L-005

マルウェアの発生時系列分布特性

Disposition characteristic of the time series related to the outbreak of malware

柏井 祐樹 †　森井 昌克 †　井上 大介 ‡　中尾 康二 ‡
Yuki Kashii　Masakatu Morii　Daisuke Inoue　Kouji Nakao

1　はじめに

マルウェアには発生過程が特徴的であったり，マルウェア同士で発生過程に相関が存在する場合がある．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は発生過程のグラフ化である．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．さらに，その発生過程および分析結果について可視化を試みる．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，可視化や解析等の効率化を図る．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．さらに発生過程の解析支援システムとして，解析レポート作成日時の古い順から検体名ごとに立法体内で発生過程の可視化を行う．

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と可視化を実現する．

3　NONSTOP データを用いた時系列分析

本章ではマルウェアの発生過程の時系列分析によりマルウェア同士の相関の評価を行う．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することが可能となる．

3.1　2 変量時系列の相互相関関数

時系列分析を行うために 2 変量時系列における相互相関関数 [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．時刻 $(t = 1, 2, \cdots, n)$ における 2 変量時系列データを $(x_{1,t}, x_{2,t})$ とする．2 つの時系列データが定常過程であり，そのラグが k のとき相互相関関数 $\rho_k(1,2), \rho_k(2,1)$ はそれぞれ

$$\rho_k(1,2) = \frac{\sum_{t=k+1}^{n}(x_{1,t}-\mu_1)(x_{2,t-k}-\mu_2)}{\sqrt{\sum_{t=1}^{n}(x_{1,t}-\mu_1)^2}\sqrt{\sum_{t=1}^{n}(x_{2,t}-\mu_2)^2}} \quad (1)$$

$$\rho_k(2,1) = \frac{\sum_{t=k+1}^{n}(x_{2,t}-\mu_1)(x_{1,t-k}-\mu_1)}{\sqrt{\sum_{t=1}^{n}(x_{1,t}-\mu_1)^2}\sqrt{\sum_{t=1}^{n}(x_{2,t}-\mu_2)^2}} \quad (2)$$

で求めることができる．また，$\mu_i(i = 1, 2)$ は 2 変量時系列データの平均であり

† 神戸大学大学院工学研究科，Graduate School of Engineering,Kobe University
‡ 独立行政法人情報通信研究機構，Natioanal Institute of Infomation and Communications Technology

表 1　代表的な検体名の例

Adware & Downloader	adware.*,downloader, etc..
Virus	w32.sality.*,w32.virus.*, etc..
Trojan	trojan.*,packed.generic.*, etc..
Worm	w32.virut.*,w32.rahack.*, etc..
Unknown	Unknown,Noname

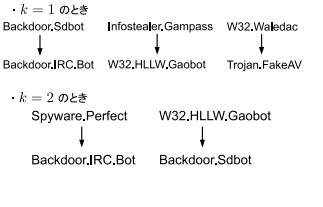

図 1　k=1,2 の場合の相関例

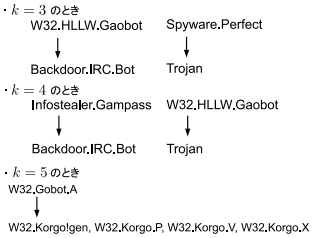

図 2　k=3,4,5 の場合の相関例

$$\mu_i = \frac{1}{n}\sum_{t=1}^{n} x_i \tag{3}$$

で求めることができる．

前述したように相互相関関数は 2 つの時系列データが定常過程の場合で用いることができる．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_i は t に依存しない
2. すべての k に対して，共分散関数 Cov_i は t に依存しない

$$Cov_i = \frac{1}{n}\sum_{t=k+1}^{n}(x_{i,t}-\mu_i)(x_{i,t-k}-\mu_i) \tag{4}$$

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相互相関分析を行う．

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，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，そのまま可視化や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，可視化や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．

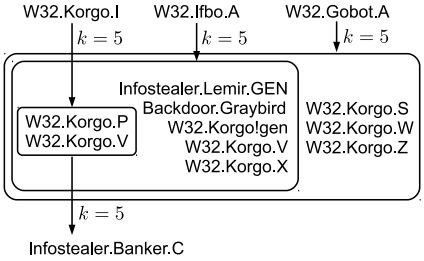

図 3　W32.Korgo 型を含む相関

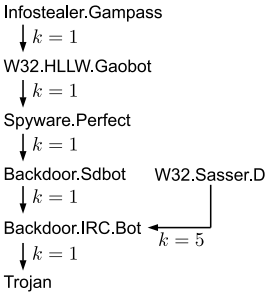

図 4　多階層型の相関

マルウェアの発生過程は非定常時系列データの可能性が高いため，取得した検体名ごとの発生過程に対して相互相関関数を用いることができる可能性は低い．そこで，マルウェアの発生過程に対して 1 次の差分をとることで定常時系列データに変換する．差分をとった時系列データ $y_{i,t}$ は

$$y_{i,t} = x_{i,t+1} - x_{i,t} \tag{5}$$

で求めることができる．定常過程となった時系列データ $y_{i,t}$ に対して，相互相関関数を用いることで検体名ごとの相関の有無を評価する．また，マルウェアの発生件数が少なかったり発生過程がインパルス応答型の場合は相互相関関数から正しい相関を得ることができない．そこで，評価を行う前に閾値を下回るマルウェアの発生過程は調査対象から除外する．

3.3　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．3.2 節で述べたように発生件数が少なければ正しい相関を得ることができない．そこで，2 年間で発生した総検体数が 20 以下のマルウェアは除外する．発生過程がインパルス応答型の場合も同様にして，発生過程をグラフ化して視認することで除外する．本稿では除外されなかったマルウェア 104 種に対して評価を行う．相互相関関数の値の最大値が 0.45 以上のとき相関が存在するとし，ラグ $k = 1, 2, \cdots, 5$ までの相関の有無を評価した．

マルウェアごとの相互相関関数の評価結果の例を図 1,2 に示す．図 1,2 は矢印上側のマルウェアの発生過程を k 日遅れで矢印下側のマルウェアの発生過程が追随す

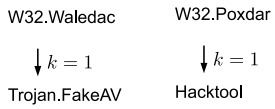

図 5　マルウェアの連鎖発生が短い相関

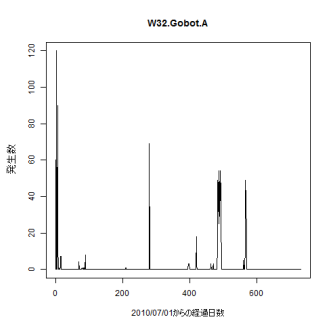

図 6　W32.Gobot.A の発生過程

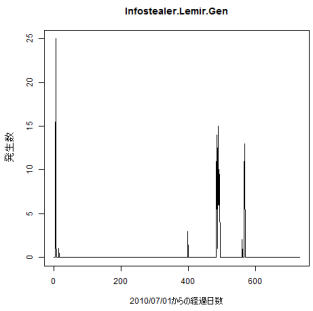

図 7　InfoStealer.Lemir.Gen の発生過程

る可能性が高いことを示す．例えば図 1 左上から Backdoor.Sdbot の発生過程を 1 日遅れで Backdoor.IRC.Bot の発生過程が追随する可能性が高いことが分かる．

評価結果を全てまとめると 3 種の相関図に分けることができた．3 種の相関図を図 3,4,5 に示す．図 3 は W32.Korgo 型を含む相関結果を示す．図 3 に示した W32.Korgo.I，W32.Ifbo.A と W32.Gobot.A が 2010 年 7 月 2 日に発生すると 5 日後の 7 日に枠線内全てのマルウェアで発生を確認した．複数のマルウェアが連鎖的に発生し，W32.Korgo 型のマルウェアは同時期に発生する可能性が高い．そのため，縦だけでなく横のつながりも存在することが分かる．W.32Korgo 型と W32.Ifbo.A は機能面で類似性が存在したが，他の 4 種については W32.Korgo 型等と類似性がみられなかった．図 4 は多階層型の相関結果を示す．図 4 に示した Infostealer.gampass が 2010 年 9 月 15 日に発生すると 1 日後の 16 日に W32.HLLW.Gaobot の発生を確認し，さらに 1 日後の 17 日に Spyware.Perfect の発生を確認した．また，Backdoor.Sdbot が 2010 年 11 月 2 日に急激な発生数の増加を示すと 1 日後の 3 日に Backdoor.IRC.Bot が急激に発生数を増加させ，さらに 1 日後の 4 日に Trojan の急激な発生数の増加を確認した．図 4 からも複数のマルウェアが連鎖的に発生していることが分かるが，図 3 の場合より縦のつながりが大きいという特徴がある．W32.HLLW.Gaobot，Backdoor.Sdbot と Backdoor.IRC.Bot は機能面で類似性が存在したが，他の 4 種については類似性が存在しなかった．ただし，Trojan はトロイの木馬型の汎用名であることからトロイの木馬型である Backdoor.Sdbot と Backdoor.IRC.Bot とは機能が類似している可能性がある．図 5 はマルウェアの発生の連鎖がすぐに停止した結果を示す．図 5 に示した W32.Waledac が 2012 年 6 月 16〜18 日の間に発生すると追随するように 17〜19 日の間で Trojan.FakeAV の発生を確認した．また，Trojan.FakeAV は W32.Waledac によってダウンロードされることがあるので相関が生じることは自明である．Hacktool はコンピュータシステムやネットワークを攻撃する可能性があ

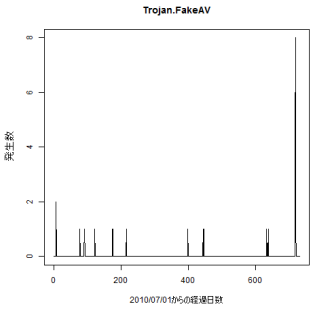

図 8　Trojan.FakeAV の発生過程

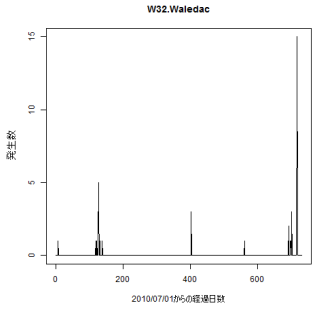

図 9　W32.Waledac の発生過程

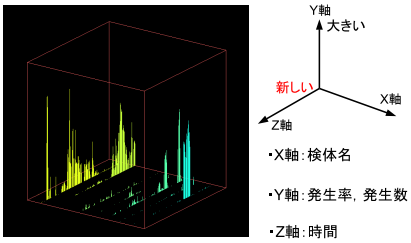

図 10　立方体各軸の構成

るプログラムの検出名のため，分散型サービス拒否攻撃を行う W32.Poxdar と相関が生じる可能性は高いと考えられる．

発生過程の相関が生じる原因の大半は機能の類似性やダウンローダ型のマルウェアの影響であると分かった．しかし，図 3,4 の一部には機能の類似性が存在しなかった．今回は機能の類似性のみで評価を行ったが，マルウェアの API ログや Symantec 社以外の検体名を用いることで相関の生じる原因が分かる可能性がある．また，k = 6 以上の相関を評価すれば新たな相関を発見することができるので類似性が生じる可能性がある．図 6,7 に W32.Gobot.A と InfoStealer.Lemir.Gen の発生過程を，図 8,9 に Trojan.FakeAV と W32.Waledac の発生過程を示す．

4　NONSTOP データを用いた発生過程可視化

1 章で述べたように，マルウェアの発生過程を単純にグラフ化するだけでは視覚的に捉えることは困難である．そこで本章では NONSTOP から得られた発生過程について 3D アニメーションでの可視化を提案する．可視化によりマルウェアの発生過程を直感的に理解することが可能となる．

4.1　システム概要

提案システムではマルウェアの解析レポートから得られた各セキュリティソフトベンダによる検体名を用いて，マルウェア発生過程の可視化を行う．可視化については，マルウェアの発生過程の流れが一目で判断できるようアニメーションを使用する．

まず，3.2 節で述べ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．そして，取得した発生過程をもとに発生過程の可視化を行う．可視化を行う

にあたり，まず 3 次元空間上に立方体を作成する．立方体の各軸の構成について図 10 に示す．x 軸は分類された検体名ごとに表示位置を与える．提案システムではユーザが可視化する検体名を入力として与え，x 軸の原点に近いほうから入力した内容順に表現する．y 軸は分類された検体ごとの発生率，もしくは発生数を与える．y 軸の上部ほど値が大きい．z 軸は時間となっており，奥部ほど新しい期間を与える．線分一本で 1 時間当たりの発生数もしくは発生率を表現し，可視化の対象期間は相互相関分析を行った期間と同じ 2010 年 7 月 1 日～2012 年 6 月 30 日である．そして，立方体の奥部から検体ごとの発生過程が手前に向けてアニメーション形式で流れる．

4.2 システム詳細

ユーザインターフェイスの実装は，より多くの情報をユーザに提供することを可能にする．以下に実装したユーザインターフェイスについて述べる．

1. 可視化する検体名を入力するためのフォームを表示
2. K,L キーによる発生率と発生数の表示モード切り替え
3. ↑，↓キーによる表示上限数の切り替え (発生数モードのみ)
4. ←，→キーによる時間経過の加減速 (巻き戻し可能)
5. スペースキーによるアニメーションの一時停止と再開
6. マウス操作による視点変更
7. モードと上限数の表示

1. では可視化用画面の右にユーザが可視化する検体名を入力するためのフォームを表示する．あいまい検索も可能であり，入力欄に backdoor% と入力することで backdoor 型の可視化を行うことが可能となる．2. ではキーボードの K, L キーによって表示形式の選択を行うことができる．K キーを押すと表示形式が発生率になり，L キーを押すと表示形式が発生数となる．3. ではマルウェアの発生数を表示しているときのみキーボードの↑，↓キーによってマルウェアの表示上限数を切り替えることができる．↑キーで表示上限数の増加，↓キーで表示上限数の減少が可能となる．4. ではキーボードの矢印キーによってアニメーションの時間経過の加減速を行うことができる．←キーで再生速度の減速，→キーで再生速度の加速が可能となる．5. ではキーボードのスペースキーによりアニメーションの一時停止と再開が可能となっている．スペースキーを一度押下することでアニメーションが一時停止し，もう一度スペースキーを押下するとアニメーションが再開する．6. ではマウス操作により視点の変更が可能になっている．マウスをドラッグして立方体の回転を行うことで，ユーザは立方体の裏側など見えずらい箇所を表示することができる．7. では表示モードと表示上限数の混乱を防ぐためにモードと上限数の表示を画面左上に行う．

4.3 システムの評価

本節では提案システムを実装し，マルウェアの発生過程を可視化した結果を示す．図 11 に発生率を可視化した際の例を示す．グラフ化では大まかな発生傾向しか把握できなかったが，提案システムでは 1 時間あたりの発生過程を表示できる．その結果流行した詳細な時期や増加傾向にあるマルウェアの発見が可能となり，マルウェアの発生傾向予測が容易になった．また，提案システムで

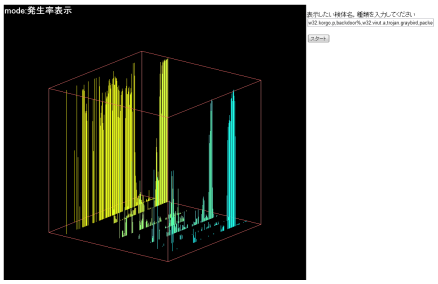

図 11　マルウェアの発生過程の可視化 (発生率モード)

はユーザが任意に表示させたい検体名を選択することができるため，相関の把握を行った後に確認をすることが容易になった．しかし，提案システムでは線分一本で 1 時間当たりの発生数もしくは発生率を表現しているため，ユーザはこれ以上詳細な発生傾向や大まかな発生傾向を知ることはできない．今後の課題として線分一本で表現する単位時間の変更も可能にする必要がある．

5 まとめ

本稿ではマルウェアの発生過程について時系列分析を行い，マルウェア同士の相関の有無を評価した．さらに，マルウェアの発生過程を 3D アニメーションで可視化するシステムを提案した．マルウェア同士の時系列分析には相互相関関数を用いた．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ことができた．発生過程の相関が生じる原因の大半は機能の類似性やダウンローダ型のマルウェアの影響であると分かった．しかし，一部の結果は機能の類似性が存在しなかった．このようなマルウェアに対して相関が生じた原因を調査していくことで今後のマルウェアの発生傾向の予測につながると考えられる．また，発生過程の単純なグラフ化では大量のデータを表示すると視覚認識能力を超えてしまう．そこで，マルウェアの発生過程を 3D アニメーションで可視化し，より詳細な発生傾向の把握が可能となった．今後はよりユーザインターフェイスの充実を図り，解析の支援を円滑にできるよう改良する．

謝辞

(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の有益なご助言，ご協力に感謝致します．

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/