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

小櫻 文彦　津田 宏　鳥居 悟

株式会社富士通研究所　ソフトウェア＆ソリューション研究所

〒211-8588 神奈川県川崎市中原区上小田中 4-1-1

E-mail: {kozac, htsuda, torii.satoru}@jp.fujitsu.com

あらまし　研究用データセット CCC DATAset 2008 の攻撃元データの分析において、ウイルス名称が付与される前後の挙動変化に注目し、攻撃元や地域などとの関連を見える化することでウイルスの攻撃挙動と現状の対策について傾向を探る。

キーワード　マルウェア，ボット，ハニーポット，ログ

Visualization of malware attacks based on their lifecycle

Fumihiko KOZAKURA, Hiroshi TSUDA, and Satoru TORII

Software and Solution Laboratories, Fujitsu Laboratories LTD.

4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki, 211-8588 Japan

E-mail: {kozac, htsuda, torii.satoru}@jp.fujitsu.com

Abstract　In the attack host log of CCC DATAset 2008, we investigate the malware activities before and after they are named by virus vendors and visualize the relation with attack host, area, etc.

Keyword　malware，bot，honey pot，log

1. はじめに

インターネット上での不正アクセス活動は活性化するとともに、マルウェアの挙動も巧妙になっている。最近では、イントラネットのシステムを狙い、利用者に特定の行動を誘導する受動的攻撃のマルウェアによる被害も深刻化している[1]。こうした多様な不正アクセス活動に対して、その動向を知り、それに基づいた対策が求められている。

本論文では、サイバークリーンセンター (https://www.ccc.go.jp/) が収集した研究用データセットCCC DATAset 2008 の攻撃元データ(**以降、CCC2008攻撃元データ**)において、ウイルスのライフサイクルに着目した分析結果を報告する。ウイルスが発生し、ウイルス対策ベンダーで名称が付与され対策パターンが作られ、終息するまでのライフサイクルにおいて、特にウイルス名称が付与される前後の挙動に注目し、攻撃元や地域などの関連情報を見える化することで攻撃や対策の動向を分析する。

2. 関連研究

インターネットにおいては、例えば警視庁@police [7] においてファイアウォールのログなどの定点観測が行われており、攻撃手法や国別の統計が公開されている。鬼頭ら[2]は、不正アクセス元のホストの地域、時間帯による変化を分析している。また、ラック社では、マルウェア用のハニーポットのログから攻撃元の地域を Google Map 上にマップして 2008/8/5 より公開している[3]。一般に日本におけるマルウェアの攻撃元は、国内、カナダ、中国、アメリカと言われる[1]。

3. ウイルスのライフサイクル分析

今回我々は、これまで多く研究が行われていた全体的、時系列的な攻撃の傾向ではなく、個々のウイルスのライフサイクルに着目した分析を行うこととした。

3.1. CCC2008 攻撃元データの概要と特徴

CCC2008 攻撃元データは、ハニーポット 112 台による 2007 年 11 月 1 日から 2008 年 4 月 30 日の 6ヶ月間のマルウェア取得のログデータが含まれる。件数は 2,942,221 件。項目は、

(時刻、ダウンロードホスト IP アドレス、利用者ポート番号/プロトコル、通信方式、ハッシュ値(SHA1)、ウイルス名称、ファイル名)

である。ここでハッシュ値はウイルスのコードから計算されるものである。分析では、同一ハッシュ値は同一のウイルスコード、異なるものは異なるウイルスコードと仮定する。

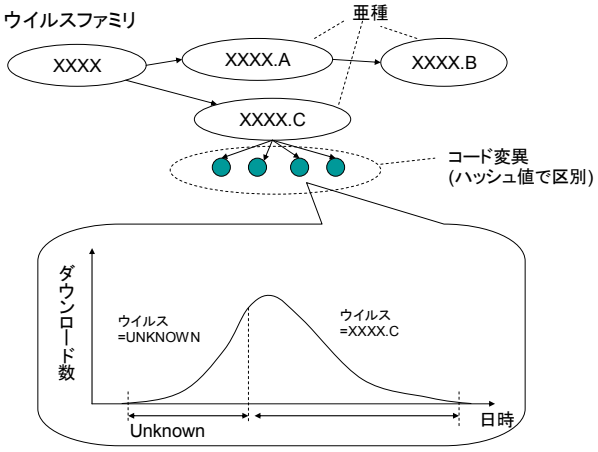

図 1　ウイルスのライフサイクル

今回特に着目したのは、ウイルス名称の項目である。ウイルスの発生当初はこの値は”UNKNOWN”であるが、ウイルス対策ベンダーにより命名され検出パターンが作られると XXXX.YY のような名称になる。ここで、XXXX はウイルスのファミリ名で、YY は「亜種」であることを表す。なお、一度名称がつくと他の名称に変更されることはなかった。

注意したいのは、同一のウイルス名であってもハッシュ値つまりコードとしては異なるものがあることである。このように、同一ウイルス名称でもコードが違うものを「コード変異」と呼ぶことにする。

また、各コードに対して、最初の出現から最初にウイルス名称がつくまでの期間を「UNKNOWN 期間」、その後最後の出現までの期間を「命名後期間」と呼ぶ(図 1)。

3.2. CCC2008 攻撃元データの基本統計

今回の分析の中心となる項目について基礎データとなる分析情報を以下に記載する。

(1) ダウンロードホスト IP アドレス
- ユニークアドレス　：258,711 個
- 最多出現　：826,962 件
- 出現 2 回　：72,108 個
- 出現 1 回　：51,788 個

(2) ハッシュ値（ウイルスの異なりコード）
- ユニーク数　：52,465 個
- 最多出現　：101,363 件
- 出現 2 回　：22,051 個
- 出現 1 回　：17,563 個
- 途中で UNKNOWN から、ウイルス名称に変更があったもの　：805 個（ユニーク数の 1.5%）

(3) ウイルス名称
- ユニーク数　：1,081 個（UNKNOWN 含まず）
- １ウイルス名称当りのハッシュ値数（平均のコード変異数）：約 47 個
- 途中で UNKNOWN 状態があったウイルス　：542 個（ユニーク数の約半分）

極めて多くダウンロードされるホストがある一方、大多数のホストは 1,2 回しかダウンロードされていない。これはアクセス数やページの被リンク数などインターネットでは良く見られる傾向である。

また、途中で UNKNOWN からウイルス名称を付与されたコードは 1.5%と少ないが、ウイルス名称で見ると、大半は途中でウイルス名称が付与されている。

3.3. コード毎のウイルス名称の変更

CCC2008 攻撃元データのうち、途中でウイルス名称が付与された 805 個のハッシュ値を持つコードに注目して次の項目を集計した。

(最初に出現した日時、最初に名称付与された日時、最後に出現した日時、UNKNOWN 期間、命名後期間、各期間における出現回数

図 2は、あるウイルスのダウンロード数を

時系列で見たものである。中ほどで命名されており、その後緩やかに減少している[1]。

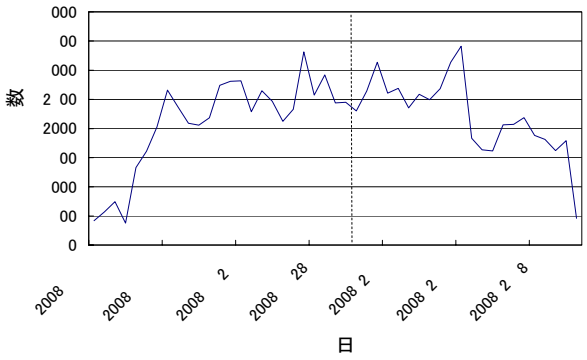

図 2 ダウンロード数の時系列変化

3.4. ウイルス名称付与前の特徴

図 3は UNKNOWN 期間の長さとウイルス出現回数の関係を表したグラフである。

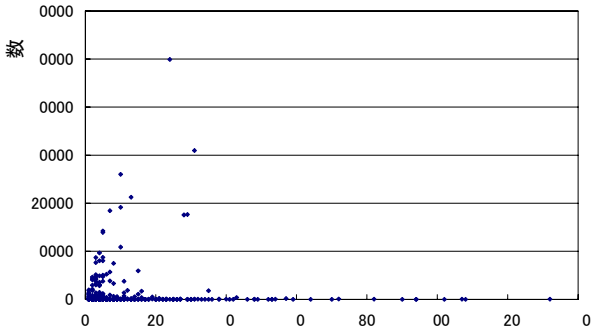

図 3 UNKNOWN 期間と出現数の関係

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。

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あることが読み取れる。これらの特徴として、ダウンロードホスト IP アドレスが特定の IP アドレスに偏っている[2]。特に上位 6IP アドレスのログはログ全体の 1/3 を占めており、関連するコードを含めるとログ全体の 1/2 も占めており、これら IP アドレスの全体に対する影響力は大きいと言える。

そこで中位以下のホストからのウイルスの特徴を見るために、上位 6 個の IP アドレス分および、関係するコードも除いたのが図 4 である。ここでは、特異的なウイルスは少なく、中位以降のダウンロードホストに関するマルウェアにおいては、UNKNOWN 期間とその出現回数はある程度の範囲内(99%は図 4 における双曲線より下)に収まっている。

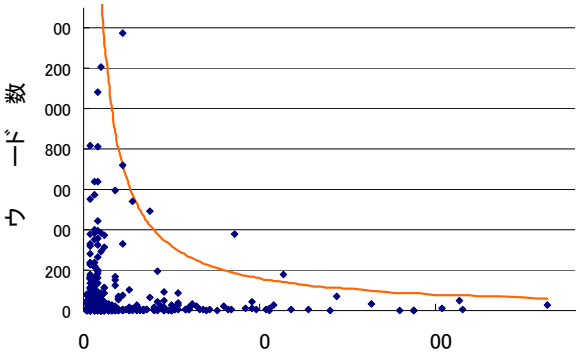

図 4 上位 6 アドレスの関連を除いた関係

3.5. ウイルス名称付与前後の変化

次にウイルス名称の付与前後の変化を見るために、出現回数をウイルス名称の付与前後に分けたグラフが図 5 である。

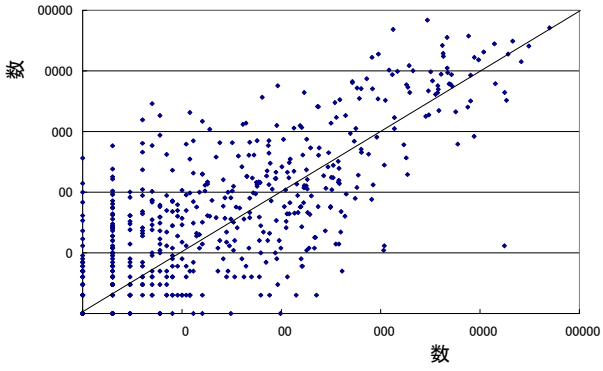

図 5 UNKNOWN 期間と命名後期間の出現回数

数減少が見られる。

[1] 命名日時は CCC2008 攻撃元データに含まれるもののみである。実際にはウイルス対策ベンダーによって命名および対策のタイミングは異なるため、本図のようにウイルスの活動状況は複雑な動きになる。

[2] 図 3 でダウンロード件数が極端に多いのが 1 点あり、それはある国のあるホストである。なお、その国では 08 年 2 月に大規模なボットネット犯罪グループが逮捕されている[4]。その影響かは不明だが、2 月後半からその国のいくつかのホストからのダウンロード

図 5 からウイルス名称の付与前後での出現回数は正の相関がある。つまり UNKNOWN 期間が長いものほど命名後期間の出現回数も多く、すなわちウイルス対策ベンダーの対応が遅いものは、パターンが出来てからのウイルスの活動が活発という仮説を立てることができる。

3.6. UNKNOWN 期間と地域との関係

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見てみよう。可視化ツールとして、富士通研究所で開発した関係メタデータの分析システムであるビジネス情報ナビゲーター[5]を使用した。ビジネス情報ナビゲーターは、RDF(Resource Description Framework)[6]形式のグラフメタデータの視覚化が可能で、人脈の検索や顧客関係の分析に用いられている。今回のように、ウイルス名称と地域の関係などグラフ形式で表現される情報には適したツールである。

UNKNOWN 期間のウイルスファミリとダウンロード元の地域（北米、欧州、アジア）を可視化したのが図 6である。

左下端に北米、中央上端に欧州、右下端にアジアと場所を固定すると、関連度の高いウイルス(名称)が自動的に近くに配置される。すると、3 地域からダウンロードされるウイルスは中央に配置される。また、北米とアジアからのみダウンロードされるウイルスは中央下端、北米と欧州からのみダウンロードされるウイルスは左上にくる。今回、右上が空ということは欧州とアジアからのみダウンロードされるウイルスがないことがわかる。また、北米やアジアは固有のウイルスが多いのに対して、欧州は少ない。北米発のウイルスがアジア、欧州に伝染するというケースが多い可能性がある。

次にあるウイルスファミリに注目して UNKNOWN 期間の亜種と国を可視化したのが図 7 である。同様に、複数の国からダウンロードされるウィルスは真ん中へんに、各国固有のものは周辺に配置される。ノードの色が黄色いものはダウンロードが活発なものを表している。やはり、多くの国で活動しているウイルスは活発である。一方、日本固有のものは、ダウンロード数はさほど多くはないが、亜種の数は相当数あることがわかる。

4. 終わりに

本論文では、CCC2008 攻撃元データの分析において、ウイルスが発生・命名・終息までのライフサイクルに着目した分析を行った。ウイルス対策ベンダーの方針もあろうが、命名のタイミングについてはバラツキが多く、特に日本を含めて地域固有のウイルスについては、数が多いにもかかわらず、影響が限定的なこともあり対応が遅い傾向がある。

日本は世界的に見て比較的ウイルス感染率が低いとされている[8]ものの、今後のマルウェアの攻撃の多様性を考えていくと、こうしたローカルなウイルスに対しては、ウイルス対策ベンダーだけではなく、政府や国内 ISP における連携した対策が必要となってきているのではないだろうか。

文　献

[1] “情報セキュリティ白書 2008”, 独立行政法人情報処理推進機構,2008.6.

[2] 鬼頭, 仲小路, 寺田, 菊池, “インターネット上の不正ホスト分布に関する社会的レイヤからの考察”, 情報処理学会研究報告, 2007-CSEC-38, 2007.7

[3] ラック社, セキュリティ情報, http://www.lac.co.jp/info/attacks-now.html

[4] Botnet Burst in Canada, McAfee Avert Labs blog, 2008.2.21 (http://www.avertlabs.com/research/blog/index.php/2008/02/21/botnet-bust-in-canada/)

[5] 松井,津田,片山, ナレッジマネジメントツール:ビジネス情報ナビゲーター, FUJITSU, pp.325-330, Vol.57. No.3, 2006.5

[6] RDF (Resource Description Framework), http://www.w3.org/RDF/

[7] 警察庁セキュリティポータルサイト@police, http://www.cyberpolice.go.jp/detect/observation.html

[8] Microsoft Security Intelligence Report, July through December 2007,

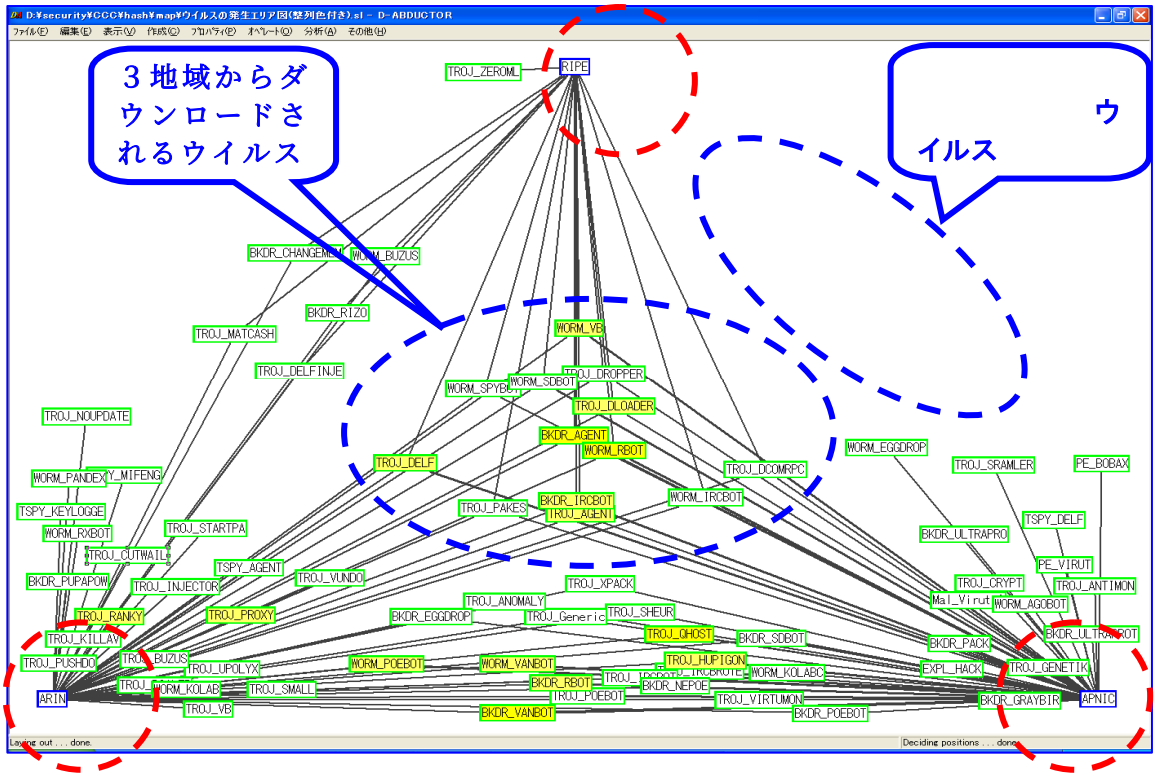

図 6 ウイルスファミリとダウンロード元（地域別）

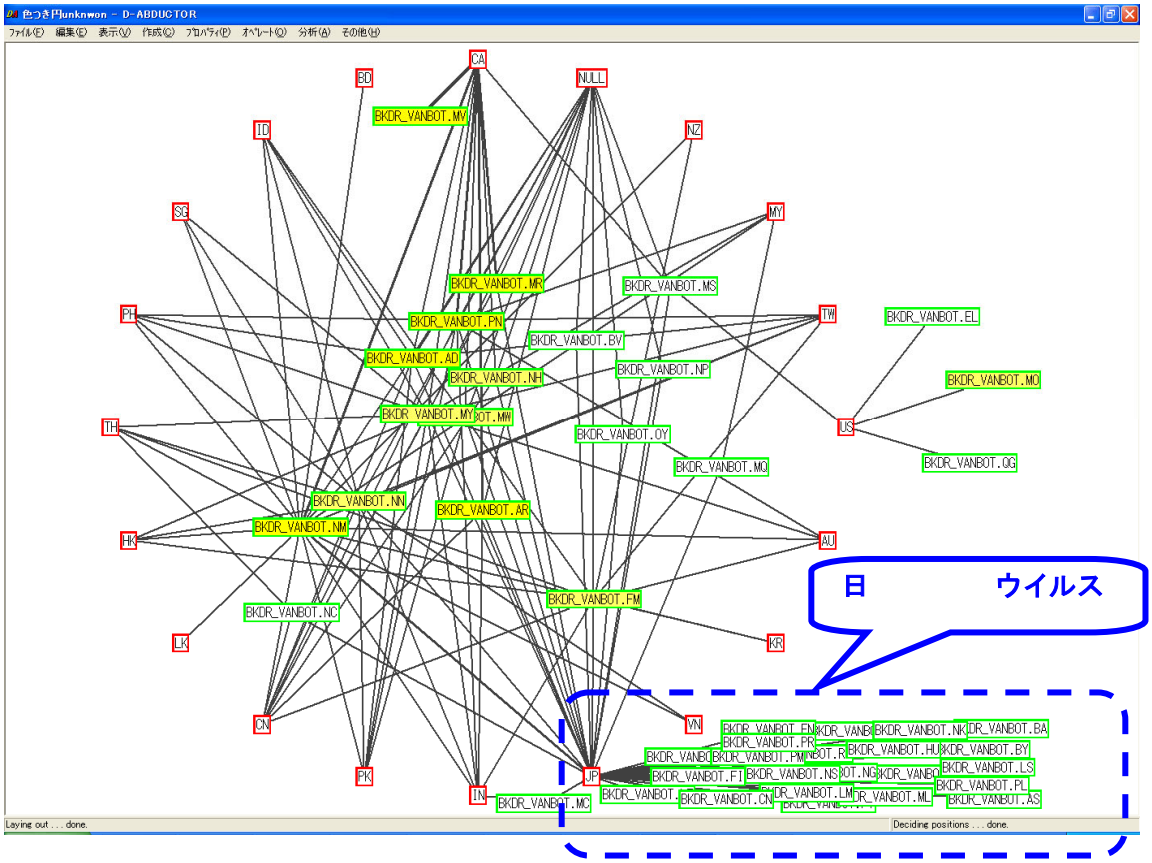

図 7 あるウイルスファミリの亜種とダウンロード元（国別）