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

畑田 充弘[†1]　中津留 勇[†2]　寺田 真敏[†3]　篠田 陽一[†4]

[†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] 株式会社 日立製作所　〒212-8567 神奈川県川崎市幸区鹿島田 890
[†4] 北陸先端科学技術大学院大学　〒923-1211 石川県能美市旭台 1-1

E-mail: [†1] m.hatada@ntt.com, [†2] office@jpcert.or.jp, [†3] masato.terada.rd@hitachi.com, [†4] shinoda@jaist.ac.jp

あらまし 近年のマルウェアによる脅威は複雑化しており，予防･検知･防御･回復といった対策の研究が盛んに行われている．しかしながら，それぞれの研究の評価には，独自にハニーポットを設置して収集したデータが用いられることが多く，客観的な評価が困難である．本稿では，サイバークリーンセンターで収集している近年のマルウェアによる脅威を捉えた研究用データセット（CCC DATAset 2008）の概要を述べ，データセットを用いた研究成果が共有されたワークショップ（MWS 2008）の概要を報告する．また，MWS 2009 に提供される CCC DATAset 2009 の概要を述べ，研究用データセットの要件と課題をまとめる．

Dataset for anti-malware research and research achievements shared at the workshop

Mitsuhiro Hatada[†1]　You Nakatsuru[†2]　Masato Terada[†3]　Yoichi Shinoda[†4]

[†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] Hitachi, Ltd.
890 Kashimada, Saiwai-ku, Kawasaki, Kanagawa, 212-8567 Japan
[†1] Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1211 Japan

E-mail: [†1] m.hatada@ntt.com, [†2] office@jpcert.or.jp, [†3] masato.terada.rd@hitachi.com, [†4] shinoda@jaist.ac.jp

Abstract There has been a lot of research on prevention, detection, protection and recovery against the complicated threats by malware. However, objective evaluation is likely to be hard because each evaluation data was made in their own way such as collecting by their original honeypot and its operation. This paper describes CCC DATAset 2008, a dataset for recent anti-malware research, and research achievements shared at MWS 2008. Furthermore, CCC DATAset 2009 for MWS 2009 and the requirements of dataset are outlined.

1. はじめに

マルウェアによる脅威は複雑化しており，予防･検知･防御･回復といった対策の研究が盛んに行われている．それぞれの研究の評価には，独自にハニーポットを設置して収集したデータが用いられることが多く，客観的な評価が困難である．一方で，組織のポリシーによる制約から，ハニーポットによる研究用データの収集が困難である場合も多く，マルウェア対策に関わる研究の発展を阻害する一因となっている．研究用データに関して，文献 1)ではネットワークセキュリティ技術の研究，開発，性能検証のために必要となるデータセットの調査が行われ，公開されている代表的なものとして侵入検知システムの研究のためのトラフィックデータ：DARPA Intrusion Detection Evaluation Data Sets[2]を挙げている．1998 年，1999 年，2000 年の三つのデータセットが，学習用と学習後の検証用として提供されている．しかしながら，近年のボットネットを代表例とするマルウェアの高度な機能や，その運用による攻撃の手法や傾向は異なっており，現状の脅威に対する研究用データとしては適切とはいえない．その他，文献3)ではマルウェアの解析用データセットとして，時系列でのOS内部観測によるAPI呼び出しとその引数を抽出して行列を生成したものもある．

本稿では，サイバークリーンセンター（CCC）[4]で収集しているデータを元に作成した研究用データセット（CCC DATAset 2008）の概要を 2 章で述べる．3 章では，本データセットを用いた研究成果が共有されたマルウェア対策研究人材育成ワークショップ 2008（MWS 2008[5]）の概要を報告する．また，4 章では MWS 2009 に向けて提供される CCC DATAset 2009 について概要を述べ，5 章で研究用データセットの要件と課題をまとめる．

2. CCC DATAset 2008

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、それぞれについて概要を述べる ．

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）一つをテキスト形式で記載したファイルである．マルウェア検体の選定にあたっては，IRC 接続や自己アップデート，DDoS 攻撃の実行など機能が豊富であり，仮想マシン環境の検知などのアンチデバッグ機能やパッキングなど耐解析性が高いという方針で選定している．

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，Windows 2000 と 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 2008 年 4 月 28 日と 2008 年 4 月 29 日，総パケット数が 15,901,943 パケット，約 2.8GB のデータサイズである．

2.3. 攻撃元データ

2007 年 11 月 1 日から 2008 年 4 月 30 日までの 6 ケ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1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．攻撃通信データのデータ収集環境と同等のハニーポットの構成をとり，国内の複数の ISP にそれぞれ接続された 112 台のハニーポットで記録された約 390MB のデータである．攻撃元データの基本情報を表 2 に示す．

通信方向と利用ポート番号の関係を図 1 に示す．通信方向：PULL では，脆弱性に対する攻撃成立後に，ハニーポットからダウンロードホストにマルウェア検体を要求し，その要求時のダウンロードホストのポート番号が利用ポート番号となる．通信方向：PUSH では，ダウンロードホスト，攻撃元ホストや C&C サーバなどからの命令によりポートを開き，ダウンロードホストからマルウェア検体が転送されるのを待つ．その際の待ち受けポート番号が利用ポート番号となる．その際，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．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．なお，ハニーポットを識別できる情報（IP アドレスなど）は提供されていない．

表 1 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	2007-11-01 00:02:01
ダウンロードホスト IP アドレス	**.10.167.74
利用ポート番号／TCP または UDP	6251/TCP
通信方向	Pull
マルウェア検体のハッシュ値（SHA1）	*****a7e7edca3b787624c4edb6cc74d4dbd1b8f
ウイルス名称	PE_VIRUT.XV
ファイル名	C:\WINNT\system32\cwgbiw.exe

表 2 基本情報

項目	件数
全レコード数	2,942,221
TCP によるダウンロードレコード数	2,846,053
UDP によるダウンロードレコード数	96,168
ダウンロードホスト IP アドレス種類数	258,711
マルウェア検体のハッシュ値種類数	52,465
ウイルス名称種類数（UNKNOWN 含まない）	1,081

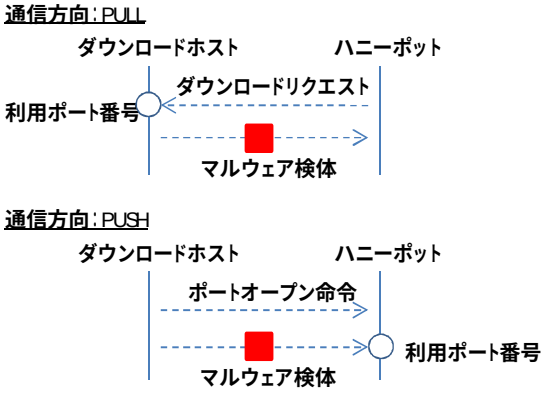

図 1 通信方向と利用ポート番号

3. マルウェア対策研究人材育成ワークショップ 2008（MWS2008）

MWS 2008 では CCC DATAset 2008 の「マルウェア検体」を用いて 6 件，「攻撃通信データ」で 8 件，「攻撃元データ」で 8 件，合計 22 件の研究発表が行われた．以下，それぞれの研究成果の概要について，得られた知見を中心に報告する．

3.1. マルウェア検体

文献 6)では，模倣 DNS サーバ(NINFD)なし／ありの環境で検体を 10 分間動作させ，パケットキャプチャと検体が展開されている領域のメモリダンプを比較している．パケットキャプチャでは，ある DNS クエリが行われ，NINFD なしでは応答がなく 10 秒間隔で再度同じクエリを行う．一方で NINFD ありでは応答した偽 IP アドレスの 6667/tcp への接続を試みるが，当該ポートが開いていないため，3 秒，6 秒，1 秒の順に間隔を置いて 90 回接続を試み，再度クエリを行い 6667/tcp への接続を試みる動作を繰り返す．メモリダンプでは，NINFD なしの方が strings コマンドで得られる情報は多く，この差は，検体の実行 10 分経過時点のもののため，NINFD なし／ありによる DNS クエリ後の動作が変更されたと類推している．

文献 7)では，独自のページフォルトハンドラを用いて出力した「書き込まれた後に実行される」メモリページを候補として，確率モデルによるコンパイラ出力コードの尤もらしさからオリジナルコードを特定する．アンパッキング結果として，当該検体は起動した後に三回の再起動を繰り返し，二番目のプロセスでオリジナルコードが出現した後にマルウェア自身がシステムディレクトリへコピーされそこから再起動している．つまり一度目の再起動はパッカーによるもの，二度目の再起動はアンパッキングされたマルウェアによるものということがわかる．三番目のプロセスが起動し，パッカーにより四番目のプロセスが起動後，ボットとしての本来の活動が開始される．

文献 8)では，システムフォルダへマルウェア検体自身を複製する自己ファイル READ と，複製後に元のマルウェア検体自身を削除する自己ファイル DELETE をボットの挙動と定義した検知方式を提案し，マルウェア検体と著者らが独自に収集した検体による検知実験を行っている．検知実験ではマルウェア検体のみ自己ファイル DELETE を検出できていないが自己ファイル READ は検出しており，実行時に仮想マシンチェックを行った上でその後の活動を続けるようなボットとは異なる動作であるとしている．

文献 9)では，アンチデバッグ機能を無効化しつつ解析作業を行えるステルスデバッガは，ゲスト OS のさらに下の階層で高い権限で動作し，従来の OS や CPU のデバッグ支援機構に頼らない．パッキングされたマルウェアが OEP (Original Entry Point) へジャンプする直前の特徴的な動作をシグネチャとして定義し，デバッガをブレークさせることで効率的にアンパッキングを行えるとしている．解析結果として，パッカーの中から呼び出される API の先頭アドレスに 0xCC がセットされていないかのチェック，意図的に例外を発生させるために独自例外ハンドラの登録，シングルステップ状態で実行されていないかの pushf, popf を利用したフラグチェック，自身が仮想マシン内で動作しているかのチェック，といった複数のアンチデバッグ機能を備えているとしている．

文献 10)では，擬似インターネット環境を含む動的解析環境において，実行したボットが IRC サーバへ接続した時に，ボットが呼び出す文字列処理関数を操作し，ボットがハーダーからのコマンドをチェックする際の文字列比較において比較する文字列が完全に一致する場合も必ず false を返すようにすることで，全てのコマンド文字列を抽出できる．また，あらかじめ決められた文字列をパラメータにセットしたコマンドをボットに送信し，細工された文字列処理関数などで引数を推定することで，コマンドパラメータを得る．マルウェア検体の解析結果とし約 20 分間で実行環境の情報収集，IRC 接続，HTTP や FTP によるダウンロードやスピードテスト，DDoS などコマンド(各パラメータ含む)を 99 個抽出することができている．

文献 11)では，標的型攻撃に実装されている代表的な耐解析機能としてパッキング，独自 IAT (Import Address Table)，文書ファイルなどのデータ形式ファイルの媒介，デバッガ検出などが挙げられている．解析効率化のため，耐解析機能を自動的に解除し従来と同様の解析が可能な実行形式ファイルを生成し，関連情報をレポートするコアエンジンなどが必要となる．その一機能として実装したインストラクショントレーサーにより，マルウェア検体実行時の任意の箇所での停止･再開，レジスタ･メモリの参照，停止時のサブツールの実行などの正常動作を確認している．

3.2. 攻撃通信データ

文献 12)では，シェルコードが数十から数キロバイト連続した NOP 命令を含むことを利用して，シェルコード候補としてパケットを抽出し，NOP 命令の連続数と連続した NOP 命令群の出現パターンとパケットサイズにより分類を行っている．分類された各シェルコード群からそれぞれ複数個を選び，利用するシステムコールによる動作タイプや，XOR による暗号化有無，UNICODE による文字列処理などの構成の特徴により，12 種類にシェルコードを分類している．NOP 命令は 0x90 を使用したものが大半であったが，あるシェルコード分類では 0x41～0x44 を使用しているものもあった．シェルコード分類と取得するマルウェアの関係から，あるシェルコード分類は多種多様なマルウェアにおいても再利用されていることがわかる．当該シェルコード詳細分析から，マルウェア本体のダウンロードが始まる前にシェルコード中に埋め込まれたキーとなる値が送信されていることがわかった．

文献 13)では，攻撃通信データの解析により得たバイナリコードを比較し，一致するコードの割合をもとにファイル間の相互の類似性を算出し，その値を使って実行ファイルの分類を行っている．また実行ファ

イルのメモリダンプから逆アセンブルにより実行ファイルが使用する API の抽出を行っている．UNKNOWN と判定された実行ファイルのバイナリコードの類似度判定結果から，コードの類似度が高い実行ファイルはダウンロードホストの IP アドレスが同一であったことが確認されている．

文献 14)では，特徴的な挙動として，何回かの攻撃通信が到達したが感染していない，攻撃が成功したがダウンロードホストからの応答がなく要求の再送を繰り返す，攻撃から他ホストへの攻撃拡大まで一通りの動作をする，多重感染して活動するといったケースを挙げている．DNS クエリの分析では従来の研究成果を検証している．Windows 2000 では IP アドレスを指定した正引きが多数行われているが，XP では同様の DNS クエリは見られない． XP ではリゾルバの逸脱が確認できたが，2000 では同様の挙動は見られない．普通でない qtype（MX，AXFR，IXFR）は両ハニーポットで確認されていない．また，ハニーポットが設置されている ISP 内のホスト名を名前解決しようとしている傾向が見える．

文献 15)では，攻撃コードの受信と感染後の実行ファイルダウンロードにおいて，HTTP によるダウンロードと同程度の回数利用されている独自のファイル転送プロトコルが確認され，その詳細が述べられている．TCP セッション確立後にファイル送信側が 380 バイトの固定長 Hello メッセージを送信し，ファイルを受信する側が 4 バイト固定長の Ack メッセージを送信し，その後ファイルが送信される．Hello，Ack はそれぞれ 1 パケットで送信されている．Hello の 380 バイトは先頭から 168 バイトと末尾 200 バイトは全ての通信で同一であり，間の 12 バイトが拡張子を含むファイル名の指定と考えられる．Ack は送信ホスト，送信元ポート，実行ファイルの種類が同じ場合のみ同一のデータが送信されていることも確認されている．

文献 16)では，ボットの多重感染活動を見分けるために宛先ポート番号を可視化の要素に組み込み，通信のインバウンドとアウトバウンドを見分けるためにパケットの向きも可視化の要素とし，2 台のハニーポットのトラフィックを分離して並べて可視化している．両ハニーポットでほぼ同時刻に同じメールサーバへのパケットを送信している様子や，それと同時に異なる IP アドレスを持つノードに対するスキャン活動など，同調活動を確認している．同調活動とみなす許容時間差を 1 秒とした場合に約 700 件，10 秒とした場合に 5,800 件と多くの同調活動を検知している．

文献 17)では，攻撃通信データと攻撃元データから検体の感染時刻の間隔に着目して，閾値を 3 分に設定した連鎖感染の分析を行っている．攻撃通信データからハニーポット 1 台の 1 日当りの平均として 3 検体の連鎖感染が確認でき，最大 20 検体の連作感染も見られた．連鎖感染ツリーから，連鎖の最初は PE ファイル感染型である割合が高い，未知検体は種類数が少ないがほとんどの連鎖に含まれる，未知検体は連鎖の 2 番目に現れる傾向が強い，未知検体からの連鎖は 80 番ポートを使用する割合が高い，などを報告している．また，連鎖感染ツリーをノードにより重ね合わせた連鎖感染マップからは，既知の検体と関連性が全くない未知検体の連鎖，多数の既知検体と関連している未知検体の存在が示唆されている．

文献 18)では，ボットネットの多段追跡システムの一部として，一定時間エッジルータを通過する全パケットを記録して，ボット PC 及び C&C サーバの IP アドレスの特定を行う．攻撃通信データの解析から得た 15 種類の検体種類（うち 5 種類は未知検体として個々に扱っている）毎に接続先サーバを調査した結果，複数回異なる C&C サーバ及びダウンロードサーバに接続する通信と，15 種類のボットが，特定の 4 サーバのうちの 1 つ以上に接続していることを確認した．これら 4 サーバの名前解決を行う通信を利用し，DNS クエリの送信元をボット PC 特定のルールとし，クエリ結果を C&C サーバ特定のルールとして，侵入検知システムによる検知実験を行い全 15 種類のボットに関してボット PC と C&C サーバを特定している．

文献 19)では，ハニーポットから発信される中継サーバへのパケットで 1 つでも宛先 IP アドレスが一致する場合を同じボットネットに属するとして，1 つのボットネットを構成する通信データから，17 重の中継サーバの冗長化が行われ，3 つ以上のボットからアクセスされる活発な 7 重の中継サーバの存在が確認された．また，1 つのコードを配布する中継サーバ数は平均 2.1 台で，最大 69 台であり，攻撃元データを母数とすると 25.7%のコードが分散管理されていることになる．各コードに感染して 10 分後のコードの追加・変更数とレジストリまたは hosts ファイルの追加・変更・削除の状況分析から，顕著なものには数百のコード変更とレジストリ操作を行うコードや，ウイルススキャナのパターンファイル配布サーバのホスト名に対する IP アドレスを 127.0.0.1 あるいは 255.255.255.255 に hosts ファイルを変更するコードがあり，ボット駆除の耐性も有している．

3.3. 攻撃元データ

文献 20)では，攻撃元データのうち，最初はウイルス名称が UNKNOWN だった 805 個のハッシュ値に注目し，UNKNOWN 期間はウイルスによってばらつきが多く長いものは 100 日以上にわたること，また UNKNOWN 期間の 10 日前後と 30 日前後に出現回数が多いウイルスコードが存在すること，ダウンロードホスト IP アドレスが特定の 6IP アドレスに偏っていることが示されている． UNKNOWN 期間のダウンロードホスト IP アドレスからダウンロード元の地域との関係を可視化・分析し，欧州とアジアからのみのダウン

ロードは少ないこと，北米やアジアは固有種が多く欧州は固有種が少ないことがわかる．日本を含めて地域固有の数が多い場合でも，影響が局所的なこともあり対応が遅い傾向があると指摘している．

文献21)では，多次元データ分析ツールにより攻撃元データを分析している．ダウンロードホストの現地時間に補正した時間帯別での攻撃数の変化からは顕著な特徴が見られなかった．マルウェア検体を対象に，著者らが構築･運用している自動解析システムでの解析結果として，hosts ファイルの最後に 1 バイト(\x0a)追加や，模倣 DNS サーバへのある DNS クエリ後に IRC サーバへのログインなどが報告されている．このような挙動をカテゴリデータへ変換してベクトル化し，ハミング距離による挙動の類似性に基づくマルウェアの名称推定では，著者らの蓄積した約 7,600 検体のデータとの類似性から実際の名称とは一致しなかった．実行環境の違いや比較対象とした蓄積済解析データの偏りが原因と推定している．

文献 22)では，攻撃元データと攻撃通信データ及び ISDAS の定点観測データの比較から，平均到着間隔(スキャン密度)は 11%の誤差で一致が見られたが，宛先ポートの分布とスキャン速度では ISDAS で観測しているものは全てがボットによる攻撃ではないこと，またポートスキャンのタイプはマルウェアではなく C&C サーバからの命令に依存するとしている．決定木学習によりペイロードを見ない通信パターンの特徴量に基づいて，再現率 93%，適合率 94%の精度でスキャンパターンの同定が可能であり，識別には総送信パケット数とユニーク宛先アドレス数が有益であること，再現率 79%，適合率 69%の確からしさでウイルス名称の同定が可能，識別にはパケット数やユニーク発信元アドレス数が有益であるとしている．

文献 23)では，攻撃元データを用いてマルウェアの配布元ノードの国別特性，時差を考慮した活動状況，特定の配布サイトを分析している．日本の配布元ノードは全体の 79.8%を占め，そのうち 96.4%が PULL 型の転送方法だが，アメリカでは 94.3%，韓国では 92.0%が PUSH 型の転送であり，大きな偏りが見られる．日本では多くの家庭で NAPT 機能などにより外部からの通信が制限されている環境に起因するものとしている．日本での時間毎の平均的な攻撃元ノード数の分布から，NSPIXP2 のトラフィック傾向と類似していること，台湾やアメリカとの比較から，時差を補正しても分布に類似性が見られないとしている．

文献 24)では，攻撃元データを用いて人間が調査･分析するための User Interface (UI) の要件として，大域情報としての傾向情報の提供，情報視覚化による認知支援，段階的な調査方法の提供，対話機能による調査作業の簡易化の 4 点が挙げられている．これらの要件をもとに実装した MWLT Browser (MalWare Log Trend Browser) を使用した攻撃元データの調査事例についても報告している．MWLT Browser の使用により未知のマルウェアが活発に活動を開始した兆候を捉えることができたとしている．

文献 25)では，2008 年 4 月の複数観測データ間の IP アドレス重複分析結果として，攻撃元データと定点観測データの一致数が他の一致数に比べて多いこと，スパムメール発信元に対して攻撃元データの一致数は少ないこと，攻撃元とスパムメールの誘導先がごく短い期間で連動していること，攻撃元データとマルウェア感染ブラックリストから 12 種類の URL が同一の IP アドレスであり攻撃元として期間中まばらに観測されたこと，攻撃元とスパムメール誘導先及びフィッシングサイトとして 5 個の IP アドレスが 1 週間程同時に活動していたこと，攻撃元とスパムメール発信元及び定点観測データで一致した 7 個の IP アドレスの 1 つに着目するとスパムメール発信元としての利用は頻度が少ないことなどが報告されている．

文献 26)では，著者らが MITF (Malware Investigation Task Force) で観測している 1ISP 内の局所的なデータと攻撃元データを比較分析した結果が報告されている．両者において共通で観測されたマルウェアはどちらの環境においても活発な活動を行っているマルウェアを多く含むとしている．また，ネットマスク長を距離として，攻撃元データに固有のマルウェアにおけるダウンロードホストの IP アドレスから MITF 観測点範囲までの距離とマルウェア取得件数の関係，ならびに MITF 攻撃元データに固有のマルウェアにおけるダウンロードホストの IP アドレスから MITF 観測点までの距離とマルウェア取得件数の関係から，攻撃元データに固有のマルウェアについてはそのほとんどは MITF 観測点から遠く離れていたことがわかり，マルウェアの感染活動には局所性があることを示している．

文献 27)では，攻撃元データを用いてダウンロードホストに着目したマルウェアの特徴を分析した結果が報告されている．ダウンロードホストの活動期間は，長期間，短期間，及び複数期間の 3 種類に分類でき，ダウンロード回数とマルウェアの種類数の関連性は薄く，ウイルススキャナのパターンファイル配布前後でダウンロード回数の変化が小さい．しかしマルウェアの更新前後ではダウンロード回数が変化し，1 日の平均ダウンロード回数が多いマルウェアの活動期間は短い傾向にあり，ダウンロードホストの活動期間が長いほどマルウェアの種類数が多い傾向にあるなどマルウェアの活動傾向が広く分析されている．

4. CCC DATAset 2009

CCC DATAset 2008 と同様の研究テーマを想定した三つのデータから構成される．また，CCC DATAset 2008 も参考情報として提供される．CCC

DATAset 2008 に対する CCC DATAset 2009 の差異比較を表 3 に示す. 主な差異としては「マルウェア検体」の対象とする検体数拡大,「攻撃元データ」の期間拡大とデータ項目の追加(ハニーポット ID)が挙げられる. なお, 攻撃通信データのハニーポットは, 対応するハニーポット ID で示している.

4.1. 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) 10 個をテキスト形式で記載したファイルである. マルウェア検体は以下の観点で選定しており, それぞれハッシュ値を記載している.

(1) 解析結果を照合できる検体:9 検体

(2) 関連性のある複数の検体:5 検体

(3) 特徴的な機能を有する検体:5 検体

(1)は事前に静的解析が完了している検体であり, 解析精度の評価に活用することを考慮した要件である. (2)は静的解析に基づいて, 何らかの関連性を持っている複数の検体を選定している. 検体間の関連性分析の評価に活用することを考慮した要件である. (3)は静的解析に基づいて, 耐解析機能や特殊な通信機能などの特徴的な機能を有する検体を選定している. 検体の特徴分析の評価に活用することを考慮した要件である. なお, (1)～(3)の間で重複して記載されているハッシュ値もある.

4.2. 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 2.2 と同じ構成をとり, 同じ ISP に接続しているが, ハニーポット自体は異なり, honey003 が Windows XP SP1, honey004 が Windows 2000 である. なお, CCC DATAset 2008 では honey001 が Windows XP SP1, honey002 が Windows 2000 である. データ収集日は 2009 年 3 月 13 日と 2009 年 3 月 14 日の 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.

4.3. 攻撃元データ

2008 年 5 月 1 日から 2009 年 4 月 30 日までの 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 4 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. 攻撃通信データのデータ収集環境と同等のハニーポットの構成をとり, 国内の複数の ISP にそれぞれ接続された 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表 5 に示す.

図 1 で示した通信方向:PULL の場合は送信元がハニーポットとなり, 宛先がダウンロードホストとなる. 通信方向:PUSH の場合は送信元がダウンロードホストとなり, 宛先がハニーポットとなる. ただし, 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～honey094) に置換されて記載されている. その他の項目については 3.3 と同様である.

表 3 CCC DATAset 2008／2009 の差異比較

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	honey001, honey002	honey003, honey004
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 4 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	2009-04-01 00:01:58
送信元 IP アドレス	honey035
送信元ポート番号	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)	*****86f2ec74727b1400 1cfe0b88af718797c91
ウイルス名称	WORM_AUTORUN.CZU
ファイル名	C:\WINDOWS\system32 \ptkj.exe

表 5 基本情報

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	63,820
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
ウイルス名称種類数(UNKNOWN 含まない)	1,335

5. 研究用データセットの要件と課題

CCC DATAset 2008／2009 は, マルウェア対策のための研究用データセットの整備を目的に収集されたデータではないことを踏まえて, あらためて本来の研究用データセットの要件を整理し, CCC DATAset 作成時の考慮事項と課題をまとめる.

5.1. データの種類

(要件) 幅広いマルウェア対策の研究に適したデータが必要である. ①プログラムされたマルウェアの動

作を解析できる検体そのもの，②ネットワークを介した感染手法及び感染後の挙動データ，③PC 内部における感染後の挙動データが挙げられる．また，研究の目的に応じて，④必要となる前処理を施した扱い易いデータ，⑤データ収集時点でしか得られない補足データがあることも望ましい．

(考慮事項) CCC で蓄積されているデータを使って，①はマルウェア検体，②は攻撃通信データ，④は攻撃元データとしている．マルウェア検体は安全面を考慮して，ハッシュ値による提供としている．攻撃元データは，時系列でのマルウェアやダウンロードホストの変化といった多面的な傾向分析ができることを想定してログ項目を選定した．

(課題) ファイルやレジストリの操作，サービスの起動や停止といった③PC 内部の感染後の挙動データや，DNS のレコードや各種の公開ブラックリストといった⑤データを収集した時点でしか得られない補足データが提供されていないことなどがある．

5.2. データ収集環境の網羅性

(要件) OS の種類とバージョン，パッチ適用状況，アプリケーション導入状況や受動的攻撃を受けるためのアプリケーションの操作，各種設定といった攻撃対象そのものの環境と，契約 ISP，IP アドレス帯，帯域，アクセス制御といったネットワーク接続環境により，攻撃手法や感染後の動作が異なる．そのため様々なデータ収集環境を準備し，同じ環境でも複数の環境を準備することが，より多くの攻撃を受け，より多くのマルウェア検体を収集し，より多くの感染後の動作に関するデータを収集できる．

(考慮事項) 一般ユーザが多く利用する PC 及びインターネット環境を加味して，攻撃元データでは，国内の主要な ISP に接続しているハニーポットを多数混在させた．

(課題) データ収集環境の網羅性を高くすればするほど，物理的･論理的なリソースのコストが高くなる．もちろんデータ収集環境であることを攻撃者に検知されないようにするための技術的対策も必要となる．膨大なマルウェアの解析技術の研究として自動解析も注目されており，その有効性の評価が可能な検体数を提供することや，被害が広がっている受動的攻撃に関するデータも必要となる．

5.3. データ収集の期間

(要件) マルウェアの活動の傾向は，短期間においても長期間においても変化が見られるため，長期間に渡る連続性のあるデータであることと，最新のデータを即時提供可能であることが求められる．

(考慮事項) 一般ユーザがよく PC を利用する日を考慮して，攻撃通信データは連続する 2 日間で休前日と休日を選定した．また攻撃元データは，長期での傾向分析にも対応できるよう，2008／2009 で連続し，2009 では期間を 1 年とした．

(課題) データ収集期間に依存して増加する適切なデータ量を見極めることに加えて，データ収集ならびに提供主体の所在も課題となる．継続的な管理を行えることが望ましく，また客観性や公平性という観点では第三者的な公的機関などが管理主体となることが望ましい．当然ながら民間の企業や団体，大学が管理主体となった場合でも，研究用データセットが広く活用されることが望ましい．

5.4. データ収集環境の運用情報

(要件) 研究用データセットにより何らかの傾向変化を捉えた場合に，マルウェアの活動傾向の変化とデータ収集環境の変化のどちらに起因するかを切り分け可能であることが望ましい．また，年毎のデータセットの比較を容易にするためにも，同一のデータ収集環境を維持する必要もある．データ収集環境の変化としては，ハニーポットとなるゲスト OS をクリーン状態へリセットする周期や，攻撃の宛先となる IP アドレスの割り当て方法，障害対応あるいは性能拡張といった要因でのデータ収集環境の構成変更の記録などが挙げられる．

(考慮事項) CCC の運用に関わる内容であり，データ収集環境の情報は公開していない．ただし，研究に必要な最低限の情報については，意見交換会などの活動を通して提供することとした．

(課題) データ収集環境の運用情報そのものが，管理主体にとって機密性の高い技術的ノウハウに相当する内容である場合や，攻撃者がデータ収集環境であることを検知できる条件となりえることから，環境情報とともに情報の公開範囲や公開内容を検討する必要がある．

6. おわりに

近年活発に行われているマルウェア対策研究において，研究素材あるいは客観的な評価のための研究用データセットとなりえる CCC DATAset 2008 について述べ，MWS 2008 で共有された各研究成果の概要を報告した．マルウェアの解析技術，感染手法の検知ならびに解析技術，ボットの活動傾向把握技術などの様々な研究成果が共有されるとともに，研究用データセット自身が研究者間での共通言語としての役割を担うことや，研究用データセットとともに研究に用いたツールや解析したデータが共有されれば人材育成を含む本研究分野の発展に寄与することが期待できる．また，MWS 2009 で使用される CCC DATAset 2009 について述べ，研究用データセット

の要件と課題を概観した．データ種類，データ収集環境の網羅性，期間，運用情報などの課題に加えて，現実的には研究用データセットの提供と利用に関わる手続き上の課題もある．今後，最新の脅威を捉えた研究用データセットの作成と共に，データセットの活用ならびに評価用として利用可能な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CCCの関係者各位に深く感謝致します．

参考文献

1 ）細井琢朗，他：公開ネットワークログデータセットの調査とワーム検知数の変遷調査，情報処理学会研究報告，Vol.2009-CSEC-44，No.31，pp.181-186，2009

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

3 ）安藤類央，他：Memento Project: 仮想化技術を用いたOS内部観測による解析用データセットの構築と公開，Vol.2009-CSEC-45，No.29，pp.1-6，2009

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

5 ）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/2008/

6 ）三輪信介，他：模倣 DNS によるマルウェア隔離解析の解析能向上，MWS2008，pp.19-24，2008

7 ）岩村誠，他：コンパイラ出力コードモデルの尤度に基づくアンパッキング手法，，MWS2008，pp.103-108，2008

8 ）酒井崇裕，他：自己ファイル READ/DELETE の検出によるボット検知の可能性に関する一検討，MWS2008，pp.109-114，2008

9 ）川古谷裕平，他：ステルスデバッガを利用したマルウェア解析手法の提案，MWS2008，pp.115-120，2008

10）星澤裕二，他：動的解析による BOT コマンドの自動抽出，MWS2008，pp.121-126，2008

11）鵜飼裕司，他：脆弱性を利用した標的型攻撃のための解析ツール，MWS2008，pp.127-132，2008

12）齋藤彰一，他：シェルコード解析と分類による不正侵入防止システムへの適応可能性の評価，MWS2008，pp.1-6，2008

13）勝手壮馬，他：研究用データセットにおける攻撃通信データによるマルウェア解析の一結果，MWS2008，pp.7-12，2008

14）東角芳樹，他：DNS 通信の挙動からみたボット感染検知方式の検討，MWS2008，pp.13-18，2008

15）水谷正慶，他：通信の状態遷移に着目したボット活動の調査，MWS2008，pp.25-30，2008

16）仲小路博史，他：パケット送受信における同調活動に着目した ボット感染ノードへの指令および反応活動の可視化，MWS2008，pp.31-36，2008

17）松木隆宏：時系列分析による連鎖感染の可視化と検体種別の推測，MWS2008，pp.37-42，2008

18）三原元，他：ボットネットの多段追跡システムの構想と CCC DATAset 2008 の利用手法，MWS2008，pp.43-48，2008

19）竹森敬祐，他：ボットネットおよびボットコードセットの耐性解析，MWS2008，pp.49-54，2008

20）小櫻文彦，他：ウイルスのライフサイクルに着目した攻撃挙動の見える化，MWS2008，pp.55-59，2008

21）永田大，他：OLAP(多次元データ分析)を利用した攻撃元データの分析と検体の自動解析，MWS2008，pp.61-66，2008

22）小堀智弘，他：マルウェアの通信履歴と定点観測の相関について，MWS2008，pp.67-74，2008

23）金井瑛，他：マルウェアの転送ログを利用したボットの活動分析，MWS2008，pp.75-80，2008

24）高田哲司，他：人間による Honeypot の攻撃元ログ調査を支援する User Interface の提案，MWS2008，pp.81-86，2008

25）畑田充弘，他：複数観測データを用いたボットネットの活動分析に関する一考察，MWS2008，pp.87-92，2008

26）永尾禎啓，他：観測網の大小に基づく結果の比較とその差異に関する一考察，MWS2008，pp.93-95，2008

27）石井宏樹，他：ダウンロードホストに着目したマルウェアの活動傾向分析，MWS2008，pp.97-102，2008