2 日間で，総パケット数が 3,511,850 パケット，約 580MB のデータサイズである．

●攻撃元データ

2008 年 5 月 1 日（木）から 2009 年 4 月 30 日（木）までの 1 年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 -1 のログ項目を 1 レコードとして記録した csv 形式のファイルである．なお，ハニーポットの IP アドレスは対応する ID（honey001 ～ honey094）に置換して記載されている．攻撃通信データのデータ収集環境と同様のハニーポットの構成をとり，国内の複数の ISP に接続された 94 台のハニーポットで記録された約 348MB のデータである．攻撃元データの基本情報をまとめた表 -2 において，ダウンロードホストとはマルウェア検体を取得した外部のホストであり，マルウェア名称種類数には UNKNOWN（アンチウイルスソフトでマルウェア取得時に検出できなかった場合に付与した名称）は含まない．

項目	2008	2009
マルウェア検体		
検体数	1	10
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
攻撃通信データ		
ハニーポット数	2 台	2 台
収集日	2008/4/28, 2008/4/29	2009/3/13, 2009/3/14
攻撃元データ		
ハニーポット数	112 台	94 台
ハニーポット ID	なし	あり
収集期間	2007/11/1 ～ 2008/4/30	2008/5/1 ～ 2009/4/30

表 -3　CCC DATAset 2008/2009 の比較

CCC DATAset 2008 ではマルウェア検体の検体数が 1 種類，攻撃元データの収集期間が半年間でログ項目にはハニーポット ID がなかった．MWS2008 の開催を通して得た関係者の要望を CCC DATAset 2009 に反映しており，その主な差異について表 -3 にまとめる．このような研究用データセットの量的・質的な改善に伴って，MWS2009 ではマルウェアの自動分類といった新たな研究や，MWS2008 の研究成果を踏まえたマルウェアの検知や攻撃傾向の可視化に関する研究の発表が行われた．

MWS2009

2009 年 10 月 26 日（月）から 10 月 28 日（水）の 3 日間，富山国際会議場にて MWS2009 を開催した（図 -1）．28 件の発表（うち学生による発表が 15 件）と 1 件のパネルディスカッションに加え，研究用データセットを用いた新たな取り組みとして MWS Cup 2009（本特集の「コラム：MWS Cup 2009」参照）も開催し，研究成果の共有ならびに切磋琢磨する環境として大変有益なワークショップとなった．CCC DATAset 2008/2009 と MWS2008 を総括する発表 1 件を除き，研究発表における CCC DATAset 2009 の各データの利用件数は，マルウェア検体：7 件，攻撃通信データ：14 件，攻撃元データ：6 件であった．以下，発表の内容を写真とともに振り返って紹介する．

●マルウェア検体を用いた発表

マルウェア検体を用いた発表では，ネットワーク型侵入検知システムによるトラフィック解析とマルウェアに感染したホストから取得した実行プロセスのコンテキスト情報を連携させた感染検知方式や，マルウェアがホストに侵入する際に繰り返し行われる挙動に着目した感染

《ハニーポット》

OS やアプリケーション（AP）の脆弱性を残したまま，攻撃を受けてマルウェアへの感染 PC を装う仕組みの総称であり，おとり PC とも呼ばれる．ハニーポットによって，攻撃コードやマルウェアにかかわる多くの情報を得ることができる．ハニーポットの実現方法には，攻撃側との対話レベルによる分類と攻撃の受け方による分類ができる．

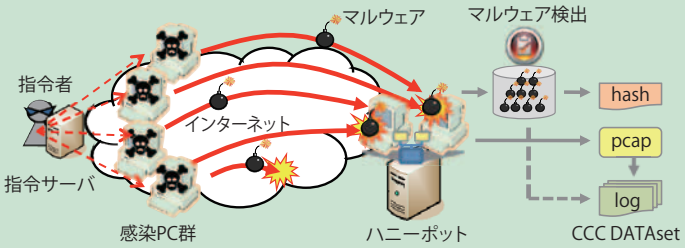

図 -5　ハニーポットと CCC DATAset

対話レベルによる分類では，実際に稼働している OS や AP の脆弱性を攻撃させるハイ・インタラクション型とエミュレートした脆弱性を攻撃させるロー・インタラクション型がある．ハイ・インタラクション型は攻撃を受けてマルウェアに感染した後の挙動まで捕捉することができ多くの情報を得ることができるが，ハニーポットの外部に攻撃を行ったりハニーポットの制御を奪われてしまったりするリスクを伴う．一方でロー・インタラクション型は攻撃を受けた後のマルウェアの実行が制限されるため，リスクは低いが得られる情報は少ない．攻撃の受け方による分類では，受動的に攻撃を待ち受けるサーバ型と能動的に攻撃を受けに行くクライアント型がある．サーバ型は脆弱性のある OS や AP を操作することなくワームなどによる感染活動を待ち受ける．クライアント型はブラウザなどを操作して攻撃コードの仕掛けられた Web サイトを閲覧することで攻撃を受ける．

CCC で運用しているハニーポットはハイ・インタラクション型のサーバ型ハニーポットであり，その規模や運用については本特集の「ボット対策プロジェクト：サイバークリーンセンターからみた国内のマルウェア対策」を参照いただきたい．ハニーポットで収集したマルウェアから選定したマルウェアのハッシュ値が「マルウェア検体」，pcap が「攻撃通信データ」となる（図 -5）．またアンチウイルスソフトによるマルウェア検出結果と pcap の解析により「攻撃元データ」を作成している．

検知方式の提案と評価が行われた．

日々大量に新たなマルウェアが発見され，手動による静的解析の限界から，実際にマルウェアを実行する動的解析にかかわる研究が盛んに行われている．動的解析にて得たさまざまなログ情報を自動解析して感染動作のみを抽出する研究や，完全に隔離した実行環境と安全な通信のみインターネット接続を許可した実行環境との違いによる解析結果の比較，実行時の挙動を網羅的に解析するための試行回数を調査した実験結果が報告された．

一方で，マルウェア検体のプログラムコードの類似性を機械語命令単位で高速に算出する自動マルウェア分類システムの提案と分類結果も示され注目を集めた（本特集の「研究用データセット：マルウェア検体編」参照）．

●攻撃通信データを用いた発表

攻撃通信データを用いた発表では，通信の変化の様子を可視化するツールや，攻撃通信データを動的解析によって得た通信挙動ログと見なして擬似クライアントによる模擬通信を行い実サーバからの応答を蓄積するシステムの提案が行われた．

研究テーマとして最も多かったのが，マルウェアの通信挙動の特徴抽出と感染検知である．ダウンロードホストの応答時間や DNS の応答内容，パケットのヘッダ情報のみを用いた時系列での傾向，一般的な OS では利用されない TCP フィンガープリントなどを特徴として，感染検知の方式提案と評価が行われている．また，解析に有用となる統計情報や特徴のデータベース化による解析の効率化の提案もあった．


図 -1　MWS2009 会場風景

学生優秀論文発表賞を受賞した桑原和也氏（東海大学）の発表（図 -2）では，攻撃通信データから 14 種類の特徴量を抽出し，発見した規則（ポートスキャン，連鎖感染，マルウェア取得時の通信方向）による条件分岐から，感染有無を判定する手法が提案され有効性が示された．

優秀論文発表賞を受賞した竹森敬祐氏（（株）KDDI 研究所）の発表（図 -3）では，9 種類の侵入フェーズ，5 種類の指令・配布フェーズ，10 種類の攻撃フェーズを通信要素として攻撃通信データの調査を行っている．101 種類の通信シナリオ（通信要素のパターン）が観測され，マルウェア検体を用いて通信シナリオを抑制する対策の効果を比較評価している．

●攻撃元データを用いた発表

攻撃元データを用いた発表では，未知検体のダウンロ

図 -2　桑原氏の発表の様子

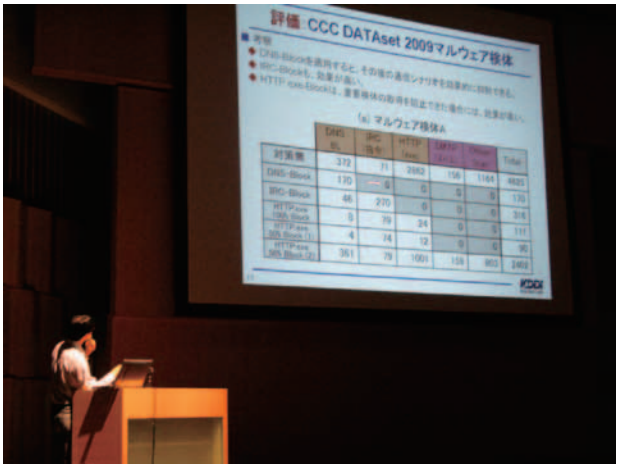

図 -3　竹森氏の発表の様子

ードホストの地理的位置に関する時間的変化や，時間的な関連性に着目した連鎖感染の可視化（本特集の「研究用データセット：攻撃元データ編」参照）の発表があり，他のデータセットを用いた発表も含めて，可視化が重要なテーマとなっていることがうかがえる．

ダウンロードホストの IP アドレスとダウンロード時刻の相関から独立した 4 つのボットネットと 5 つの活動フェーズがあったとの分析や，独自に収集したデータとの比較によりマルウェア感染活動の局所性を示した発表もあった．

また，マルウェアの種類ごとに複数ハニーポットによる捕捉特性（ハニーポットの IP アドレス情報があれば AS ごとの感染特性を示すことができる）を可視化して，動的解析結果を関連付けてネットワークリソースへの影響度を推定する被害予報システムも提案された．

●パネルディスカッション

MWS の新たな展開に向けて，マルウェアの動的解析時に得られる動作記録データの必要性や活用方法に焦点を当てたパネルディスカッションも行った（図 -4）．MWS2009 プログラム委員長の門林雄基氏（奈良先端科学技術大学院大学，（独）情報通信研究機構）をコーディネータとして，動作記録データの提供視点で真鍋敬士氏（JPCERT/CC）と岩村誠氏（NTT 情報流通プラットフォーム研究所），利用視点で筆者と佐々木良一氏（東京電機大学）がパネリストを務めた．動作記録データによってマルウェア解析のハードルを下げることへの期待や，一方で静的解析ができる人材の育成にも目を向けるべきといった議論が会場を交えて行われ，MWS2009 は幕を閉じた．

CCC DATAset 2009 と MWS2009 を通して得られたものは研究成果だけではない．MWS2009 開催後の意見交換会で共有された静的解析の結果は，マルウェア検体を利用した研究者へのフィードバックとなるとともに，利用していない研究者にとっても今後の研究に向けた教材として有用な情報となるだろう．「マルウェア対策研究にかかわる日本の研究者が一堂に会することでお互いの顔が見えるようになった」，「同じデータを利用していることによって認識の違いが生まれにくい」といったことも参加者の声として聞こえてきている．継続的かつ効果的に対策研究を行っていくためにも，進化するマルウェアの脅威に対応したデータの収集環境や種類など研究用データそのものの研究分野の発展に期待している．


図 -4　パネルディスカッションの様子

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) Sangster, B. et al. : Toward Instrumenting Network Warfare Competitions to Generate Labeled Datasets, 18th USENIX Security Symposium CSET '09 (Aug. 2009).
3) 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 (Oct. 2009).

（平成 21 年 12 月 26 日受付）

畑田充弘（正会員）
m.hatada@ntt.com
2003 年早稲田大学大学院理工学研究科修士課程修了．同年，NTT コミュニケーションズ（株）入社．以来，マルウェア対策をはじめとするネットワークセキュリティの研究開発に従事．