改ざんサイトの傾向分析に基づいた検出手法

大西 祥生† 常松 直樹† 永塚 学† 野口 雅矢†

†株式会社 MC セキュリティ
690-0816 島根県松江市北陵町 51-2
css2013@sdl.hitachi.co.jp

あらまし 近年，改ざんサイトによるマルウェア感染が相次いでいる．特に Drive By Download 攻撃などの攻撃が頻発し，大きな問題となっている．また Blackhole Exploit Kit などの攻撃ツールが一般に出回っており，高度な技術を持たなくても攻撃が可能となっている．このような攻撃による被害拡大を防ぐためには，改ざんサイトを早期検出する必要がある．そこで本論文では，改ざんサイトの傾向に着目した新たな検出手法を検討する．また検討した検出手法を IPS に実装し，その有用性について検証を行う．

A Study of the detection method based on trend analysis of falsificated web site

Sachio Ohnishi† Naoki Tsunematsu†

Manabu Eitsuka† Noguchi Masaya†

†MC Security Co., Ltd.
51-2 Hokuryou-cho, Matsue-city, Shimane 690-0816, JAPAN 212-8567, JAPAN

Abstract Recently, the infection of malware infection caused by the falsificated website one after another. Attacks such as Drive-by download is increased, it has become a serious problem. Tools which attack website has spread too, for example, Blackhole Exploit Kit. It is possible to attack without advanced skill. We need to find falsificated site early to prevent from attack. In this paper, we focus on the trend of destination about falsificated website. We also implement module of IPS, which find tampered site. Finally we verify this method.

1 はじめに

技術の発達により，年々情報通信分野の重要性が増している．PC だけでなく，スマートフォンやタブレット端末の普及により，インターネットは人々にとって，より近いものとなった．その一方で，インターネットを介したマルウェアの被害が，年々増加している[1]．例えば Drive-By-Download 攻撃[2]の一種である Gumblar を始めとした，web サイトの不正改ざんによるマルウェア感染が，数多く確認されている．

マルウェア感染による被害を防止するためには，改ざんされたサイトを早期発見しなければ

ならない. しかし多様化するマルウェアに対して, 検出が困難になってきている. また複雑に難読化されたコードに対して, 既存の手法では対処できない場合もある. そのため新たな検知手法の確立が求められている.

本稿では, 改ざんサイトの転送傾向や改ざんコードの記述法に着目し, 新たな改ざんサイト検出手法の確率を目指す.

2 改ざん傾向分析

改ざんサイトの特性として, iframe タグをウェブサイトに組み込み, リダイレクトを行うものや, 難読化コードを挿入し, リダイレクトを行うものが挙げられる[3]. そこでリダイレクトに関する情報を分析し, 改ざんサイトの傾向, 及び特徴の調査を行う.

また改ざんされた web サイトには, 特徴的なコードが記述されていることが多い. これらの web サイトの解析を行うことで, 改ざんされた挿入コードの特徴を抽出し, 改ざんサイトの判定を行う.

3 関連研究

改ざんサイトの分析には以前より, 様々な手法が用いられている. 例えばマルウェアを可視化することにより, 攻撃に対する傾向分析が行われている. 改ざんサイトの傾向を分析することにより, サイトの特徴や, マルウェアに感染する危険性の高い国などが解析されている[4][5].

またドメインや IP アドレス, AS 番号などを用いた改ざんサイトの分析も行われている[6].

本研究では, リダイレクト時のパケットの流れや URL などに注目して, マルウェアの傾向分析を行う. そして危険性のある改ざんサイトの特徴を調査する.

4 転送経路分析実験

4.1 実験の目的

収集した改ざんサイトを, ページ遷移傾向やパケット通信解析を用いて分析を行う. それらの分析結果から, 改ざんサイトを検知し, 改ざんサイトからの被害防止を目的とする.

4.2 環境構築

実験を行うにあたり, 弊社サーバ上で実験用のための環境構築した.

環境構築には VMware vSphere を用いた. vSphere で仮想化環境を作成し, その上で仮想 OS を動作させた. OS には windows 7 を使用した. 一方クライアント側では Virtual Box を用いて, 仮想的に環境を構築した. なおクライアント側の仮想 OS は windows7 である.

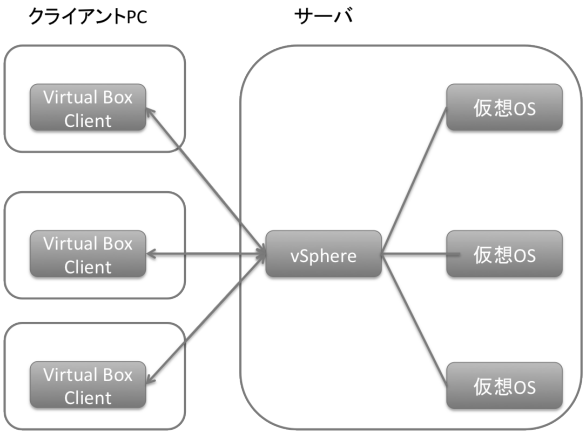

図 1 ネットワーク図

改ざんサイトの URL は, マルウェアサイトを収集している複数の web サイトより, 情報を取得した[7-10].

4.3 実験方法

構築した実験環境から収集した URL へアクセスを行い, wireshark を用いたパケット解析, リダイレクトの有無, 及び改ざん手法について調査する.

仮想 OS 上の処理が終わったのち，仮想 OS の環境を初期状態に戻し，再度収集を行う．

仮想 OS を用いることで作業の効率化，及び改ざんサイトからの被害防止を図る．

4.4 改ざんサイトの分析

「国名」，「文字数」，「拡張子」，「ランダム性」の四項目について着目し，改ざんサイトの傾向分析を行った．また四項目とは別に，それぞれの改ざんサイトに挿入されたソースコードを確認し，改ざんサイトの傾向を分析した．

4.4.1 国名

改ざんサイトおよび，リダイレクト先のドメインより国名を求め，記録する．

特定の国，及び言語圏で改ざんされる確率やリダイレクト前後の国名の変化に着目し，国別調査を行った．

リダイレクト前後の URL において，国名が異なる場合，改ざんが行われた可能性があると考えた．

4.4.2 URL 文字数

改ざんサイトよりリダイレクトを行う際，パラメータを用いて情報を転送している可能性がある．そこで URL の文字数を調べ，パラメータの有無や，改ざんサイトからのリダイレクトの調査を行った．

4.4.3 拡張子

改ざんサイトがリダイレクトを行う際，リダイレクト先のファイルはプレーンな HTML だけでなく，javascript や PHP など，別のプログラミング言語の場合がある．そこでリダイレクト前後の URL で拡張子が異なる場合，改ざんが行われた可能性があると考え，拡張子の遷移についても調査した．

4.4.4 ランダム性

リダイレクトされる URL は，ウイルス等が埋め込まれた，特殊なサイトである可能性が高いと考える．そのため．他の web サイトと比べ，URL が特徴的である可能性がある．そこで URL 表記の規則性について改ざんサイトとの関連性を調査する．

4.4.5 挿入コード

改ざんサイトによる攻撃では，難読化コードや，iframe などにより，リダイレクトさせるものが広く知られている．そこで web ページ内を調査し，難読化コードや iframe など，リダイレクトに繋がる部分の有無を調査する．

5 転送経路分析の結果

5.1 URL 文字数

改ざんサイトの URL 文字数を集計したものが図 2 である．

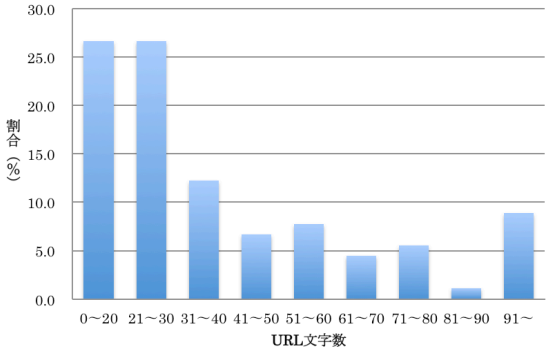

図 2 URL 文字数

図2より，URL の平均的な長さは 76 文字で，標準偏差は 37.41 である[11]．今回の結果については，62%の URL が平均的な URL 長を逸脱し，38%が標準偏差内である．

5.2 国名

改ざんされたサイトの国別の集計データが図 3 である.

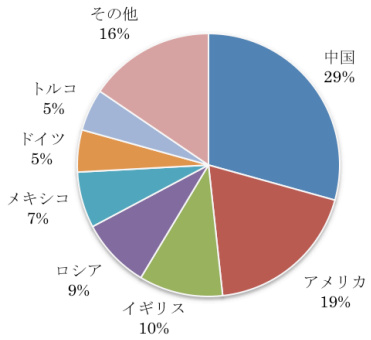

図 3 国別統計

改ざんされたサイトの数では，中国が最も多く，ついでアメリカやイギリスなどの国が並んでいる.

5.2.1 国名の遷移

リダイレクト前の URL とリダイレクト後の URL を比較し，国名の遷移をまとめたものが図 4 である.

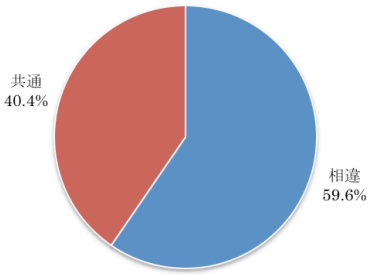

図 4 国名の遷移

図 4 より，リダイレクト後の国名が異なるものが 59.6%であり，半数以上の URL で国名の変化が確認できた．別の国へリダイレクトされた URL については，主に中国やロシアといった国に対し，リダイレクトが行われていることが確認できた.

5.3 拡張子

取得した URL に対して，拡張子を調査したものが図 5 である.

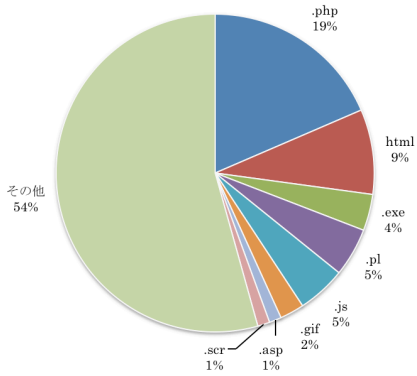

図 5 拡張子

図5より，全体の 45.7%が，何らかのファイルの拡張子を伴ったアドレスである．また，html ファイルが 8.6%であったのに対し，それ以外の形式のものが 37.1%あった.

5.3.1 拡張子の遷移

リダイレクト前の URL とリダイレクト後の拡張子を比較し，その遷移を調べたものが図 6 である.

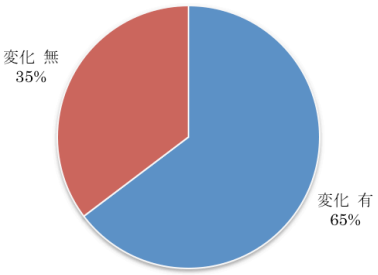

図 6 拡張子の遷移

図6より，全体の 65%がリダイレクト前後で，拡張子が異なることが分かった．そのためリダイレクト時に拡張子が遷移していた場合は，改ざんサイトである確率が高い.

5.4 ランダム性

URL 表記の規則性について改ざんサイトとの関連性を調査したものが，図7である．

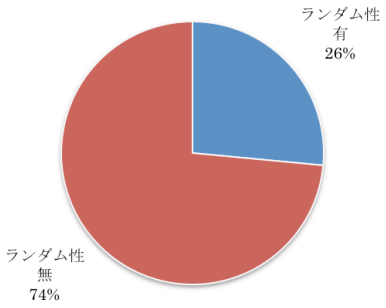

図 7 URL 表記の規則性

図7より，URL 中でランダム性があるものは26%程度であり，ランダム性が無いものが多い，という結果になった．

5.5 挿入コード

改ざんサイトに挿入されたソースコードについて調査したものが，図8である．

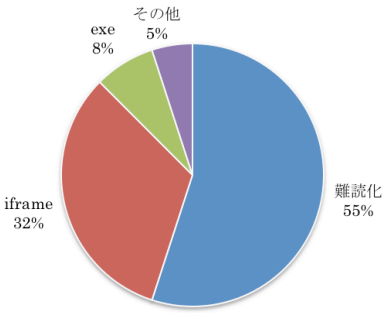

図 8 攻撃手法

図8より，挿入コードでは難読化を用いて，他のサイトへ転送するものが全体の半数以上となった．また iframe を用いて，ページ閲覧者が気づかないうちに，別のサイトへリダイレクトを行うものが 32%と大きな割合を占めた．

またファイル自体が exe 形式であり，悪影響をおよぼすものや，scr 形式（スクリーンセーバー）のような特殊な形式のファイルにアクセスを行う事例などを，確認することができた．

6 転送経路分析実験の考察

6.1 URL 文字数

URL 長はパラメータの有無や，ドメインの長さによって大きく差が生じた．ただし今回アクセスしたページの多くが，web サイトのトップページであるため，パラメータが書かれていない可能性が高く，必然的に URL 文字数が少なくなる傾向がある．そのためトップページの URL 文字数に注目し，さらなる検討を行う必要がある．

6.2 国名

実験結果で特徴的なことは，中国の割合である．Web サイトで使われている言葉の 55%は英語であり，中国語は 4%であった[12]．そのため，アメリカやイギリスといった英語圏の国々が，最上位になると推測できる．しかし図3の国別統計データより，中国が上回るという結果となった．このことより，他国に比べ中国の web サイトは，改ざんの危険性が高いと思われる．

また図4より，他国へのリダイレクトが行われた場合，マルウェア感染の危険性が高まると言える．

6.3 拡張子

ディレクトリ名を指定した時にだけ表示されるページ(index.html など)は，ファイルの拡張子が不明であるため，除外して考察を行う．

図5より，html 形式に比べ，それ以外の形式のファイルが多い．そのため html 以外の形式であった場合，html 形式よりも危険な可能性がある．

また通常のリダイレクトでは，アクセスすることが考えづらい exe ファイルにも，アクセスしていると判明したため，検出の際には注意をして分析する必要がある．

拡張子の遷移が行われない場合に比べ，行われた場合のほうが多い．そのためリダイレクト時に拡張子が変異していたときは，改ざんサイトの可能性がある．

6.4 ランダム性

図7においてランダム性の無い場合が多い理由は，リダイレクト時にパラメータを用いない場合が多いことが考えられる．5.4 の実験結果より，ランダム性は，改ざんサイトとの関連性が薄いと思われる．しかしリダイレクト時にパラメータを用いる場合や，特殊な文字列の web サイトへとアクセスすることも有り，一概に無関係とは言えないため，今後さらなる検討をしていく必要がある．

6.5 挿入コード

5.5 の実験結果より，挿入されたコードのほとんどは，他サイトへのリダイレクトによるものだと分かった．そのためリダイレクトを検知し，適切に遮断をすることで，改ざんサイトからの被害を抑えることが可能である．

7 改ざんコード傾向分析実験

7.1 実験の目的

html ソース中の記述より，改ざんサイトの傾向を分析し，検出をはかる．

マルウェアに感染させる手法としては，大きく難読化コードを用いるものと，iframe を html 中に挿入するものがある．本実験ではこの両者の検知が可能か確認する．

難読化コードは，特定の記号で区切るなど，種類によって特徴的な記述がある．そのため文字列検出を行うことで，挿入コードの検知を行い，改ざんサイトを検出する．また不自然な位置に書かれた iframe を検知し，改ざんサイトと判定する．

実験の対象としては，検体のデータセットである，D3M2013，D3M2012，D3M2011 を用いた．

7.2 検知パターンの設定

検出のためのパターンを設定し，改ざんサイトの検出を行う．難読化コードは，特定の区切り文字を用いるなど，特徴的な記述がある．パターンの設定については，区切り文字やスクリプト中のコメントを基に行っている．また iframe はソースコードの末尾など，不自然な箇所にあった場合について，改ざんサイトと判定する．

8 挿入コード傾向分析の結果

検知パターンを用いて，改ざんサイトの検知を試みた（図 9）．

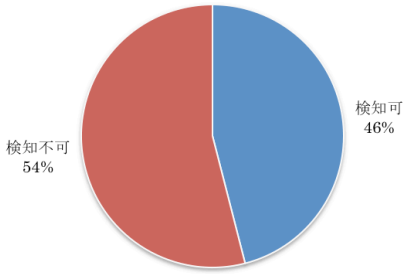

図 9 傾向分析結果

対象とした改ざんサイトにおいては，半数近くの実験対象で検出が可能であった．その一方で，パターンの特定が難しい事例もあった．

9　挿入コード傾向分析の考察

8章の実験結果より，複雑なパターンへの対処が難しいことが分かった．例えばハッシュを用いてリダイレクトを行うものや，不規則に区切り文字を変更する難読化コードについては，対処が難しかった．今回収集した改ざんサイトについては，例のような複雑な改ざんのものは少なかった．しかし今後複雑な改ざん手法が現れた場合，対処できない可能性がある．そのため検知パターンの拡充を進める必要がある．

10 まとめ

本研究では，転送傾向と，挿入コードのパターンをもとに，改ざんの傾向について調査を行った．

転送傾向については URL の記述と，ドメインより得られる国名の情報に着目した．5.1 の結果及び，国名については 5.2 の結果より，国によって改ざんサイトに遭遇する可能性が異なるという結果が得られた．特に中国，アメリカで51%を占めており，これらの国のサイトにアクセスする場合は改ざんサイトに遭遇する可能性が高まる．またリダイレクト時に国の遷移が生じた場合，更に可能性が高まるデータが得られた．

挿入コード傾向分析では，図 9 より検知可能な挿入コードが 46%であった．この結果から挿入コードのパターンマッチングのみでも，多様化する改ざんサイトからの被害防止に一定の効果があった．

今後も継続してデータを収集し，検知パターンの拡充，及び新たな改ざんサイトの検知手法を確立し，検知プログラムの改良を続けていく必要がある．

謝辞

本稿の作成にあたり，ご協力してくださった皆様に深く感謝いたします．

参考文献

[1] Microsoft.(2012)「マルウェアの進化と脅威の状況 -10 年間の振り返り」 Microsoft Security Intelligence Report.
[2] A. Niki.(2009),Drive-by download attacks: Effects and detection methods. PhD thesis, Master's thesis, Royal Holloway University of London.
[3] Provos, N., et al. (2007), The ghost in the browser analysis of web-based malware: USENIX Association.
[4] 松木隆宏，新井悠,(2009)，「CCC DATAset 2009 によるマルウェア配布元の可視化」, MWS2009.
[5]金子博一，(2011),「地理的可視化を用いたマルウェアの統合解析」, MWS2011.
[6]福島祥郎，堀良彰，櫻井幸一，(2010),「ドメイン情報に着目した悪性 Web サイトの活動傾向調査と関連性分析」, MWS2010.
[7]「urlquery」 <http://urlquery.net/>(2013/8/16 アクセス)
[8] 「web inspector」 <http://app.webinspector.com/>(2013/8/16 アクセス)
[9] 「malware domain list」 <http://www.malwaredomainlist.com/>(2013/8/16 アクセス)
[10] 「malware patrol」 <http://www.malware.com.br/>(2013/8/16 アクセス)
[11] 「Supermind Consulting」 <http://www.supermind.org/blog/740/average-length-of-a-url-part-2>(2013/8/16 アクセス)
[12]「Usage of content language for website」 <http://w3techs.com/technologies/overview/content_language/all> (2013/8/16 アクセス)