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

勝手 壮馬 †　　安本 幸希 †　　曽根 直人 ‡　　森井 昌克 †

† 神戸大学大学院工学研究科
657-8501 兵庫県神戸市灘区六甲台町 1-1

skatsute@stu.kobe-u.ac.jp　yasumoto@stu.kobe-u.ac.jp　mmorii@kobe-u.ac.jp
‡ 鳴門教育大学 高度情報研究教育センター
772-8502 徳島県鳴門市鳴門町高島 748

naosone@naruto-u.ac.jp

あらまし　本調査解析ではハニーポットで捕らえられた通信データからマルウェアの実行ファイルを復元し，それらに対してバイナリデータの類似性による分類やメモリダンプによる静的解析を適用した．今回の解析により，UNKNOWN と分類されたいくつかのマルウェアは「脅威なし」と判定されるが実際にはマルウェア的な挙動を示すことが判明した．また類似性による分類を用いることでさらにいくつかの小グループへと分類可能であった．

A result of Malware analysis in CSS2008 dataset

Soma Katsute†　　Koki Yasumoto†　　Naoto Sone‡　　Masakatu Morii†

†Graduate School of Engineering, Kobe University
1-1, Rokkodai, Nada-ku, Kobe, Hyogo 657-8501, Japan

skatsute@stu.kobe-u.ac.jp　yasumoto@stu.kobe-u.ac.jp　mmorii@kobe-u.ac.jp
‡Advanced Information Research and Education Center, Naruto University of Education
748 Naruto-cho, Naruto, Tokushima 772-8502, Japan

naosone@naruto-u.ac.jp

Abstract

In this paper, we restore executable files of the malware from the communication data captured by the honey pot. In addition, we apply our classification technique of the malware and static analyses of the malware to them. By our analysis, malwares assumed to be UNKNOWN were classified into small groups with some similarities.

1　はじめに

本調査解析ではサイバークリーンセンター (CCC) で収集されたボット観測研究用データセット “CCC DATAset 2008” にて提供される

(1) マルウェア検体

(2) 攻撃通信データ

(3) 攻撃元データ

のうち，(2) 攻撃通信データおよび (3) 攻撃元データを利用しマルウェアの分類や解析を試みた結果，得られた知見について述べる．利用したデータの概要を表 1 に示す．

本稿ではまず具体的な攻撃通信データの解析手順について説明する．次に解析によって得られた結果を示し，最後に考察を行う．

2　攻撃通信データの解析手順

攻撃通信データの解析は以下の手順で行った．

1. **実行ファイルの復元**
   攻撃通信データから実行ファイルを復元

表 1: 利用したデータの概要

CCC DATAset 2008	
(2) 攻撃通信データ	• ハニーポット 2 台への通信を 2 日分フルキャプチャしたデータ (Win2 k, WinXP SP1+2005 年末までにリリースされたパッチ適用済み，ネットワーク接続環境：FTTH，動的 IP アドレス) • 収集日：2008 年 4 月 28 日〜29 日 • 提供形態:バイナリ形式，日毎に 1 ファイル，計 2 ファイルで約 2.8GB
(3) 攻撃元データ	• ハニーポット 112 台による 6ヶ月間のマルウェア取得時のログデータ (時刻, ダウンロードホスト IP アドレス, 利用ポート番号／プロトコル、通信方向、ハッシュ値 (SHA1)、ウイルス名称、ファイル名) • 収集日： 2007 年 11 月 1 日〜2008 年 4 月 30 日 • 提供形態：テキスト形式、1 ファイルで約 390MB(294 万レコード)

する．

2. **攻撃元データとの対応**
   復元した実行ファイルと攻撃元データとの対応を行う．

3. **バイナリコードの比較による分類**
   バイナリコードにより，復元した実行ファイルを分類する．

4. **メモリダンプを用いた静的解析**
   実行ファイルのメモリダンプを取得し静的解析を行う．

それぞれの手順での詳細については以下の節で述べる．

2.1 実行ファイルの復元

攻撃通信データ (pcap 形式，今回の解析では 4 月 28 日分を利用) から以下に示す．手順に従いダウンロードされたバイナリファイルを求める．

step1 “CCC DATASet 2008” にて提供される pcap ファイルに対して snort(ルールセット:snapshot-2.8) を適用し，アラートを得る．

step2 得られたアラートを snortalog にて処理し，レポートを作成する．

step3 レポートから対象となるホストの IP アドレスを求める．

step4 対象となる IP アドレスとの通信を攻撃通信データから切り出す．

step5 切り出したファイルからバイナリファイルを構成する．このとき攻撃元データに記録されたダウンロード記録と照合するための目安としてバイナリファイルのパケットストリームの開始時間を取得する．

step3 で求めた全ての IP アドレスに対して step4,5 の処理を行うことでダウンロードされた全てのバイナリを取得することができる．

今回の解析では上記の step1〜step5 で得られたバイナリファイルを復元に成功した実行ファイルとして取り扱う．

2.1.1 バイナリファイルの構成

step5 では，step4 によって切り出した特定のホストとの通信データから，TCP/UDP のそれぞれのセッションを取り出し，それぞれのセッションの先頭パケットからシグネチャーとのマッチングを行う．マッチングに成功した場合，マッチした位置からセッションの終了までを一つのバイナリファイルとして復元する．

なお，通信データにおいて図 1 に示したように 1 セッションで複数のファイルが含まれていた場合，正常にファイルが抽出されない可能性がある．今回の解析ではこのようなファイルについて考慮していない．

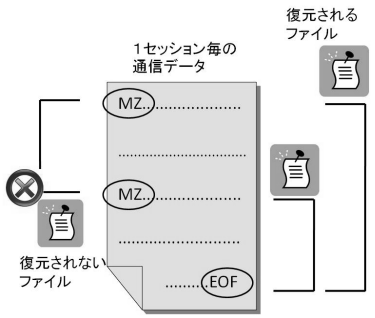

図 1: バイナリファイルの再構成方法

2.2 攻撃元データとの対応

2.1 節で抽出した実行ファイルはハニーポットの通信データから取り出している．したがってハニーポットにダウンロードされたマルウェアを記録している攻撃元データにも対応するマルウェアダウンロード時のデータが記録されていると考えられる．2.1 節の操作で得た実行ファイルが適切にダウンロードファイルを抽出しているか確認を行うため，攻撃元データとの対応付けを行なった．

2.1 節で抽出した実行ファイルと攻撃元データとの対応付けは容易ではない．攻撃元データに記載されているハッシュ値と得られた実行ファイルのハッシュ値が一致しないためである．2.1 節の抽出アルゴリズムでは得られた実行ファイルにはその本質的な内容に関係しないデータ列が加わっており，それを取り除くことは容易ではないからである．

攻撃元データと 2.1 節で得た実行ファイルを対応付けるために

- ダウンロードが行われた IP アドレス
- ファイルがダウンロードされた時間
- マルウェア名

を 2.1 節で得たホスト IP アドレスおよび実行ファイルのダウンロード開始時間を比較することによって可能とした．

対応付けの一例として，実行ファイル “IPaddr_A.exe” を攻撃元データと対応づける．この実行ファイルはハニーポットと IPaddr_A の通信から得られたもので，ファイルが含まれた通信セッションの開始時刻は “2008-04-28 01:14:01” であった．また，この実行ファイルをトレンドマイクロ社の提供するオンラインスキャンで検査したところ，“Mal_Allaple” というマルウェアであると判定された．

攻撃元データからダウンロードホスト IP アドレスとして IPaddr_A が記録されているレコードを検索では 2 件の一致するレコードが存在し，それぞれ時刻は “2008-04-28 01:14:01”，マルウェア名は “Mal_Allaple” と記録されていた．これは我々が攻撃通信データから得た実行ファイル “IPaddr_A.exe” の検査により得た情報と時刻，マルウェア名が一致していることを確認した．

2.3 バイナリコードの比較による分類

マルウェアの解析は逆アセンブルなどを用いた静的解析や隔離した検証環境で挙動を観測する動的解析によって行なわれている．しかし，そのような解析には時間がかかるため，検体を入手してもそれが新種のマルウェアなのか既存マルウェアの亜種であるのかを判断することは困難である．そこで時間やコストをより削減させる分類手法としてバイナリコードの比較による分類が提案されている [1]．

バイナリコードの比較による分類では，検体が未知のマルウェアであったとしても既に解析されているマルウェアの亜種であった場合，利用する API には類似性があり，バイナリコードが一致する割合が高くなると仮定し，分類を行う．本節では 2.1 節で得た実行ファイルを時間や経験が必要となるマルウェアの静的解析ではなく，プログラムにより自動化されたバイナリコードの比較手法による分類について述べる．

2.3.1 バイナリコード比較アルゴリズム

実行ファイル間で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う．

実行ファイル間の類似度を判定するために図 2 に示すアルゴリズムにより，バイナリコードの類似度を求める．

```
1: files[] = { 実行ファイル 0,..., 実行ファイル n}
2: for x in (0 ... n)
3: for y in (0 ... n), y ≠ x
4: 一致率 = 0
5: for i in (1 .. t)
6: sample[] = files[x] から l バイトの検査コードをランダムに m 個抽出
7: 一致率 += files[y] に sample[] のコードが含まれる割合
8: end
9: 類似度 = average(一致率)
10: end
11: end
```

図 2: バイナリコード比較アルゴリズム

本稿では各試行で抽出する検査コードの個数 m=100，長さ l=16 バイト，試行の繰り返し回数 t=5 回に設定し解析を行った．

2.4 メモリダンプを用いた静的解析

マルウェアのコードには暗号化 (パッキング) が施されているものが存在する．これらのマルウェアは直接バイナリコードを読み取ることができないため，通常の静的解析手法を適用できない．このような暗号化を施されたマルウェアの解析手法の一つとして，これまで筆者らが開発してきたメモリ上に展開されたコードを利用するシステムを用いることができる [2].

本節ではマルウェアを実行することでメモリ上に展開される実行ファイルコード（メモリダンプ）を取得し，逆アセンブルによってメモリダンプから実行ファイルが使用する API の抽出を行う解析手法について述べる．解析手順を以下に示す．

1. マルウェアのメモリダンプ取得
: マルウェアを実行し，メモリ上に展開されるダンプコードを一定時間間隔で取得する．最後に取得したメモリダンプの逆アセンブルを行う．

2. 逆アセンブルコードの解析
: 逆アセンブルコードの先頭から命令に従い走査する．マルウェアが実行する順に API とその引数 (文字列) を抽出する．

3. 解析レポートの作成
: マルウェアの動作に関するレポートを作成する．レポートには，ファイルやレジストリ操作, メールの送信，使用するネットワークプロトコル等マルウェアの動作把握に有用な情報が記述される．

3 解析結果

本章では第 2 章で述べた解析手法によって得られた結果を示すと共に，取得した結果に応じた解析の方針について述べる．

3.1 攻撃通信データの解析

2.1 節の手法により攻撃通信データから得られた結果を表 2 に示す．

攻撃通信データに対して snort を適用することで 13398 個の攻撃アラートが報告された．攻撃アラートから攻撃元の IP アドレスを抽出し 6242 個の IP アドレスを得た．攻撃通信データから取得した 6242 個の IP アドレスとの通信をそれぞれ切り出し，実行ファイルの抽出作業を行なった．その結果 914 個の実行ファイルが得られた．但しそのうち 97 個の実行ファイルはサイズが数十バイト以下と非常に小さく，解析に適さないものと判断し解析の対象外とし，残りの 817 個の実行ファイルを解析の対象として扱った．

表 2: 攻撃通信データからの実行ファイル抽出

	個数
検出した攻撃アラート	13398
取得した IP アドレス	6242
抽出した実行ファイル	914
除去した実行ファイル	97
解析対象実行ファイル	817

3.2 攻撃通信データとの対応

表 2 に示した解析対象実行ファイルに対してトレンドマイクロ社の提供するオンラインスキャンを実施し，解析対象実行ファイルの判定を行った．オンラインスキャンの判定結果を表 3 に示す．

347 個の実行ファイルはオンラインスキャンにより，マルウェアの種類が特定された．残りの 470 個については「脅威がない」ファイルとして判定された．ここで，これらのファイルを「UNKNOWN 候補ファイル」と定義する次節では「脅威がない」と判定された UNKNOWN 候補ファイルについて攻撃元データで UNKNOWN と分類されているものとの対応付けを行い，さらに解析を行う．

3.3 UNKNOWN ファイルの解析

攻撃元データで UNKNOWN と分類されるファイルの解析を目的とし，攻撃通信データからの抽出を試みた．

攻撃通信データ取得期間中 (2008 年 4 月 28 日) を対象として，攻撃元データから UNKNOWN ファイルのダウンロードが行なわれた IP アドレスを検索することにより 174 個の IP アドレスを得た．さらに既に抽出済みの UNKNOWN 候補ファイルと 174 個の IP アドレスの対応付け作業を行った結果，UNKNOWN 候補中 119 個の

表 3: オンラインスキャンによる定義結果

	個数
PE_VIRUT.A	85
WORM_BOBAX.BD	58
PE_BOBAX.AK	54
PE_BOBAX.AH	21
PE_VIRUT.D-1	20
PE_VIRUT.D-4	14
PE_VIRUT.D-2	13
PE_VIRUT.XW	12
BKDR_VANBOT.AHH	9
TROJ_ANOMALY.MB	7
WORM_KOLAB.AX	7
PE_VIRUT.B	7
TROJ_DLOADER.ETK	6
WORM_RBOT.QK	6
WORM_VANBOT.AX	5
その他 (28 種)	58
脅威がない (UNKNOWN 候補)	470
合計	817

表 4: UNKNOWN とされるファイル

	個数
時間，IP アドレス共に一致	119
時間不一致	69
合計	188

実行ファイルが攻撃元データでは UNKNOWN と分類されたファイルに相当すると考えられる (表 4).

UNKNOWN と対応付けられた 119 個の実行ファイルはオンラインスキャンによる判定では全てのファイルについて「脅威はない」という結果が得られている．しかし，これらのファイルはハニーポットに対してダウンロードされており，何らかの挙動を示すことが考えられる．そこで我々は得られた実行バイナリに対してさらに解析を行い，脅威の有無について検証を行った．

3.3.1 バイナリデータの比較による分類

前節で UNKNOWN と対応付けられたバイナリファイル 119 個 (表 4) に対して，バイナリコードの比較による分類を適用し，いくつかのグループへの分類を試みた．コードの一致率が 80%以上のバイナリを同じグループと分類す

表 5: コード比較による UNKNOWN ファイルの分類

No	ファイルが抽出された IP	個数
Group1	IPaddr_U1	52
Group2	IPaddr_U2	38
Group3	IPaddr_U2	7
Group4	IPaddr_U3	4
Group5	IPaddr_U4	4
Group6	IPaddr_U4	2
Group7	IPaddr_U3	2
Group8	IPaddr_U3	2
Group9	IPaddr_U4	2
Group10	IPaddr_U2	2
Group11	IPaddr_U4	1
計		117

ることで表 5 に示す[1]ように 119 個のバイナリファイルは 11 グループへ分類された．表 5 中の IPaddr_U1〜IPADDR_U4 はそれぞれバイナリファイルが抽出された IP アドレスを表している．

3.3.2 静的解析による分析

UNKNOWN ファイル 119 個に対してメモリダンプによる静的解析を行った．その結果，47 個のファイルについてはメモリダンプから API の取得が可能であった．

例えば UNKNOWN ファイルの一つである，”IPaddr_U3.exe” からは以下の情報が得られた．

- 書き込み・参照するレジストリ
  `SOFTWARE\Microsoft\\Windows\\CurrentVersion\\Explorer`
  Trend 定義：TROJ_SPYWAD.B が書き込みを行うレジストリに一致
- 作成するファイル removalfile.bat
  Trend 定義：TROJ_DLOADER.DIY が作成するファイル名に一致
- 文字列として IRC で使われる文字列を持つ．
  “NICK ikullytw”（使用するニックネームと考えられる）

[1]表 5 の合計が 119 ではなく 117 となっているのは解析中のエラーにより 2 ファイル分の類似度データを失ったためである．

“proxim.ntkrnlpa.info”（ボットが接続する IRC サーバ名）
“send connect receive”(パケットのやり取りを行う関数名)

4 考察

“CCC DATASet 2008” の解析を行なうことにより次の考察を与える．

- 攻撃通信データの解析により実行ファイルを取得し，攻撃元データとの対応付けから UNKNOWN と判定されるファイルを取り出し，解析を行った．その結果，既存のマルウェア対策ソフトでは「脅威なし」と判定されたファイルもマルウェア的な挙動が見られることが判明した．
- バイナリコードの類似度判定システムにより，UNKNOWN と判定されたファイルの分類を行った．その結果，コードがよく似ているファイルはダウンロード先の IP アドレスが同じであった．ファイルが UNKNOWN と判定されていることから，まだ一般には広く拡散していないマルウェアの一種であると考えられるが，既にその時点で複数の亜種と共にダウンロードが行われている．つまり，マルウェア作者は初期の段階から複数の亜種を作成し，拡散を行っていると考えられる．
- バイナリコードの類似度判定システムを用いることにより，未知の検体であっても過去に解析された検体との類似性を指摘できる．これにより，マルウェア判定の誤検知 (false negative) を減少させられると考えられる．

5 むすび

本調査解析では，サイバークリーンセンター (CCC) によって収集された “CCCDATASet 2008” から実行ファイルの抽出および解析を行った．

得られた実行ファイルに対し，バイナリデータの類似性を用いた分類を行った．また実行ファイルのメモリダンプを取得し，逆アセンブルによる静的解析を行った結果から，CCC によるデータ取得環境で UNKNOWN と分類されるマルウェア実行ファイルを，性質毎により小さなグループへと分類可能であることを示した．

参考文献

[1] 安本幸希，森井昌克，中尾康二，“コードの類似度判定を用いたマルウェア分類法”, コンピュータセキュリティシンポジウム (CSS2008) 予稿集，2008 年 10 月．

[2] 岡田隼人, 伊沢亮一, 森井昌克, 中尾康二, “ウイルスコード自動解析システムの開発,” 2007 年 暗号と情報セキュリティシンポジウム (SCIS2007) 予稿集, 2007 年 1 月．