Computer Security Symposium 2016
11 - 13 October 2016

大規模ライブネット観測データを用いたマルウェアの時系列解析に関して

村井 健祥[1]　森井 昌克[1]　池上 雅人[2]　長谷川 智久[2]　石川 堤一[2]

概要：銀行における不正送金や個人情報漏えい等，不正アクセスの被害は深刻の一途をたどっている．その不正アクセスを担っている方法の中核がマルウェアの利用である．マルウェアは日々進化しているとはいえ，新しい検体が流行の中心とは言えず，数年前に作成，さらに流行したマルウェアが再度利用されたりその亜種が流行することも多い．また，マルウェア同士も亜種だけにとどまらず，部分的に同じソースコードを利用することも多く，発生過程や発生状況も似通ることも少なくない．本稿ではマルウェアの時系列相関に関して考察を与え，亜種のマルウェアだけでなく異種のマルウェアでもその発生に関して相関があることを示す．実際に日本国内で発生したマルウェアの時系列分析を行い，互いに相関のあるマルウェアを与える．さらに世界各国のマルウェアの時系列分析を行い，その発生過程において似通った国々をグループに分割する．これらの結果はマルウェアの発生予測を行う上で重要な指標を与えるものである．

キーワード：マルウェア，時系列解析，相関解析

Time Series Analysis of Malware using Large-Scale Live Network Data

Kensho Murai[1]　Masakatu Morii[1]　Masato Ikegami[2]　Tomohisa Hasegawa[2]　Teiichi Ishikawa[2]

Abstract: Illegal remittance and personal information leakage　in the banking, etc. Damage of unauthorized access has gotten severely steadily. Malware is evolving day-to-day, but new malware can not be said to the center of the epidemic. Created a few years ago, prevalent malware is available again or its variants are often prevalent. Because it is also often utilize partially the same source code, generating process and generating situation may be similar. This paper provides a discussion about the time series correlation of malware is also shows that there is a correlation to occur in malware heterogeneous not only the malware of subspecies. These results are to provide a key indicator in performing the malware of occurrence prediction.

Keywords: Malware，Time Series Analysis，Correlation Analysis

1. はじめに

コンピュータやネットワーク技術の発展に伴いマルウェアによる被害の激増が問題となっている近年では，その対処が非常に重要な課題となっている．今や世界中で 1 日に数万から数十万のマルウェアが発生していると言われており，将来的には更なる発生の増加が危惧される．さらにマルウェアの持つ機能は多種多様であり，個人情報の流出からシステムの破壊に至るまで様々である．

マルウェアには亜種と呼ばれる機能の一部を少し変更した検体がマルウェア毎に数多く存在する．亜種の関係にあるマルウェアは基本的な機能が類似している場合が多く，それらの発生傾向には有る程度の相関が見られる．また，新種のマルウェアに関しても種類毎に機能や構造が全く異なるというわけでなく，過去に発生したマルウェアのソースコードを利用する例もある．したがって，異種の関係に

[1] 神戸大学
Kobe University
[2] キヤノン IT ソリューションズ株式会社
Canon IT Solutions Inc.

あるマルウェアであっても機能や発生傾向の類似は十分に考えられる．一つのマルウェアから連鎖的に複数のマルウェアが発生する場合を考える．最初に発生したマルウェアと相関の高いマルウェアが判明していれば，その後の発生の予測は容易となり，被害の防止に役立つ．

そこで本稿では，特定の 2 つの検体の発生数の推移が時期をずらして強く類似している場合，両検体には相関があると考える．そしてこの考えに基づき，大規模なライブネット観測データ（マルウェアの発生時系列データ）を利用したマルウェアの発生時系列の相関解析を行った．相関解析には相互相関関数を用いることで，定量的な類似度の評価によって膨大な検体の組み合わせから有意な検体の関係を効率良く絞り込む．この解析によって，相関のあるマルウェアの組を 4 組発見し，これらの組はいずれも異種検体の関係であった．さらに一国内における検体の相関解析から規模を広げ，世界各国を対象とした相関解析も行った．その結果，発生傾向の類似する国がいくつかのグループに分かれる事例を得た．これは，マルウェアの発生は必ずしも世界各国で同時かつ同様に起こる訳ではなく，発生に地域性を持つことを示唆している．これらの結果はマルウェアの発生予測を行う上で重要な指標を与えるものである．

2. 大規模ライブネット観測データ

”Eset Live Grid”はスロバキアの ESET 社並びにキャノン IT ソリューションズ株式会社によって提供されている，マルウェアの検知収集及び早期警告システムである．このシステムは ESET 製品を通して世界各国に展開され，マルウェアの流行を調査するセンサの役割も果たしている大規模なライブネット観測データである．本稿では ESET Live Grid よって収集されたマルウェア発生数データを事例に解析を行った．解析対象期間は 2015/01/01 から 2016/07/20 までの 567 日間である．マルウェア発生数データに収録されている情報は，日付（年月日），国名，発生した検体，発生数の 4 項目である．**表 1** に上記の解析対象期間における，世界全体 (147 ヶ国) 及び日本国内のセンサによる発生数及び発生した検体の種類数を示す．世界全体はもとより，日本国内だけで見ても発生数が平均で 1 日 21000 と非常に多くのマルウェアが発生検知されていることがわかる．さらに検体は約 1 万種類観測されているが，世界での発生検体種類数と比較すると 3000 ものマルウェアが日本では一切発生していないとも捉えられる．

表 1 統計情報

解析対象期間	2015/01/01-2016/07/20
検体発生数（世界）	1177934686
検体種類数（世界）	13894
検体発生数（日本国内）	12308700
検体種類数（日本国内）	10135

3. 日本における検体間の相関解析

検体間の相関解析において先行研究が柏井ら [1] によって行われている．柏井らは情報通信研究機構の NONSTOP と呼ばれるリモート分析環境から得られるデータを用いて，マルウェアの発生過程を 0，1 の 2 値に置き換えたものに相互相関関数を用いることで相関の評価を行っている．結果としていくつかの検体間において，発生の相関を発見した．したがって本稿では，柏井らの解析手法に改良を施した上で ESET のマルウェア発生数データに対して相互相関関数を用いた相関の評価を行った．本章では日本国内で発生したマルウェアに対して行なった解析の各段階について説明した後，結果と考察を述べる．

3.1 閾値による発生数の平滑化

後に用いる相互相関関数は本来，定常過程を対象とした関数であるため，マルウェアの発生のような非定常過程に適用した場合は誤差が大きくなることが予想される．よってその誤差を抑えるため，各検体の発生過程を 0，1 の 2 値で表現する．**図 1** に平滑化の例を示す．1 日毎の発生数に対し，設定した閾値を上回っていれば 1，そうでなければ 0 で置き換える．事前に複数の検体の発生過程をグラフ化して傾向を大まかに確認したところ，発生がインパルス応答に近い検体や，曜日の影響で 1 週間の周期で増減を繰り返すものが散見された．前者の特徴を持つの検体間においては相関係数が高い結果となっていても，偶然同時期に発生したとも捉えることができる．また，後者については閾値の判定におけるノイズとなる可能性がある．そこで本稿では閾値による置き換えの後，1 → 0 となる時点と 0 → 1 となる時点の日付の差が 7 日以上の場合は別の発生の山であるとみなす．逆に 6 日以内の場合は同一の山であるとみなし，その期間は新たに 1 で置き換える．山が 2 回以上観測された検体を，相互相関関数の入力とする．図 1 の例において，山は 2 つとカウントされる．

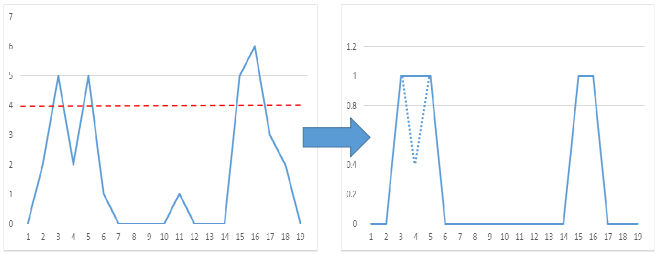

図 1 発生過程の平滑化

3.2 相互相関関数の適用

相互相関関数は対象とする 2 つの異なる系列データを

それぞれ入力として，データ間の類似度である相関係数を出力する．相関係数は-1 から 1 までの値をとり，絶対値の大きさによって相関を評価することができる．また，位相差を指定することにより，系列データをずらした場合の類似度も求めることができる．系列データを x_i，y_i $\{i = 1, 2, \cdots n\}$，x_i，y_i の相加平均をそれぞれ $\bar{x}$，$\bar{y}$，位相差を k とすると，相互相関係数 R_{xy} は以下の式で定義される．

$$R_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_{i-k} - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_{i-k} - \bar{y})^2}} \qquad (1)$$

(1) 式に対し，2 値で置き換えた発生過程の組を入力し検体間の相関係数を求める．発生数データの解析対象期間は 567 日間であるので，$n = 567$ となる．位相差が $k = \pm 5$ 以内で相関係数が 0.7 以上となる検体組み合わせを抽出した．一般には相関係数が 0.4 以上のとき相関があるとされているが，ここでも前節の誤差を考慮して 0.7 と設定した．

柏井らの解析手法では，単一の閾値において平滑化した各検体について組をとっている．しかし発生数に中程度の差がある検体間を考えた場合，平滑化した際に大きく様相が異なってしまい本来あるはずの相関が上手く現れないということが考えられる．そこで複数の閾値を設定し，各閾値によって平滑化した各検体の発生過程全ての集合から組をとり，相互相関関数を適用する．これにより，解析の網羅性の向上を図る．組の中には異なる閾値により平滑化された同一検体の組み合わせが存在することになるが，そういった事例は処理の過程で除外している．日本における発生数の元データを調査したところ，1 日の平均発生数の中央値の約半分にあたる 7000 前後の閾値において影響を受ける検体はごく僅かであったため，閾値の最大は 7000 とした．また，全体的な発生数から検知規模を推察し，20 以下の発生は検知誤差とみなした．以上より，閾値を 20 から 7000 まで 20 刻みで変化させ，各検体の発生過程を平滑化した．

最後に，相互相関関数により抽出された検体の各組に対して，発生数のオーダーについて判定を行う．これは前段階において複数の閾値で発生過程を 2 値化しているため，発生の規模が考慮されていない状態となっているからである．したがって，両検体の 1 日の最大発生数が 5 倍以上開いている組は相関が低いと判断し，除外する．

3.3 解析結果と考察

以上の解析手法の結果，1827 通りの検体の組が抽出された．発生数グラフの目視による類似度の確認が可能な範囲に収まっているため，これらの検体の組から特に高い相関を示した 4 つの組について，考察を行う．**表 2** 及び**図 2**，**図 3**，**図 4**，**図 5** に検体の組と発生数のグラフを示す．縦軸は発生数，横軸は 2015/01/01 からの経過日数を表す．

図 2 に関して，JS/Toolbar.Crossrider はブラウザ上で過去のアクティビティを考慮した広告やポップアップを表示するツールバーの一種である．また JS/Adware.Spigot も同様に害のある広告をブラウザ上に表示する機能を持っており，どちらも JavaScript で書かれたプログラムである．グラフ中の 197 日目で JS/Adware.Spigot が急激に増えており，以降は JS/Toolbar.Crossrider の発生過程に付随する傾向を示している．図 3 に関して，Win32/TrojanDropper.Addrop はトロイの木馬に分類され，対象マシンに様々なアドウェアや不適切な動作をする可能性のあるアプリケーションを呼び込む機能を持つ．一方で Win32/Adware.Imali は名前の通りアドウェアの一種であることから，ダウンロードが行なわれても不自然ではない．図 4 に関して，Win32/Bayrob メールの添付ファイルとして配布されるトロイの木馬で主

表 2 相関の高い検体ペアと発生ずれ

検体ペア		発生のずれ (日)
JS/Toolbar.Crossrider	JS/Adware.Spigot	1
Win32/TrojanDropper.Addrop	Win32/Adware.Imali	5
Win32/Bayrob	Win32/TrojanDownloader.Zurgop	5
Win32/ExtenBro	Win32/Zlader	5

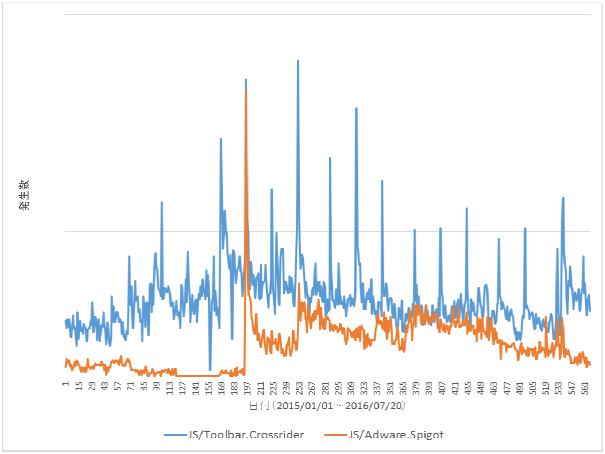

図 2 JS/Toolbar.Crossrider - JS/Adware.Spigot

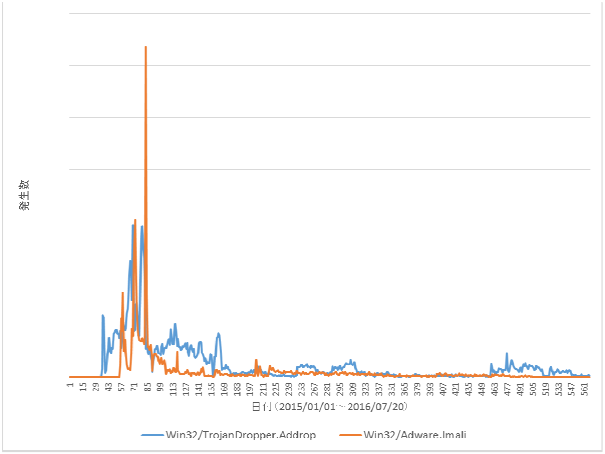

図 3 Win32/TrojanDropper.Addrop - Win32/Adware.Imali

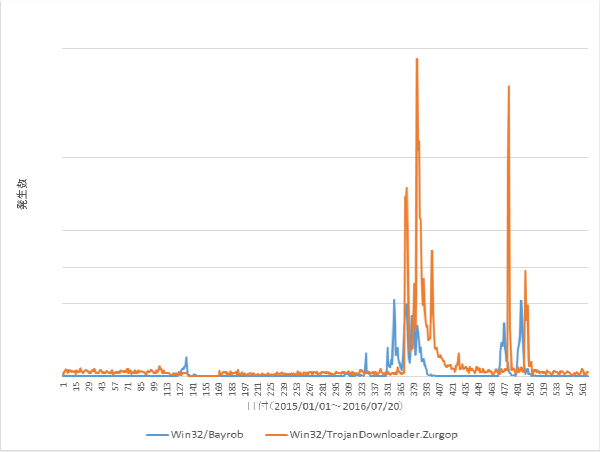

図 4　Win32/Bayrob - Win32/TrojanDownloader.Zurgop

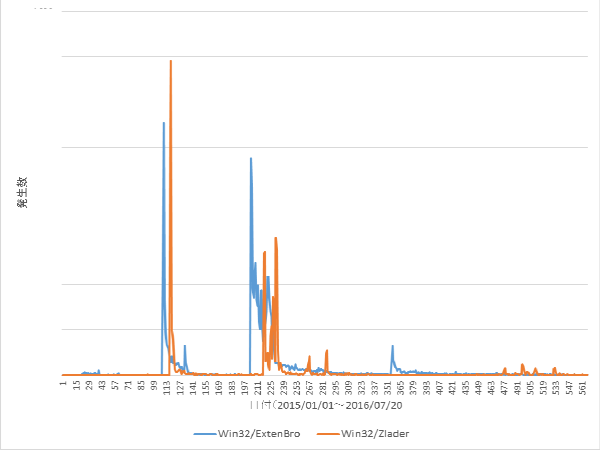

図 5　Win32/ExtenBro - Win32/Zlader

にバックドアとしてリモートサーバーに情報を送信するほか，マルウェアのダウンローダ機能も持つ．発生過程を見ても，異なる時期において同様の増加傾向を示しており，Win32/TrojanDownloader.Zurgop が連鎖的に発生する可能性は高いと言える．図 5 に関して，Win32/ExtenBro はトロイの木馬であり，特徴として感染経路がメール，P2P ネットワーク，WEb サイト等様々である点が挙げられる．Win32/Zlader は主に個人情報を盗み出すトロイの木馬である．こちらの組についても異なる時期において同様の発生傾向を示していることが分かる．日本国内における相関解析の結果から以上の 4 つの組は発生に相関を持つと考えられる．したがって今後は表 2 中左側の検体が発生した場合，数日後にそれと対応する検体が発生すると予測され，注意喚起などにより被害を抑えることができる．

4.　複数国での相関から見る発生の地域性

4.1　日本での発生傾向と他国の比較

前章では日本国内における検体間の相関解析により，4 組の検体で相関関係を発見した．そこで解析の規模を拡大し，日本において相関のある検体が他国においてどのような傾向を有しているか調査を行なった．

解析期間内に 147 ヶ国全てにおいて JS/Toolbar.Crossrider と JS/Adware.Spigot が共に発生していた．JS/Toolbar.Crossrider に関してはどの国においても同様の発生傾向であった．しかし JS/Adware.Spigot に関してはグラフ中の 197 日目（2015/07/17）でのスパイクが JS/Toolbar.Crossrider の発生数と比較して強く現れている国と，中程度現れている国，全く現れていないもしくは検知誤差と見なせる程度に少ない国の 3 通り存在した．日本と異なる発生傾向を，それぞれ 1 国を例に図 **4.1** に示す．強く現れている国のは日

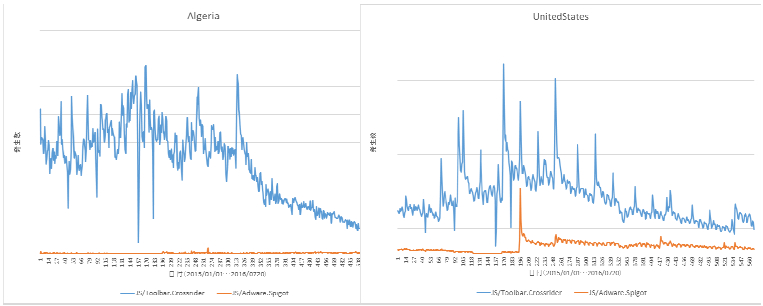

図 6　左：スパイク無し，右：中程度のスパイク

本，香港，中国，シンガポール，台湾，オーストラリアであり，中程度現れている国はヨーロッパ諸国と北米であった．スパイクの現れた国に関しては，発生のずれはいずれも 1 日前後であったことから，両検体には強い相関があり，発生の予測が可能であると言える．スパイクが強く現れた国に共通する事柄に時差の一致が挙げられる．また，日本時間で 2015/07/17 は ipodtouch 第六世代の発売日と一致していた．したがって購入者の多くがインターネット利用したことで他の諸国に比べて多く発生したと推察され，その後 JS/Toolbar.Crossrider の広告サジェスト機能により JS/Adware.Spigot の検知が続いたとも捉えられる．

Win32/Bayrob と Win32/TrojanDownloader.Zurgop は 145 ヶ国にて両検体の発生が確認されたが，発生過程に相関関係が見られた国は日本のみであった．Win32/Bayrob を添付したメールにはいくつかの言語が存在し，特定の国を狙って攻撃している可能性があるとされている．攻撃者はなんらかの理由で Win32/TrojanDownloader.Zurgop の感染を試みていると推察され，Win32/Bayrob と Win32/TrojanDownloader.Zurgop は日本特有の相関であると考えられる．

Win32/ExtenBro と Win32/Zlader は 145 ヶ国にて両検体の発生が確認された．その内，両検体が日本と同様の発生傾向である国は，オーストラリア，香港，ニュージーランド，中国，シンガポールであり，こちらも時差の一致が共通の特徴として挙げられる．ヨーロッパ諸国と北米に関してはグラフ中の 113 日目付近と 225 日目での傾向は日本

と同様であるが，447 日目付近において再度 Win32/Zlader が発生していた．ロシアに関しては他国と大きく異なる発生傾向を有しており，Win32/ExtenBro の発生数と比較して Win32/Zlader が圧倒的に少ない発生数であった．ヨーロッパ諸国と北米に共通する特徴及びロシアでの発生傾向を図 **4.1** に示す．

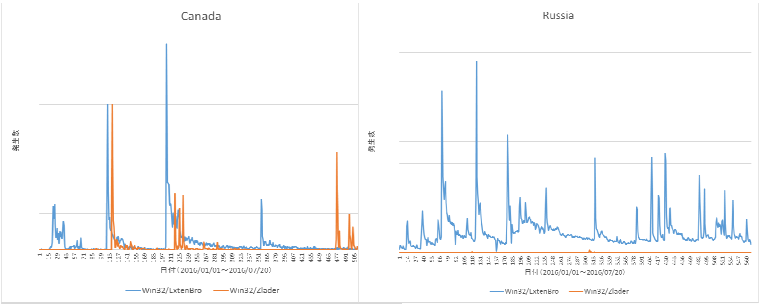

図 7　左：ヨーロッパ，北米，右：ロシア

以上より，日本で得られた検体と同様の相関が見られる国もあれば異なる相関を示す国のグループも存在し，マルウェアによる攻撃が必ずしも一様ではないと分かる．JS/Toolbar.Crossrider と JS/Adware.Spigot の事例においては，単純な大陸の区分ではなく，時差が要因となって発生の地域性が見られたことから，グループを為す要因は複数存在することが分かる．

5.　まとめ

本稿では，近年のマルウェア流行の現状から亜種及び異種マルウェア間の相関を仮定し，大規模ライブネット観測データを用いた発生時系列データの相関解析と考察を行った．解析によって，日本国内において異種関係であっても発生過程の類似するマルウェア検体の組を発見した．また，得られた相関関係を世界各国と比較した結果，検体によっては発生傾向にグループが存在することが分かった．これらの結果はマルウェアの発生予測を行う上で有意なデータであるといえる．

日本国内の相関を基準に解析を行ったが，世界各国それぞれで同様の解析を行うことで新たな検体の組が発見できる可能性が高い．今後は他国を基準とした相関解析の実施と，定量的なクラスタリング主法の適用を行う．

参考文献

[1]　柏井祐樹, 森井昌克, 井上大輔, 中尾康二” NONSTOP データを用いたマルウェアの時系列分析”　コンピュータセキュリティシンポジウム (CSS2013), 2013 年 10 月.