NONSTOP データを用いたマルウェアの時系列分析

柏井 祐樹 †　森井 昌克 †　井上 大介 ‡　中尾 康二 ‡

† 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
y.kashii@stu.kobe-u.ac.jp
mmorii@kobe-u.ac.jp

‡ 独立行政法人情報通信研究機構
184-8795 東京都小金井市貫井北町 4-2-1
{dai},{ko-nakao}@nict.go.jp

あらまし マルウェアには特徴的な発生傾向として，マルウェア同士で発生過程が類似していることが挙げられる．そのようなマルウェアは時間的な相関が存在すると仮定できる．特徴的な相関や傾向を有するマルウェアを発見できれば，今後のマルウェア対策に非常に有効となる．本稿ではマルウェア発生傾向の把握を目的として発生過程に対して時系列分析を行う．マルウェアの発生過程を統計的に分析することでマルウェア同士の相関を評価し，マルウェア同士の時間的な関連性を発見することが可能となる．

Time series analysis of malware using NONSTOP data

Yuki Kashii†　Masakatu Morii†　Daisuke Inoue ‡　Kouji Nakao ‡

†Graduate School of Engineering, Kobe University
1-1 Rokkodai-cho Nada-ward,Kobe 657-8501,JAPAN
y.kashii@stu.kobe-u.ac.jp
mmorii@kobe-u.ac.jp

‡National Institute of Infomation and Communications Tecnology
4-2-1 Nukuikitamashi, Koganei-city, Tokyo 184-8795, JAPAN
{dai},{ko-nakao}@nict.go.jp

Abstract

Malware occurs in distinctive trend,for example developmental processes of malware are similar to others. We can presume that temporal correlation exists such malwares. If we can find the malware that has characteristic tendency or correlation, it is very effective anti-malware in the future. I do the time-series analysis of developmental processes in order to understand the occurrence tendency of malware in this paper. I evaluate the correlation malwares by analyzing statistically developmental processes. As a result, it is possible to find the relevant temporal malwares.

1　はじめに

マルウェアには特徴的な発生傾向として，マルウェア同士で発生過程が類似していることが挙げられる．例えばダウンローダ型と呼ばれるマルウェアは他のマルウェアをダウンロードし，コンピュータに新たなマルウェアを感染させるマルウェアである．ダウンローダ型のマルウェアに感染すると連鎖的に別のマルウェアに感染する．その際，ダウンローダ型のマルウェアと

新たに感染したマルウェアに対して相関が生じる．また，ダウンローダ型以外のマルウェア同士においても発生過程に相関や特徴的な傾向がみられる可能性がある．特徴的な相関や傾向を有するマルウェアを発見できれば，今後のマルウェア対策に非常に有効となる．マルウェアの発生過程を把握する際に最も有効な手段として発生過程のグラフ化が挙げられる．しかし，グラフ化は情報量が少ない場合には有効な手段であるが，情報量が増加すると人間の視覚認識能力を超える問題がある．さらに，検体名は亜種も含めると数百種存在し，これを全てグラフ化して発生傾向を把握することは困難である．

本稿ではマルウェア発生傾向の把握を目的として発生過程に対して時系列分析を行う．マルウェアのデータには独立行政法人情報通信研究機構 [1] のインシデント対策センタ (nicter) が開発したリモート分析環境 Nicter Open Network Security Test-Out Platform(NONSTOP) のサーバ内から得られる解析レポートを用いる．解析レポートの中には各セキュリティソフトベンダの検体名が記載されているテキストファイルが存在する．テキストファイルに記載されている検体名をデータベース化してファイルの作成日時順で管理することにより，解析の効率化を図る．作成したデータベースから検体名ごとにマルウェアを分類し，検体名ごとの発生過程から時系列分析を行う．時系列分析によりマルウェア同士の相関を導出することで，今後のマルウェアの発生傾向を予測することが可能である．

2 NONSTOP

近年では大半のマルウェアが難読化やアンチデバッキングといった技術を利用して解析を困難にしているため，研究者が独自に安全な解析を行うことは困難である．そこで，nicter はコンピュータ上でマルウェアの可能性のあるプログラムを正確かつ高速に判別する解析機能を開発し，その解析結果を外部の共同研究者に提供している．その際に仲介するシステムが NONSTOP と呼ばれるオープンプラットフォームである．NONSTOP によって，マルウェア検体やトラフィックデータなどのネットワークセキュリティの研究に不可欠となる膨大なデータ群が安全に利用可能となる．現在利用が可能な NONSTOP 内のデータは 3 種類存在し，マクロ解析システムから得られたデータ，ミクロ解析システムから得られたデータとマクロ-ミクロ相関分析システムから得られたデータとなっている．マクロ解析システムとはネットワークを観測し，ネットワーク攻撃をリアルタイム自動分析して得られた結果をデータベースに蓄積するシステムである．ミクロ解析システムとは外部システムから飛来したマルウェア等を隔離環境の中で動作させ，解析結果をデータベースに蓄積するシステムである．マクロ-ミクロ相関分析システムとはマクロ解析システムによって検知された新たな攻撃やインシデントの予兆と，ミクロ解析システムで解析されたマルウェアの相関を調べた結果を蓄積するシステムである．

本稿では NONSTOP サーバ内からマルウェアの解析レポートを収集し，利用することでマルウェア同士の時系列分析を実現する．

3 時系列分析と相互相関関数

本章ではマルウェアの発生過程の時系列分析の際に用いる手段について説明する．マルウェアの発生過程を統計的に分析することでマルウェア同士の相関を評価することが可能となる．

3.1 多変量時系列データの相互相関関数

時系列分析を行うために多変量時系列データに対して相互相関関数 [2] を用いる．相互相関関数とは一般に 2 つ以上の時系列データの類似度を量的に表す尺度であり，値が大きいほど相関が強いといえる．多変量時系列データ $\{x_t | t \in \mathcal{Z}\}$ はある確率過程に従う確率変数の実現値とみなす．多時系列データが定常過程であり，そのラグが k(すべての整数) のとき相互相関関数 $\rho_k(i,j), i \neq j$ は

$$\rho_k(i,j) = \frac{E[(x_{i,t} - \mu_i)(x_{j,t-k} - \mu_j)]}{\sqrt{E[(x_{i,t} - \mu_i)^2 (x_{j,t-k} - \mu_j)^2]}} \quad (1)$$

表 1: 代表的な検体名の例

Adware & Downloader	Adware.* Downloader, etc..
Virus	W32.Sality.*, W32.Virut.*, etc..
Trojan	Backdoor.*, Trojan.*, etc..
Worm	W32.IRCBot.*, W32.rahack.*, etc..
Unknown	Unknown,Noname

で求めることができる．μ は多変量時系列データの平均であり

$$\mu = E[x_t] \tag{2}$$

で求めることができる．

前述したように相互相関関数は多変量時系列データが定常過程の場合に用いることができる．多変量時系列データが定常過程になる条件は以下の 2 つを満たす場合である．

1. 平均 μ は t に依存しない
2. すべての k に対して，相互共分散関数 $Cov_{i,j}$ は t に依存しない

$$Cov_{i,j} = E[(x_{i,t} - \mu_i)(x_{j,t-k} - \mu_j)] \tag{3}$$

マルウェアの発生過程は時間に依存して推移するため，マルウェアの発生過程を時系列データとして扱うことが可能である．本稿では 2 種のマルウェアに対する発生過程を 2 変量時系列データとして時系列分析を行う．

3.2 相互相関関数の発生過程への導入

検体名ごとの発生過程を取得するために，NONSTOP サーバ内に保存されているマルウェアの解析レポートを用いる．解析レポートにはマルウェアの分類を行う際に必要な検体名が記載されている．記載された検体名をもとにマルウェアの分類を行い，得られた発生過程に対して相互相関関数を用いることでマルウェア同士の相関の有無を評価する．NONSTOP サーバ内から取得できる解析レポートの中にはセキュリティソフトベンダである Symantec 社 [3]，McAfee 社 [4]，Trend Micro 社 [5] による検体名が記載されている．解析レポートはテキスト形式のデータとなっており，そのままマルウェアの分類等の解析に用いるには不向きである．そこで，マルウェアごとに記載されている検体名を各社ごとに解析レポートの作成日時順でデータベース化して管理することにより，解析の効率化を図る．次に，前述したデータベースを用いてマルウェアの発生過程を検体名ごとに取得する．本稿では Symantec 社の検体名をもとに行った．代表的な検体名を表 1 に示す．

3.1 節では時系列モデルに対して相互相関関数を示したが，実際はモデルではなくマルウェアの発生過程に対して相関を評価しなければならない．マルウェアの発生過程 $\{x_t | t = 1, 2, \cdots, n\}$ が定常時系列の実現値であると仮定すると標本相互相関関数 $\hat{\rho}_k(i,j)$ は相互相関関数の推定値を与える．標本相互相関関数は

$$\hat{\rho}_k(i,j) = \frac{\sum_{t=k+1}^{n}(x_{i,t} - \hat{\mu}_i)(x_{j,t-k} - \hat{\mu}_j)}{\sqrt{\sum_{t=1}^{n}(x_{i,t} - \hat{\mu}_i)^2}\sqrt{\sum_{t=1}^{n}(x_{j,t} - \hat{\mu}_j)^2}} \tag{4}$$

で求めることができる．ただし，$-n < k < n$ とする．$\hat{\mu}$ は多変量時系列データの標本平均であり

$$\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} x_t \tag{5}$$

で求めることができる．

マルウェアは一定の期間で集中的に発生する場合が多いため，発生過程は非定常時系列データの可能性が高い．標本相互相関関数は非定常時系列データに対しては相関値の誤差が大きくなるために用いることができない．そこで，マルウェアの発生過程を

$$x'_t = \begin{cases} 1 & (x_t \geq T) \\ 0 & (x_t < T) \end{cases} \tag{6}$$
$$T = \{5, 10, 20, 30, 40, 50\}$$

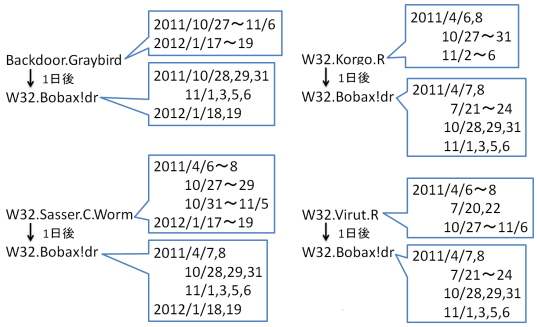

図 1: 閾値を 5 とした場合の相関

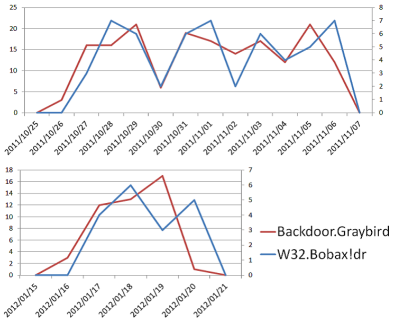

図 2: Backdoor.Graybird-W32.Bobax!dr

に平滑化し，新たなマルウェアの発生過程 x'_t を与える．T は平滑化する際の閾値を示す．

4 標本相互相関関数から得られた発生傾向

マルウェアの発生過程を取得する期間は 2010 年 7 月 1 日〜2012 年 6 月 30 日の 2 年間とする．総検体数は 1263227 であり，2011 年の 1 年間で 550 種のマルウェアを確認した．マルウェアの発生過程は 1 日あたりの検体数を 2010 年 7 月 1 日〜2012 年 6 月 30 日の 2 年間分取得する．よって，データ数は $n = 731$(閏年の影響) となっている．マルウェアの発生件数が少なければマルウェアごとの発生傾向を把握することが困難となる．そこで，2 年間で発生した総検体数が 50 以下であるマルウェアの発生過程は調査対象から除外した．また，発生過程が 1 回のみ T を越えていた場合も標本相互相関関数から発生傾

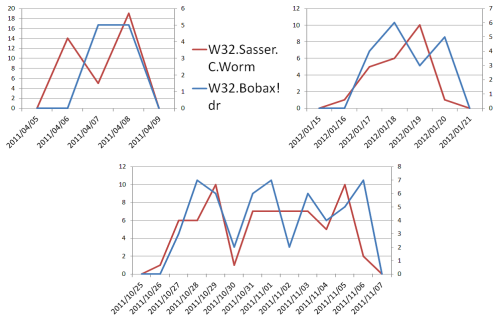

図 3: W32.Sasser.C.Worm-W32.Bobax!dr

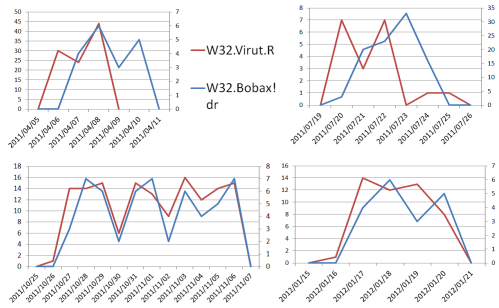

図 4: W32.Virut.R-W32.Bobax!dr

向を把握することは困難である．そこで，マルウェアの発生過程は T を越える場合が 2 回以下の場合も調査対象から除外する．

4.1 発生日にずれが生じたマルウェア

標本相互相関関数の値の最大値が 0.4 以上のとき相関が存在するとし，ラグ $k = \pm1, \pm2, \cdots, \pm5$ までの相関の有無を評価した．

閾値を 5 とした場合，1 日あたりの発生数が閾値以上を示した日付が 2 回以上一致する組み合わせは 4 通り存在した．結果と一致した日付を図 1 に示す．また，発生過程をグラフ化した結果を図 2〜4 に示す．図 1 から Backdoor.Graybird，W32.Korgo.R，W32.Sasser.C.Worm，W32.Virut.R が発生した 1 日後に W32.Bobax!dr が発生する可能性が高いことが分かる．このことから 4 種のいずれかが発生した後に W32.Bobax!dr の発生に注意する必要がある．また，図 1 から 4 種の発生する日が同時期になる可能性が高いことも読み取ることができる．図 2，4 から Backdoor.Graybird および W32.Virut,R は W32.Bobax!dr と発生過程にラグが生じていない状態でも発生過程が類

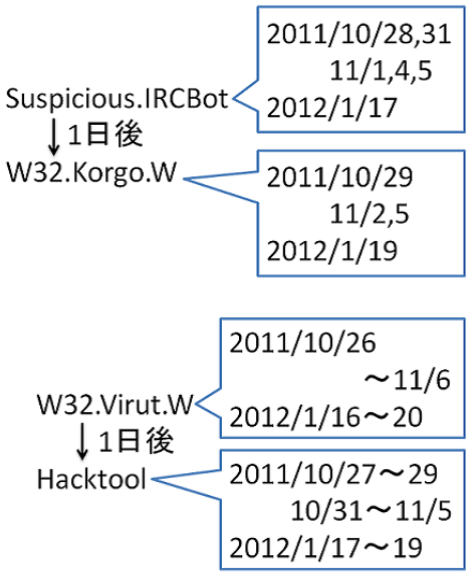

図 5: 閾値を 20 および 30 とした場合の相関

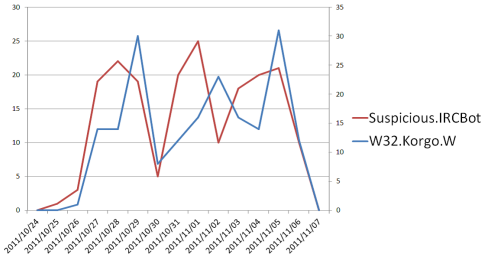

図 6: Suspicious.IRCBot-W32.Korgo.W

似している場合がある．

閾値を 20 とした場合，1 日あたりの発生数が閾値以上を示した日付が 2 回以上一致する組み合わせは 1 通り存在した．結果と一致した日付を図 5 の上側に示す．また，発生過程をグラフ化した結果を図 6 に示す．図 5，6 から Suspicious.IRCBot が発生もしくは増加した 1 日後に W32.Korgo.W が発生もしくは増加していることが明らかに分かる．このことから Suspicious.IRCBot が発生した後に W32.Korgo.W の発生を予測することは容易である．

閾値を 30 とした場合，1 日あたりの発生数が閾値以上を示した日付が 2 回以上一致する組み合わせは 1 通り存在した．結果と一致した日付を図 5 の下側に示す．また，発生過程をグラフ化した結果を図 7 に示す．図 7 から W32.Virut.W と Hacktool は発生過程にラグが生じないことが分かる．これは発生過程に閾値を設けた結果，

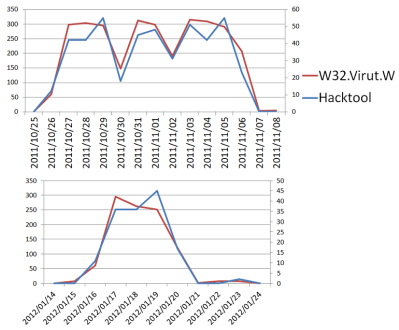

図 7: W32.Virut.W-Hacktool

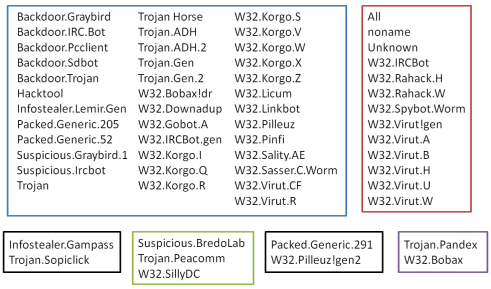

図 8: 閾値が 5 の場合

閾値以下の発生数が考慮されなかったことが原因であると考えられる．

4.2 同時期に発生したマルウェア

標本相互相関関数の値の最大値が 0.4 以上のとき相関が存在するとし，ラグ $k = 0$ の相関の有無を評価した．結果を図 8～10 に示す．

発生過程が類似しているマルウェア同士でグループ分けを行った結果，多数のマルウェアが同時期に発生していることが分かる．同時期に発生するマルウェアはそのほとんどが亜種関係にあることが分かる．閾値が増加していくとグループ内で分裂が起こっている．また，閾値を変えても相関が高いマルウェアの種類に大きな差異がないことが分かる．

5 まとめ

本稿ではマルウェアの発生過程に対して時系列分析を行い，マルウェア同士の相関の有

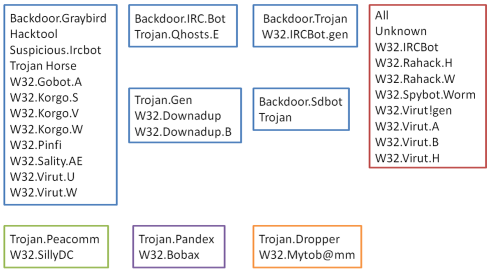

図 9: 閾値が 20 の場合

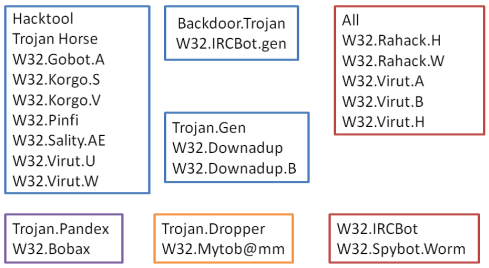

図 10: 閾値が 40 の場合

無を評価した．マルウェア同士の時系列分析には標本相互相関関数を用いた．標本相互相関関数は時系列データが定常過程の場合にマルウェア同士の相関を評価することが可能である．しかし，マルウェアの発生過程はほとんどが非定常過程である．そこで，マルウェアの発生過程に対して閾値を与えて平滑化を行った．相関が存在したマルウェアの組み合わせを 6 通り発見することができた．そのうち 5 通りは今後のマルウェア発生過程を予測する上で非常に有効であると考えられる．また，Backdoor.Graybird，W32.Korgo.R，W32.Sasser.C.Worm，W32.Virut.R および W32.Virut.W，Hacktool は発生過程にラグが生じていない．しかし，発生過程を区切る間隔をより狭くした場合に相関が存在する可能性がある．同時期に発生したマルウェアを調査した結果，数多くのマルウェアで類似していた．また，閾値を変えても相関が高いマルウェアの種類に大きな差異はなかった．

謝辞

有益な御討論を頂いた（独）情報通信研究機構衛藤将史氏ならびに伊沢亮一氏をはじめとする（独）情報通信研究機構ネットワークセキュリティ研究所サイバーセキュリティ研究室各位に感謝する．

参考文献

[1] 独立行政法人情報通信研究機構, http://www.nict.go.jp/

[2] Peter.J.Brockwell and Richard.A.Davis, Introduction to Time Series and Forecasting, Second edition, pp224〜pp257, Springer 2002.

[3] Symantec, http://www.symantec.com/ja/jp/index.jsp

[4] McAfee, http://home.mcafee.com/

[5] Trend Micro, http://jp.trendmicro.com/jp/home/