時系列分析による連鎖感染の可視化と検体種別の推測

松木　隆宏 †

†株式会社　ラック　サイバーリスク総合研究所
〒 105-7111　東京都港区東新橋 1-5-2　汐留シティセンター 11 階

あらまし　研究用データセット CCC DATAset 2008 の攻撃通信データと攻撃元データを重ね合わせた時系列分析によって，近年増加しているダウンローダ等を用いた連鎖感染の可視化とを検体の種類の推測を行う．これにより，解析の対象選定や優先度付け，連鎖感染マルウェアによる脅威の全体像の把握を試みる．
キーワード　マルウェア，ボット，時系列分析，可視化

Visualization of Chain Infection and Guess of Sample Classification by Time Series Analysis

Takahiro Matsuki †

†Little eArth Corporation(LAC) Co., Ltd.
Shiodome City Center 11F, 1-5-2, Higashi Shinbashi, Minato-ku, Tokyo, 105-7111, Japan

Abstract　The infection of chain-driven malware is increasing through the downloader, presently. In this paper, we've analysed traffic data and sources of attack in a way of time series and visualization. In addition, we've identified the perspective of the infection of chain-driven malware.
Keywords　Malware，Bot，Time Series Analysis，Visualization

1　はじめに

昨今，ダウンローダーという別のマルウェアをインターネットからダウンロードして感染させることに特化したマルウェアが急増している．これにより，マルウェアによる脅威は，高度に複雑化され，脅威となるマルウェアやその配布元サイト，攻撃者の存在が隠蔽されている．このようなマルウェアの感染は，Web アプリケーションとクライアントアプリケーションの脆弱性が合わせて悪用されている例が多く見られるが，ボットネットによって新たなマルウェアが拡散されることも懸念される．攻撃者は，ダウンローダーやボットネットを用いて任意のタイミングで新たなマルウェアを感染させるインフラを構築している可能性がある．

ダウンローダーに感染した場合，時間経過とともに次々に新種のマルウェアに感染，アップデートされる恐れがあり，大きな脅威である．日々出現する新種のマルウェアの数は年々増加し続けており，ウイルス対策ベンダの解析負荷は高まっている．これに対抗してベンダ側は，検体の解析を自動化するシステムなどを開発，利用することによって迅速なパターンファイルの配布を実現している．しかし，自動化された解析では，ウイルス対策ベンダ間で解析結果や検体に付与される名称にばらつきが生じることが懸念される．また，標的型攻撃に対する調査

でも，自動解析では，ダウンローダーを用いた連鎖感染マルウェアによる脅威の本質が分析しきれない，的確な対応ができない場合があるという指摘もある [5].

2 CCC ハニーポットの傾向

サイバークリーンセンターでは，活動実績レポートを毎月公開している．このレポートによると，2007 年の後半から，ポリモーフィック機能を持ち，かつファイル感染型のマルウェア（代表例：PE_BOBAX.AK [1], PE_VIRUT.D [2] [3]）が増加している [4]．これらは，他の実行可能ファイルに感染する性質とポリモーフィックという性質のため，マルウェアとしての本質的な機能は変化していないにも関わらず，ファイルハッシュが異なる検体が多数出現し，ウイルス対策ソフトウェアで検知できない亜種も存在する．これらの検体によって，他の未知検体の解析が妨げられるという問題が考えられる．未知検体に対してその種類が推測できれば，解析の優先度付けができ，新たな脅威となるマルウェアを優先的に解析できる可能性がある．

本稿では，研究用データセット CCC DATAset 2008（以降、CCC2008 データ）の攻撃通信データと攻撃元データ (以降、攻撃通信データ、攻撃元データ) を重ね合わせ，主に検体の感染時間に着目して時系列分析を行った．これにより，ダウンローダー等によるマルウェアの連鎖感染の分析と可視化および未知検体の種類の推測を試みた．

3 時系列分析

3.1 分析対象データ

CCC2008 データの攻撃通信データには，ハニーポットの識別情報（IP アドレス）が含まれているが，検体名称が含まれていない．攻撃元データには，ウイルス対策ソフトウェアによる検体の名称が含まれている．そのため，攻撃通信データから TCP および UDP のセッション情報を抽出，また，攻撃元データから，2008 年 4 月 28 日と 2008 年 4 月 29 日分のデータを抽出し，これらのデータの時刻情報と IP アドレス，ポート番号等の情報を照合することで，2 つのハニーポット（以降，honeypot1，honeypot2）における 2008 年 4 月 28 日，2008 年 4 月 29 日のマルウェア感染ログを作成し，これを分析の対象データとする．作成したデータの概要を表 1 に示す.

表 1: 分析対象データの概要

IP	日付	攻撃数	検体種類	未知種類
honeypot1	4/28	280	40	5
	4/29	475	55	4
honeypot2	4/28	362	41	7
	4/29	427	54	9

3.2 連鎖感染の判定法

対象データから連鎖感染を定義するパラメータとして以下の 5 つが考えられる．

1. **感染時間の間隔**
   感染の時間間隔に閾値を定め，一定時間内であれば連鎖と判定する．あるいは，一定の周期で感染しているものを連鎖と判定する．

2. **攻撃元 IP アドレスの一致度，パターン**
   攻撃（ダウンロード）元の IP アドレスが同一で感染が連続していれば連鎖と判定する．あるいは，IP アドレスの出現パターンに一定の特徴が見られる場合，連鎖と判定する．

3. **ソースポート番号の連続性**
   ソースポート番号が連続していた場合，連鎖と判定する．あるいは，ソースポート番号に一定のパターンが見られる場合，連鎖と判定する．

4. **検体名称**
   検体名称の出現パターンに一定の特徴が見られる場合，連鎖と判定する．

5. **検体ファイルサイズ**
   検体のファイルサイズの遷移に一定の特徴が見られる場合，連鎖と判定する．

各々のパラメータの変化や関連性を統計的に調査分析することで，連鎖感染の判定を数式化できる可能性があるが，いずれのパラメータについても，特定のパターンや条件を抽出するには，2 ハニーポットの 2 日間のデータでは十分でないと考えられる。

そのため，本稿では，最も単純な連鎖感染の判定法として，上記 1. の感染時間の間隔に着目し，対象データの時系列分析を行った．マルウェア検体の動的解析において，限られた時間の中で大量のマルウェアを解析するためには，検体の実行時間を制限せざるおえない．動的解析における検体の実行時間は，経験的に 3 分程度で C&C への接続や別のマルウェアの自動的なダウンロードといった挙動の大半が分析できるという知見がある．このことから，感染時間の間隔の閾値を 3 分に設定し，ある任意のマルウェアが感染してから 3 分以内に感染した別のマルウェアの感染を連鎖とみなした．この条件で抽出される連鎖感染の概要を表 2 に示す．

表 2: 連鎖感染の概要

IP	日付	連鎖の数	平均	最大
honeypot1	4/28	53	3.42	9
	4/29	96	2.95	10
honeypot2	4/28	69	3.16	20
	4/29	110	2.97	10

4 連鎖感染の可視化

マルウェアが連鎖感染する場合，その脅威は複数のマルウェアによって実現されると考えられる．しかし，ウイルス対策ベンダの提供する情報では，複数のマルウェアの関係性を一見して把握することが難しく，利用者にとってわかりづらいものとなっている．そこで 3 章で抽出した連鎖感染のデータをグラフとして可視化することで，複数のマルウェアの関係性や連鎖感染の傾向を把握する．グラフの凡例を 3 に示す．

表 3: グラフ凡例

	色	説明
ノード	赤	未知検体
	紫	名称に BOT を含む
	緑	PE ファイル感染型
	茶	トロイの木馬
	灰	その他
エッジ	赤	ポート 80
	黒	ポート 80 以外

4.1 連鎖感染ツリー

まず，個々の連鎖感染データを並べるグラフを連鎖感染ツリーとし，マルウェアの感染順序や連鎖感染全体の傾向を調査した．

honeypot1 における連鎖感染ツリーを図 1，図 2 に示す．同様に honeypot2 について図 3，図 4 に示す．

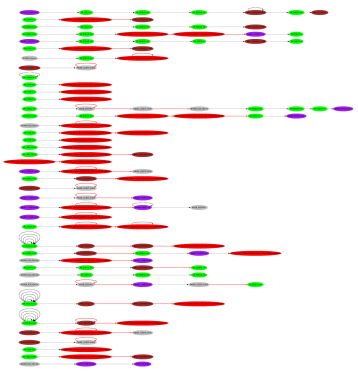

図 1: honeypot1 の連鎖感染ツリー (4 月 28 日)

これらの連鎖感染ツリーから，以下の傾向がうかがえる．

- 連鎖の最初（1 番目）の検体は，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にもかかわらず，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ポート 80 番を使用する割合が高い．

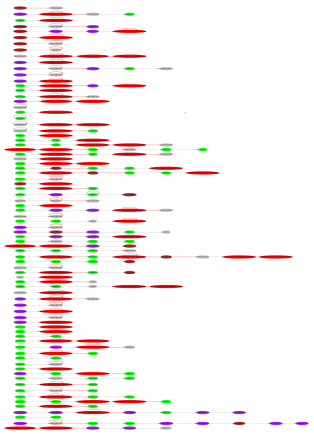

図 2: honeypot1 の連鎖感染ツリー (4 月 29 日)

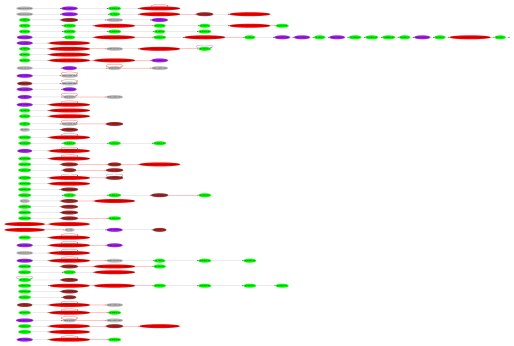

図 3: honeypot2 の連鎖感染ツリー (4 月 28 日)

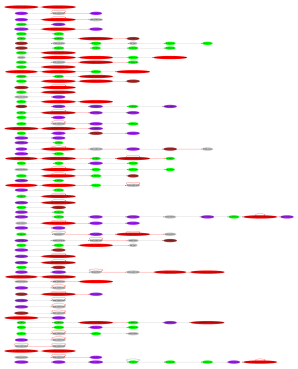

図 4: honeypot2 の連鎖感染ツリー (4 月 29 日)

4.2 連鎖感染マップ

次に連鎖感染ツリーを重ね合わせたグラフを連鎖感染マップとし，マルウェア同士の関連性を調査した．

honeypot1 における連鎖感染マップを図 5，図 6 に示す．同様に honeypot2 について図 7，図 8 に示す．連鎖感染マップから，以下のことが言える．

- 既知の検体と関連性が全くない未知検体の連鎖がある．(図 5 および図 8 の左上)
- マップの内部にある未知検体は多数の既知検体と関連している．
- BOT はマップの全域に点在している．
- 全体の半数以上はポート 80 で連鎖している．

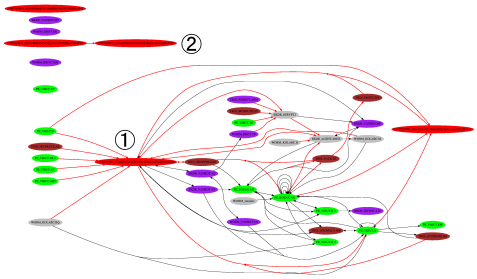

図 5: honeypot1 の連鎖感染マップ (4 月 28 日)

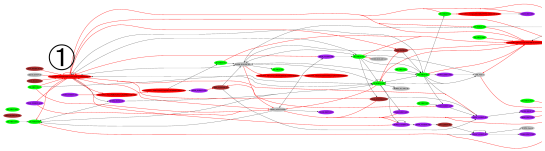

図 6: honeypot1 の連鎖感染マップ (4 月 29 日)

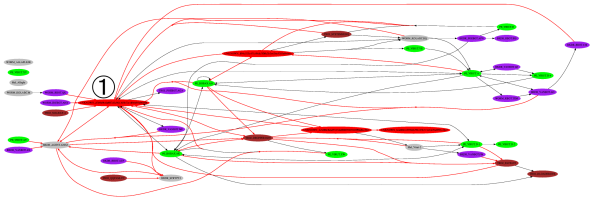

図 7: honeypot2 の連鎖感染マップ (4 月 28 日)

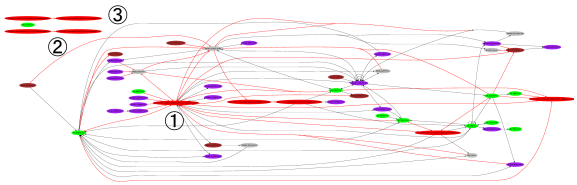

図 8: honeypot2 の連鎖感染マップ (4 月 29 日)

4.3 未知検体種類の推測

4 つの連鎖感染マップにおいて視覚的特徴を抽出できた以下の未知検体について種類の推測を試みた（表 4）.

表 4: 未知検体種類の推測

ハッシュおよび推測理由
d7b9b9b10d9f7d2c961365b72e189eb95a9f03f8 ① 4 つのマップ全てに出現し，多数の検体と関連しているため，PE ファイル感染型と推測
5037c080b4343d2d2e37c42d489ffae3866df1dc と 2b4a6bf8b9ef1c8394f7d28b29c5bbd3000ab799 の連鎖② 図 5 と図 8 に出現しており，多数の検体 と関連がなくマップの中心から外れている. 図 5 において同様の既知検体として BOT 型 の検体が存在することから BOT 型と推測
16d7e55cd173f6196cd06bebcc2bd9cb48d6856f と 52bab16ea6de92636f6ca17a5414edd1b6058e92 の連鎖③ 図 8 において，多数の検体と関連がなく マップの中心から外れている. 上記と同様に BOT 型と推測

その他のマップの内部に位置する未知検体については，目立った視覚的特徴が抽出できず，種類の推測が困難であった.

検体種類の推測が正しいかどうか検証するため，未知検体のハッシュ値を VirusTotal[6] の Hash Search を利用して調査を行った．結果を表 5 に示す.

調査の結果，未知検体のほぼ全てが接頭辞 TROJ がつくトロイの木馬型であった.
グラフ上で最も特徴的な検体であった d7b9b9b10d9f7d2c961365b72e189eb95a9f03f8 は，HOSTS ファイルを改変し，セキュリティベンダのサイトなどへのアクセスを妨害するもので

表 5: 未知検体の名称調査結果

ハッシュおよび名称
16d7e55cd173f6196cd06bebcc2bd9cb48d6856f TROJ_STARTPA.OO
2b4a6bf8b9ef1c8394f7d28b29c5bbd3000ab799 TROJ_STARTPA.PB
4c5d88c8a6d5547da0f08f5385203ff9ddbc49e1 TROJ_BUZUS.ER
5037c080b4343d2d2e37c42d489ffae3866df1dc BKDR_IRCBOT.AXA
5212ae4a28315df0a325f791c38a0d1d587dc5e2 TROJ_DROPPER.BNL
52bab16ea6de92636f6ca17a5414edd1b6058e92 TROJ_BUZUS.ES
74d2ce9d8fa7bdf3ea7c2faef9f0fc5738f774ab TROJ_BUZUS.ES
7fddf4269da2975d05d457b93d8d8923890752e0 TROJ_VUNDO.BJN
894c525c471c94dc7f08c2b3e636e5203af46bea TROJ_DROPPER.BNL
bca08616f2966b29b135b721e34e56df5f1a1ba7 TROJ_VUNDO.BUA
d7b9b9b10d9f7d2c961365b72e189eb95a9f03f8 TROJ_QHOST.LD

あった．
5037c080b4343d2d2e37c42d489ffae3866df1dc については，名称が BKDR_IRCBOT.AXA であり，BOT 型という推測は正しいといえるが，可視化情報を一見して未知検体の種類を推測することは困難であるといえる．既知検体の名称と未知検体の関連性はマップからは読み取りづらいため，未知検体の種類推測には，蓄積した連鎖感染のデータや可視化データの統計的な分析が適していると考えられる．ただし，連鎖感染データを可視化することによって，多数の既知検体と関連するもの，あるいは，既知検体との関連性が著しく低いものなど，未知検体の解析に優先度をつけることができる．

5　まとめ

本稿では，マルウェアの連鎖感染の抽出の 1 方法の検討と，得られたデータの可視化を行った．感染時間の間隔によって抽出した連鎖感染データの可視化によって以下の傾向が把握できた．

- 連鎖感染の起点は，PE_BOBAX.AK や PE_VIRUT.D などの 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にもかかわらず，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．
- 全体の半数以上はポート 80 で連鎖しており，未知検体からの連鎖はポート 80 番を使用する割合が高い．
- 未知検体はトロイの木馬型のものが多く，HOSTS ファイル書き換えなどでセキュリティ機能を無効化を狙うものもある．

本研究では，攻撃元データにハニーポット識別情報が含まれないため，攻撃通信データを起点としてデータを分析したが，攻撃通信データだけでは見ることのできない，DROPPER 型の検体も存在すると考えられ，それらも連鎖感染の 1 ノードとして分析する必要があると考える．今後，研究用データセットとしてより多くの情報がオープン化されることを期待したい．

今後の課題としては，連鎖感染データの統計分析方法の確立や連鎖感染の中の特徴抽出を動的に行える可視化システムの検討などが考えられる．

謝辞

本研究は，財団法人 日本データ通信協会 Telecom-ISAC Japan ならびにサイバークリーンセンターの支援を受け実施している．本研究を進めるにあたり，有益な助言と協力を頂いた Telecom-ISAC Japan と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] PE_BOBAX.AK - 概　要
http://www.trendmicro.co.jp/Vinfo/virusencyclo/default5.asp?VName=PE_BOBAX.AK

[2] PE_VIRUT.D - 概　要
http://www.trendmicro.co.jp/vinfo/virusencyclo/default5.asp?VName=PE_VIRUT.D

[3] 日本 F-Secure 株式会社 ： ウィルス情報 Virus:W32/Virut
http://www.f-secure.co.jp/v-descs/v-descs3/W32.Virut.htm

[4] サイバークリーンセンター（CCC）| 2008 年 02 月度　サイバークリーンセンター活動実績 https://www.ccc.go.jp/report/200802/0802monthly.html

[5] 独立行政法人　情報処理推進機構, 近年の標的型攻撃に関する調査研究
http://www.ipa.go.jp/security/fy19/reports/sequential/seq_rep.pdf

[6] VirusTotal
http://www.virustotal.com/