Web ブラウジングデータ観測による悪性リダイレクトチェーンの構造分析

嶌田 一郎[†1]　太田 敏史[†1]　白石 訓裕[†2]　中嶋 淳[†2]　山田 明[†3]

概要：Web サイトの閲覧を媒介としてマルウェアに感染させる Web 媒介型攻撃が深刻な問題となっている．Web 媒介型攻撃は，攻撃基盤のテイクダウンに耐性を持たせるためやマルウェア配布サーバを隠しブラックリスト登録から逃れるために，複数の Web サイトを自動的に遷移させるリダイレクトを用いた構造をもつことが知られている．2018 年 6 月から，Web ブラウザ拡張を一般ユーザに配布して Web アクセス情報を収集する実験を開始し，2020 年 7 月現在において約 10,000 名を超えるユーザを集めている．本稿では，複数のユーザから収集したリダイレクト情報を構造分析することによってリダイレクト構造の特徴を明らかにし，悪性リダイレクトチェーンに固有の特徴を捉えた悪性リダイレクトチェーンを特定するための新しい方式を提案する．提案方式は，Web コンテンツを全く参照しないため，難読化の影響を受けず悪性 Web サイトへのアクセスを特定することができる．収集したリダイレクト情報を用いて提案方式を評価した結果，悪性サイトへのリダイレクトを正確に特定する上で非常に効果的であることを実証した．

キーワード：Web 媒介型攻撃，Drive-by Download，Redirection Chain，構造分析，ネットワーク分析

Structural Analysis of Malicious Redirection Chains by Observing Web Browsing Data

Ichiro Shimada[†1]　Toshifumi Oota[†1]　Kunihiro Shiraishi[†2]
Jun Nakajima[†2]　Akira Yamada[†3]

Abstract: Web-based cyber attacks, which infect computers with malware when users browse malicious websites, have become a serious problem. Multiple websites are automatically redirect in order to make them more resilient to infrastructure takedowns and to avoid blacklist registration by hiding malware distribution servers. An experiment in which a browser extension was distributed to general users was carried out in June 2018, and information for about 10,000 users was collected as of January 2020. In this paper, we propose a new method for identifying malicious redirection chains that capture the unique characteristics of them by analyzing redirection information. Since the proposed method does not reference web content at all, it is unaffected by obfuscation and can identify access to malicious websites. We evaluated the proposed method using the collected information and determined that it is very effective in accurately identifying redirects to malicious websites.

Keywords: Web-based cyber attack, Drive-by Download, Redirection Chain, Structural Analysis, Network Analysis

1. はじめに

Web サイトの閲覧を媒介としてマルウェアに感染させる Web 媒介型攻撃が深刻な問題となっている．Web 媒介型攻撃は，攻撃基盤のテイクダウンに耐性を持たせるためやマルウェア配布サーバを隠しブラックリスト登録から逃れるために，複数の Web サイトを自動的に遷移させるリダイレクトを用いた構造をもつことが知られている．攻撃は，複数のページをリダイレクトしながら遷移し悪性サイトへ誘導するため，Web 上の攻撃ページにどのように到達し，マルウェアをダウンロードさせるか観測し，ページ遷移のコンテキストを把握することが，攻撃への防御策を検討する上では重要である．本稿では，複数のユーザから収集したリダイレクト情報を構造分析することによってリダイレクトの特徴を明らかにし，悪性サイトへの Web アクセスを検知する方式を提案する．提案方式は，Web コンテンツを全く参照しないため，難読化の影響を受けず悪性 Web サイトへのアクセスを特定することができる．収集したリダイレクト情報を用いて提案方式を評価した結果，悪性サイトへのリダイレクトを正確に特定する上で非常に効果的であることを実証した．

以下，2 章では，関連研究について述べ，3 章では，使用したデータセットについて述べる．4 章では，悪性リダイレクトチェーンの特徴抽出について述べる．そして，5 章では，悪性リダイレクトチェーンの構造分析について述べる．6 章で評価について述べ，7 章でまとめを述べる．なお本稿では，Web ページ閲覧を「Web ページアクセス」，Web ページアクセスにより収集したデータを「ブラウジング情

†1 株式会社 構造計画研究所
KOZO KEIKAKU ENGINEERING Inc.
†2 株式会社 セキュアブレイン
SecureBrain Corporation
†3 株式会社 KDDI 総合研究所
KDDI Research, Inc.

報」，ブラウジング情報から抽出した Web ページアクセスでの履歴に関する情報を「Web アクセス履歴」，Web ページを訪れたユーザを別ページへ自動的に誘導・遷移する仕組みを「リダイレクト」，複数の Web サイトを繰り返し経由するリダイレクトを「リダイレクトチェーン」と定義し使用する．

2. 関連研究

Drive-by Download（DBD）攻撃の防御を開発・改良する目的でマルウェアのダウンロード経路を研究した文献として[1]がある．[1]では，DBD 攻撃対策として WebWitness と名付けられたインシデント調査システムを提案している．WebWitness では，Web 上のマルウェアのダウンロード経路をユーザのブラウジング情報から自動的にトレースバックしラベル付けしている．ラベルは，現在の攻撃傾向をよりよく把握し，より効果的な防御策を開発するために活用している．ブラウジング情報からマルウェアのダウンロードトレースバックを行うのに，ブラウジング情報から経路を再構築する必要があり，トレースバックの方法として，ログ上の Referer フィールドと Location フィールドを利用して Web アクセス間をリンクさせる方法がある．しかし，この方法では，ユーザが使用しているブラウザ，JavaScript，プラグインソフトウェアの特定のバージョンなどに依存して Referer フィールド，Location フィールドが抑制され，Web アクセス間のリンクが正しく再構築できない場合があるという課題がある．このため[1]では，”referrer indicator” を導入して Web アクセス間のリンク関係の重み付けをして，複数の経路から重要な経路を選択することで自動的にトレースバックする手法を提案している．提案手法は，大規模な学術ネットワークでの実験で，既存のブラックリスト手法と比較し DBD 攻撃での感染率を 6 倍低減している．本稿においては，Google Chrome の chrome Extension API を利用した Web ブラウザ拡張（ブラウザセンサ）により Web ブラウジングログを収集することで，リダイレクトによるページ遷移を追跡する手法[2]を用いた．リダイレクトチェーンの抽出方法については 4 章で記述する．

大規模で多様な Web ブラウジングログから HTTP リダイレクトグラフを生成し，その特徴を分析することで悪性 Web ページを検出する研究として[3-4]がある．[3]では，悪性のリダイレクトチェーンの特徴として，クロスドメイン・リダイレクトの有無，トラフィックを分散させるためのハブの存在，コモディティ・エクスプロイトキットの使用，被害者の地理的分散，などの特徴量を SVM により分類することで，重要な特徴を抽出している．評価の結果，悪性のリダイレクトチェーンを識別する効果的な特徴として，リファラーの所属国の多様性，リダイレクト先が集約されるハブサイトの存在，ドメイン名が IP アドレスであること，などを指摘している．[4]では，大規模な ISP のデータセットを用いてユーザ毎に HTTP リダイレクトチェーンを構築し，悪性のリダイレクトを統計的特徴量に基づき決定木で分類し実証している．本稿においても，攻撃先の国の多様性と集約サイトの存在に着目し，リダイレクトチェーンの構造を分析した．分析手法については 5 章で記述する．

悪性リダイレクトチェーン検出を実装しているシステムとして，この他，高対話型クライアントハニーポットが記録した URL やリダイレクトチェーンから，マルウェアを配信する URL を識別するシグネチャを生成する ARROW[5]，検索用語とは関係のないマルウェア配布サイトへのリダイレクトを特定することに焦点を当てた SURF[6]，Twitter に投稿された悪意のある URL に関連するリダイレクトチェーンとツイートのコンテキスト情報の相関から得られる特徴により不審な URL を検出する WarningBird[7]などが開発されている．

本稿では，ユーザ参加による Web 媒介型攻撃対策 (WarpDrive:Web-based Attack Response with Practical and Deployable Research InitiatiVE)[8]で開発されたシステムにより収集した Web ブラウジングログを実験に用いた．WarpDrive は，一般ユーザの参加によって観測環境を構築する点が特徴となっている．システムはブラウザセンサと分析基盤からなっており，センサで収集されたブラウジング情報を分析基盤に集め，ユーザ端末で発生している攻撃の実態把握を実現している．ブラウザセンサを一般ユーザに配布した実証実験を 2018 年 6 月から開始して，約 3 ヶ月間で 5,000 人のユーザが利用し，毎日約 15 億件のブラウジング情報が収集されている．2020 年 7 月現在において約 10,000 名のユーザを集めている．本稿では，収集したログから抽出したリダイレクトチェーンの構造を分析し，得られた特徴を用いて悪性リダイレクトチェーンの検出を試み評価した．

3. データセット

本章では，WarpDrive のシステム概要，データセットの特徴について述べる．

3.1 システム概要

図 1 に，WarpDrive のシステム構成を示す．本システムは，ユーザへ配布したブラウザセンサによりユーザのブラウジング情報を収集し，分析システムを用いて分析するための分析基盤である．分析システムに集められたブラウジング情報をもとに，悪性サイトへのアクセス分析を行い，分析結果は被害の未然防止に活用する．

3.2 データセットの特徴

収集するデータは，ブラウザがページを表示する際に得られるページに関連する情報，URL で要求した際のリクエストヘッダ，レスポンスヘッダ，レスポンス IP ステータスコード，リクエスト処理を開始した時間，リクエスト処理

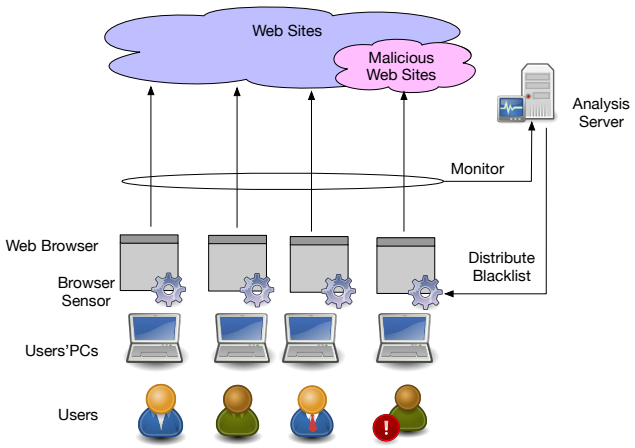

図 1 実証実験のシステム構成

を完了した時間, 表示に関わるブラウザ情報などである (cf. [2]付録 A).

収集したブラウジング情報からリダイレクトチェーンを抽出し, リダイレクトチェーンのデータセットを作成する. 参照するブラウジング情報として, リダイレクト時の tabid（ブラウザタブを識別する id), type（リソース種別:タブにトップレベルのドキュメントがロードされたか等の情報）を利用し, 以下のルールで抽出する.

(1) type が"main_frame"の履歴情報
(2) tabid が同一のログ（同じタブ上でのページ遷移）
(3) (1)(2)のデータの時系列順ソート

本稿では, データセットとして Web サーバからの 3XX 応答で発生するリダイレクトであるサーバリダイレクトを使用した（cf. [2]付録 B).

リダイレクト構造に関する本データセットの特徴は, 以下の通りである.

(1) Referer ヘッダに依らないページ遷移の把握が可能
(2) Location ヘッダと 3 XX ステータスコードに依らないサーバリダイレクト発生の識別
(3) ブラウジングに関係する全 URL の収集

本稿では, データセットの特徴に基づきリダイレクトチェーンの構造を明らかにし, 悪性リダイレクトの分析を行う.

4. 悪性リダイレクトチェーンの特徴抽出

本章では, まず, 悪性リダイレクトチェーンの特徴について述べ, 次に, データセットから悪性リダイレクトチェーンを抽出する方法について述べる.

4.1 悪性リダイレクトチェーンの特徴

Web 媒介型攻撃は, 攻撃基盤のテイクダウンに耐性を持たせるためや, チェーン内の中間ステップを柔軟に変更することでマルウェア配布サーバを隠し, 検知を回避しつつ OS やブラウザが脆弱な場合にだけ攻撃を実施するためにリダイレクトチェーンを用いた構造を持つことが知られている. 悪性リダイレクトチェーンの特徴として, [3]では, 攻撃基盤のテイクダウンに耐性を持たせるために複数のドメインを跨ぐ「クロスドメイン・リダイレクト」[a], トラフィックを分散させ集約ページへリダイレクトすることで攻撃基盤の堅牢性と新規の悪性サイトへのリダイレクトの容易性を確保するための「ハブの存在」, 攻撃者は, できるだけ多くの攻撃先へ攻撃を実施するため, 攻撃先が多様な国や地域に分散する「被害者の地理的分散」[b], などを指摘している.

一方, [2]の分析において, 電子メールアドレスなど個人情報を要求するフィッシング詐欺サイトへ至るリダイレクトチェーンから以下の結果を得た.

(1) リダイレクトに要する時間
リダイレクトの開始から最終ページへ至るまでの経過時間が数秒以内.
(2) リダイレクトの回数（リダイレクトチェーンの長さ）
数回のリダイレクトを経て最終ページへ至る.
(3) 中継サーバの所属国の多様性
国外のサーバを経由して最終ページへ至る.
(4) リダイレクトチェーンの構造的特徴の存在
リダイレクトチェーンの開始は全て同じサイトであり, 中間ステップ（中継サーバ）は複数あるが, 集約サイトにリダイレクトされ最終ページに至る.

(3), (4)の特徴は, 文献[3]の「攻撃先国の多様性」,「ハブの存在」と同じ特徴を有している.

4.2 リダイレクトチェーン抽出アルゴリズム

4.1 節の分析に基づきデータセットから悪性リダイレクトチェーンを抽出する. 図 2 に悪性リダイレクトチェーンの抽出フローを示す. まず, 候補となるリダイレクトチェーンを 3.2 節の方法で抽出する. 次に, 候補となるリダイレクトチェーンから, リダイレクトの経過時間が 5 秒以内(4.1 節(1)), リダイレクトの回数が 2 回以上(4.1 節(2))のデータを抽出する. 次に, リダイレクトチェーンを構成する各サーバのドメインの所属国が全て日本[c](4.1 節(3))である候補を除外し, [3]の「クロスドメイン・リダイレクト」の特徴を用いて, リダイレクトチェーン内に重複するドメイン名が存在する候補を除外する. 表 1 に, 除外する場合のデータ例を示す.

更に, 利用頻度の高い Web サイトは, よくメンテナンスされていることが多く, 一般的に危殆化し難いため, Alexa[9]の top 1 million（以下, Alexa top-1m）を活用して利用頻度の低いサイトを悪性リダイレクトチェーン候補とし

a) 良性リダイレクトチェーンの場合, 同じドメインのページが二つ以上存在する. 例として, 同じドメイン内のログイン認証ページへリダイレクトし, ログイン認証後にリダイレクト前のページへ戻る遷移などがある.
b) 良性の場合は, 殆どのユーザが訪問対象の Web サイトの所属国から訪れており, 国の多様性は高くない.
c) WarpDrive 実証実験は, 日本国内限定で実施している.

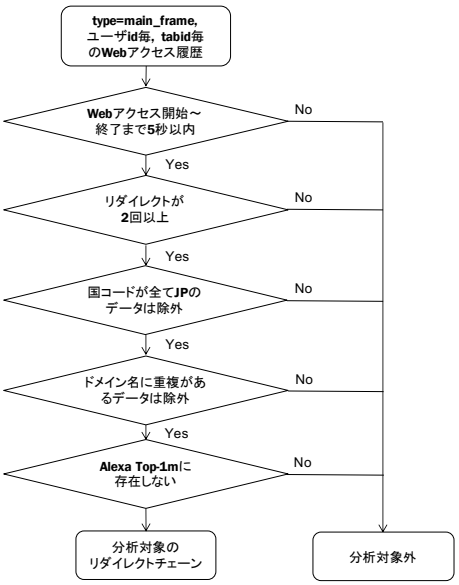

図 2 悪性リダイレクトチェーンの抽出フロー

表 1 除外する場合のデータ例（CC:国コード）

アクセス時刻	CC	URL
2019-02-04 08:21:56.011	JP	http://example.jp/A123456789/111111
2019-02-04 08:21:56.264	JP	http://www.example.com/detail.php?id=123
2019-02-04 08:21:56.294	JP	https://www.example.com/detail.php?id=123
2019-02-04 08:21:56.323	JP	https://www.example.com/error.php

て選出する．図 3 に候補選出の概念図を示す.

図 3 の選出方法では，Alexa top-1m にドメイン名が存在しなければ「利用頻度が低い Web サイト」であると定義し，リダイレクトチェーン上に 1 件でも「利用頻度が低い Web サイト」があれば，悪性リダイレクトチェーンの候補として選出する.

4.3 抽出結果

4.2 節の手法で抽出した結果の例を図 4 に示す．行（トランザクション）単位で，左から右の順で時系列・リダイレクト順にドメイン名をカンマ区切りで列挙し，ひとつのリダイレクトチェーンを構成している.

5. 悪性リダイレクトチェーンの構造分析

本章では, 抽出した悪性リダイレクトチェーンの候補を，ドメイン名の共起という観点でアルゴリズムを適用し悪性リダイレクトチェーンの分析を行う.

5.1 適用アルゴリズム

図 5 に，2019 年 2 月 3 日のリダイレクトチェーンを可視化したグラフを示す. 図 5 には, 4.1 節(4)の特徴とした，少数の起点となるサイトの存在(e.g., h**tinlethemsed.info, h**tonsfeetred.info)[d], 数回のリダイレクトと複数の経路の存在が確認でき，起点からの複数の経路上にドメイン名が共起する特徴がある．また，起点からの経路上に，リダイレクトが集約される集約サイトの存在している(e.g., 5**n7.com, a**etix.com, t**cking.marketing).

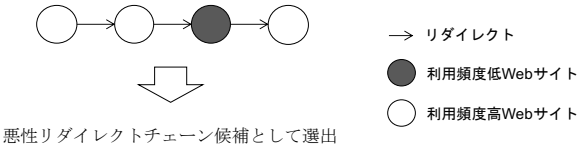

図 3 候補選出の概念図

h**tonsfeetred.info, a**etix.com, 5**n7.com
h**tonsfeetred.info, l**tds.net, a**rackingout.com
h**tinlethemsed.info, e**k.club, t**bestleadbit.com
h**tonsfeetred.info, a**etix.com, 5**n7.com, p**.doublepimp.com
h**tonsfeetred.info, p**under.bid, p**tios-raj.com
h**tonsfeetred.info, d**italdsp.com, p**affiliates.com
p**checke.club, i**brt.com, l**al-finders.com, s**rter-tech.net
p**checke.club, i**brt.com, a**tdska.pro
h**tinlethemsed.info, e**k.club, s**vanus-phe.com, t**cking.marketing, m**rosoft.com-repair-windows-system.live
h**tinlethemsed.info, l**tds.net, a**rackingout.com
h**tinlethemsed.info, f**wwiththetide.xyz, p**u5l.com
h**tinlethemsed.info, a**etix.com, 5**n7.com
h**tonsfeetred.info, a**etix.com, 5**n7.com ，t**cking.marketing, a**le.com-cleaning-os.live
h**tonsfeetred.info, c**ckredirection.com, s**rtiyada.com
v**roz.xyz, e**ionale.info, t**thandarinhec.info
h**tinlethemsed.info, f**wwiththetide.xyz, p**.doublepimp.com
h**tinlethemsed.info, l**tds.net, a**rackingout.com
h**tonsfeetred.info, a**etix.com, 5**n7.com
h**tonsfeetred.info, a**etix.com, 5**n7.com
h**tinlethemsed.info, e**k.club, t**bestleadbit.com
h**tinlethemsed.info, a**etix.com, 5**n7.com
h**tinlethemsed.info, e**k.club, s**vanus-phe.com, t**cking.marketing, m**rosoft.com-repair-windows-system.live
h**tinlethemsed.info, e**k.club, t**bestleadbit.com
h**tonsfeetred.info, t**ilter.info, c**ckssp.pro, y**rprivacy.icu
h**tonsfeetred.info, a**xchangemedia.xyz, p**u5l.com
h**tinlethemsed.info, t**ilter.info, c**mank.pro
h**tinlethemsed.info, a**etix.com, 5**n7.com ，b**wsergames2018.com
h**tinlethemsed.info, a**peratorx.com, s**at0rx99.com

図 4 抽出結果の例

上記の特徴から，適用アルゴリズムとして，ドメイン名の共起を検出する目的で，アソシエーション分析手法の Apriori アルゴリズム[10]による頻出パターンマイニング，及び SPADE アルゴリズム[11]による系列パターンマイニングを適用する．また，起点サイト，及び集約サイトの検出を目的として，ネットワーク分析手法から次数中心性と媒介中心性のアルゴリズムを適用する．[e]

5.2 Apriori アルゴリズムによる頻出ドメイン名の抽出

Apriori アルゴリズムは，頻出アイテム集合（本稿ではドメイン名の集合）とアソシエーションルール（相関ルール）を検出するアルゴリズムである．2019 年 2 月 3 日のデータセットから, 4.2 節の手法で抽出した悪性リダイレクトチェーン候補に対して Apriori アルゴリズムを適用した．表 2 に，confidence 値 1.0，support 値 0.05 で適用した結果の 78 件中の support 値上位 10 件を例として示す. 適用にあたり，相関ルールが複数のトランザクションに存在することを重視し support 値優先で適用した．検出した相関ルール（条件部，結論部の両方）から頻出ドメイン名をユニークに抽出し，ドメイン名の悪性判定を VirusTotal[12]により行った結果を表 3 に示す．　抽出した結果，14 ドメイン中の 7 ドメインが悪性判定された．一方で，7 ドメインは悪性判定されないため，Web 上で一般公開されている公開レポートを

d) 本稿ではドメイン名の一部を*で置き換えて表記する.

e) R の arules, arulesSequences, igraph パッケージを利用している.

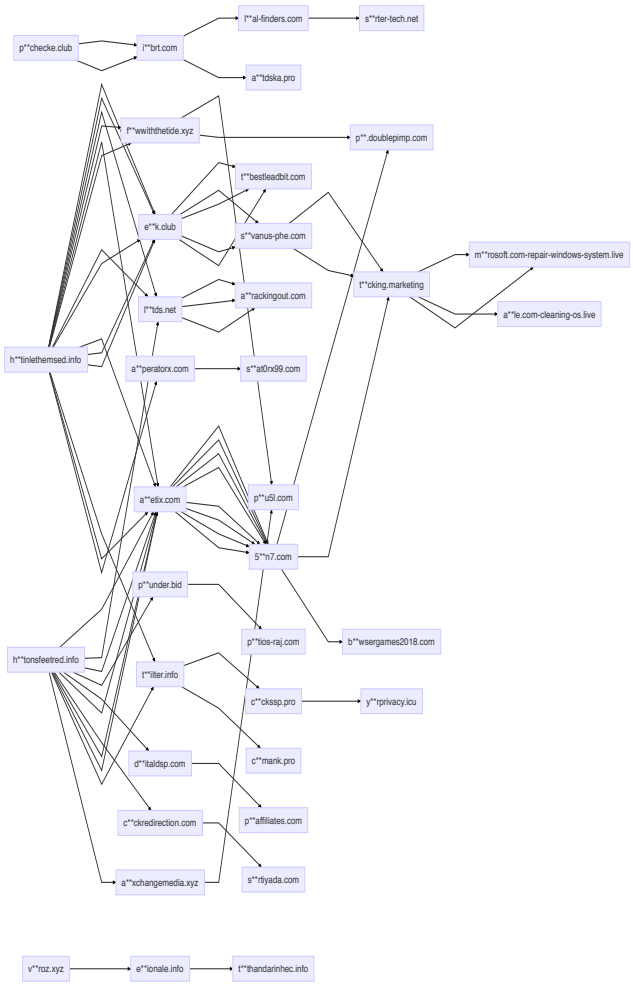

図 5 リダイレクトチェーンのグラフ

表 2 Apriori アルゴリズムの適用結果（TOP 10）

条件部	結論部
{5**n7.com}	{a**etix.com}
{a**etix.com}	{5**n7.com}
{e**k.club}	{h**tinlethemsed.info}
{5**n7.com, h**tonsfeetred.info}	{a**etix.com}
{a**etix.com, h**tonsfeetred.info}	{5**n7.com}
{a**rackingout.com}	{l**tds.net}
{l**tds.net}	{a**rackingout.com}
{t**bestleadbit.com}	{e**k.club}
{t**bestleadbit.com}	{h**tinlethemsed.info}
{e**k.club, t**bestleadbit.com}	{h**tinlethemsed.info}

表 3 抽出ドメイン名と VirusTotal の判定結果

頻出ドメイン名	悪性判定エンジン数/判定エンジン総数
5**n7.com	2/80
e**k.club	1/80
h**tinlethemsed.info	1/71
h**tonsfeetred.info	4/71
i**brt.com	1/79
m**rosoft.com-repair-windows-system.live	1/69
p**checke.club	2/70

表 4 抽出ドメイン名と公開レポートでの報告有無

頻出ドメイン名	報告サイト	報告概要
a**etix.com	MALWEARETIPS[13]他	redirect adware
a**rackingout.com	MALWEARETIPS 他	redirect adware
f**wwiththetide.xyz	ANY.RUN[14]	malicious activity
l**tds.net	MALWEARETIPS 他	redirect adware
s**vanus-phe.com	TROJAN KILLER[15]他	redirect adware
t**bestleadbit.com	報告なし	–
t**cking.marketing	ANY.RUN	malicious activity

調査した結果，VirusTotal で悪性判定されなかった 7 ドメイン中の 6 ドメインがリダイレクトアドウェアなどの悪性サイトとして報告されており，悪性のリダイレクトであることが確認できた（表 4）．

5.3 SPADE アルゴリズムによる系列パターンの抽出

SPADE(Sequential PAttern Discovery using Equivalence classes)アルゴリズムは，系列パターンマイニング(Sequential Pattern Mining)の一つであり，時系列の順序関係を考慮した相関ルール検出アルゴリズムである．リダイレクトチェーンには，時系列の Web アクセス順序があるため，リダイレクトチェーンを系列データと見做し SPADE アルゴリズムを適用し，リダイレクトチェーン上のドメイン名の頻出パターンを抽出した．2019 年 2 月 3 日のデータセットに対して適用した結果の頻出パターンを表 5 に示す．表 5 は，confidence 値 1.0，support 値 0.05 で適用した結果の件中の support 値上位 10 件である．Apriori アルゴリズムと同様に，検出した相関ルール（条件部，結論部の両方）から頻出ドメイン名をユニークに抽出し，ドメイン名の悪性判定を VirusTotal 行った結果を表 6 に示す．抽出した結果，12 ドメイン中の 7 ドメインが悪性判定された．一方で，5 ドメインは悪性判定されなかった．一般公開レポートで調査した結果，VirusTotal で悪性判定されなかった 5 ドメイン全てがリダイレクトアドウェアなどの悪性サイトとして報告されており，悪性のリダイレクトであることが確認できた．

5.4 中心性指標による頻出ドメイン名の抽出

ネットワーク分析とは，事象を頂点と頂点を結ぶ線（エッジ）で抽象化しネットワーク構造を分析する手法である．本稿では，頂点を Web サイト（ドメイン名），エッジをリダイレクトとして分析を行った．ネットワークから重要な頂点を検出する手法として中心性指標がある．本稿では，起点サイトを検出するために，他の頂点と接続するエッジ数（次数）が多い頂点を検出する次数中心性（degree centrality），及び集約サイトを検出するために，2 つの頂点間に現れる頂点数を指標とする媒介中心性（betweenness centrality）を用いて分析を行った．媒介中心性の定義は次の通りである．頂点vの媒介中心性 $CB(v)$ は，始点 s と終点 t の頂点ペア$(s,\ t)$の最短路の個数をσ_{st}，始点 s と終点 t の頂点ペア$(s,\ t)$の最短路のうち頂点vを通る経路の個数を

表 5 SPADE アルゴリズムの適用結果（TOP 10）

条件部	結論部
{a**etix.com}	{5**n7.com}
{h**tonsfeetred.info, a**etix.com}	{5**n7.com}
{l**tds.net}	{a**rackingout.com}
{h**tinlethemsed.info, a**etix.com}	{5**n7.com}
{s**vanus-phe.com}	{m**rosoft.com-repair-windows-system.live}
{s**vanus-phe.com, t**cking.marketing}	{m**rosoft.com-repair-windows-system.live}
{h**tinlethemsed.info, t**cking.marketing}	{m**rosoft.com-repair-windows-system.live}
{e**k.club, t**cking.marketing}	{m**rosoft.com-repair-windows-system.live}
{h**tinlethemsed.info, s**vanus-phe.com, t**cking.marketing}	{m**rosoft.com-repair-windows-system.live}
{e**k.club, s**vanus-phe.com, t**cking.marketing}	{m**rosoft.com-repair-windows-system.live}

表 6 抽出ドメイン名と VirusTotal の判定結果

頻出ドメイン名	悪性判定エンジン数 /判定エンジン総数
5**n7.com	2/80
e**k.club	1/80
h**tinlethemsed.info	1/71
h**tonsfeetred.info	4/71
i**brt.com	1/79
m**rosoft.com-repair-windows-system.live	1/69
p**checke.club	2/70

表 7 抽出ドメイン名と公開レポートでの報告有無

頻出ドメイン名	報告サイト	報告概要
a**etix.com	MALWEARETIPS 他	redirect adware
a**rackingout.com	MALWEARETIPS 他	redirect adware
l**tds.net	MALWEARETIPS 他	redirect adware
s**vanus-phe.com	TROJAN KILLER 他	redirect adware
t**cking.marketing	ANY.RUN	malicious activity

$\sigma_{st}(v)$としたとき，

$$CB(v) = \sum_{v \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

で表される指標である．ただし，頂点ペア$(s,\ t)$の最短路を頂点vが通らないとき$\sigma_{st}(v) = 0$とする[16-18]．集約サイトとなる複数のリダイレクトチェーンに共通に存在するドメイン名を，媒介中心性により検出する．

2019 年 2 月 3 日のデータセットから，4.2 節の手法で抽出した悪性リダイレクトチェーン候補に対してアルゴリズムを適用し抽出したドメイン名を表 8，表 9 に示す．表中の上から指標が大きい順に記載している．次数中心性の抽出条件は，起点となる頂点を抽出するために，出次数が 2 以上，入次数が 0 という条件としている．媒介中心性の抽出条件は，指標が 2 以上且つ複数の経路で媒介されている頂点を抽出している．抽出した結果，11 ドメイン中の 7 ドメインが悪性判定された．一方で，4 ドメインは悪性判定されなかったため，Web 上で一般公開されている公開レポートを調査した結果，4 ドメイン全てがリダイレクトアドウェアなどの悪性サイトとして報告されており，悪性のリダイレクトであることが確認できた．

表 8 次数中心性の適用結果

抽出ドメイン名	悪性判定エンジン数 /判定エンジン総数
h**tinlethemsed.info	1/71
h**tonsfeetred.info	4/71
p**checke.club	2/70

表 9 媒介中心性の適用結果と悪性判定

抽出ドメイン名	悪性判定方法	VirusTotal:悪性判定数 公開レポート:サイト名
5**n7.com	VirusTotal 判定	2/80
t**cking.marketing	公開レポート有無	ANY.RUN
a**etix.com	公開レポート有無	MALWEARETIPS 他
t**ilter.info	VirusTotal 判定	1/79
s**vanus-phe.com	公開レポート有無	TROJAN KILLER 他
e**k.club	VirusTotal 判定	1/80
i**brt.com	VirusTotal 判定	1/79
l**tds.net	公開レポート有無	MALWEARETIPS 他

Webアクセス順序

アクセス1	***p**checke.club***	***i**brt.com***	l**al-finders.com	s**rter-tech.net
アクセス2	***h**tonsfeetred.info***	c**ckredirection.com	s**rtiyada.com	
アクセス3	v**roz.xyz	***e**ionale.info***	t**thandarinhec.info	
アクセス4	***h**tonsfeetred.info***	***t**ilter.info***	c**ckssp.pro	y**rprivacy.icu
アクセス5	***h**tinlethemsed.info***	a**peratorx.com	***s**at0rx99.com***	

図 6 リダイレクトチェーンの悪性判定方法の例
(***Bold Italic*** : VirusTotal 悪性判定ドメイン名)

6. 評価

本章では，評価用データセットの作成方法，評価方法，及び評価結果と課題について述べる．

6.1 評価用データセットの作成方法

評価用データセットとして，評価の母数となるリダイレクトチェーンを実証実験のデータセットから抽出する．評価用データセットは，一定時間内にリダイレクトによるページ遷移を行ったデータとする．図 7 に，評価用データセットの抽出フローを示す．まず，候補となるリダイレクトチェーンを 3.2 節の方法で抽出する．次に，候補となるリダイレクトチェーンから，リダイレクトの経過時間が 5 秒以内，リダイレクトの回数が 2 回以上のデータを抽出し，図 4 の構成で保存し評価用データセットとする．

6.2 評価方法

評価は，5 章で示した各アルゴリズムで導出したドメイン名が，評価用データセットから何件検出されたかで行っ

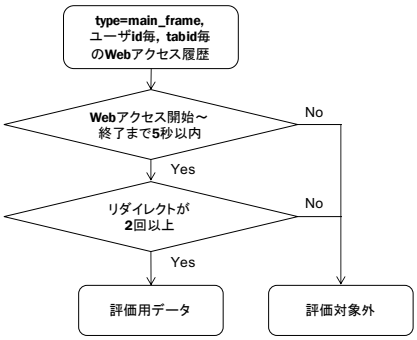

図 7 評価用データセットの抽出フロー

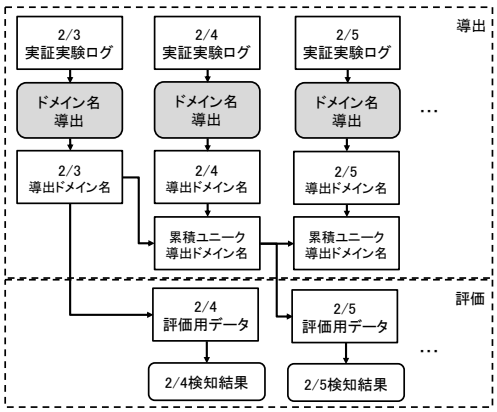

図 8 評価方法

た．なお検出数は，リダイレクトチェーン内に 1 件でも導出ドメイン名が存在した場合に検出とし，該当リダイレクトチェーンの件数をカウントしている．本稿では，Web アクセスの良性・悪性判定を，1 回の Web アクセスに対してではなく，リダイレクトチェーン全体として判定する．判定方法の概念図を図 6 に示す．リダイレクトチェーン上で，良性ドメイン名と悪性ドメイン名が共起する場合があるため，リダイレクトチェーン上にひとつでも悪性判定ドメイン名が存在する場合，悪性リダイレクトチェーンと判定する．図 6 の例では，アクセス 1-5 の全てのリダイレクトチェーンにおいて悪性判定ドメインが存在しており悪性リダイレクトチェーンと判定する．

評価方法を図 8 に示す．導出ドメイン名は，過去分の導出ドメイン名を加え累積させている．件数は，ドメイン名が同じものは除外したユニーク件数である．また，本評価実験では，2019 年 2 月 3 日から 2019 年 2 月 10 日の期間を評価期間として設定した．設定した期間のアクティブユーザ数は 900 名前後である．

6.3 評価結果

導出ドメイン名を用いた実験結果を表 10 に示す．表 10 の①は各適用アルゴリズム毎の導出ドメイン名数，②は累積ユニーク導出ドメイン名数，③は評価用データ（リダイレクトチェーン）数，④は Apriori アルゴリズムによる導出ドメイン名による検出数と検出率，⑤は SPADE アルゴリズムによる導出ドメイン名による検出数と検出率，⑥は中心性アルゴリズムによる導出ドメイン名（次数中心性と媒介中心性で抽出したドメイン名のユニーク）による検出数と検出率を示している．④-⑥の検出率はリダイレクトチェーンに VirusTotal で悪性判定されたドメイン名が存在すれば検出と判定している．

評価実験の結果，中心性アルゴリズムは，アソシエーション分析による手法（④，⑤）を検出率のうえで常に上回っており，より効率的に悪性リダイレクトチェーンを検出できることが分かった．また，中心性アルゴリズムでは，より少ない導出ドメイン名数で，より多くの悪性リダイレクトチェーンを効率的に検出できることを確認した．これは，悪性リダイレクトチェーン上で重要な起点にあたるサーバのドメイン名と，中継のハブとなるサーバのドメイン名が，中心性アルゴリズムにより構造的に重要な頂点として抽出できたことに起因している．更に，評価期間中で検出率が最も低かった 2 月 9 日について，VirusTotal で検出されなかったリダイレクトチェーンに含まれるドメイン名を，各アルゴリズム毎に公開レポートの有無で確認した．中心性アルゴリズムでは，6 ドメイン（2/9 の⑥:悪性として検出されなかったリダイレクトチェーン 5 件中に含まれていたユニークドメイン）全てで悪性サイトとして報告されており，潜在的な悪性ドメイン名であることが確認できた．表 11 に，中心性アルゴリズムで検出できなかったリダイレクトチェーンに含まれていた全 6 ドメイン名を示す．一方，アソシエーション分析による手法（④，⑤）では，22 ドメイン（2/9 の④，⑤:悪性として検出されなかったリダイレクトチェーン 15 件中に含まれていたユニークドメイン）中の 50%程度しか公開レポートに報告がなかった．このことから，中心性アルゴリズムによる手法は，潜在的な悪性リダイレクトチェーンを含め，より正確且つ効率よく悪性リダイレクトチェーンを検出していることが分かった．

今後の課題は，僅かながら VirusTotal の悪性判定でも，公開レポートでも報告されないリダイレクトチェーンが検出された．検出例としては，国外のサーバを経由したコミックや音楽ファイルのダウンロードサイトがあった．このようなグレーゾーンに属するサイトの良性・悪性判定方法がひとつの課題である．また，本稿で使用したのは，Web サーバから 3XX 応答で発生するリダイレクトを追跡したデータセットである．この他 meta tag refresh や JavaScript での location 書き換え，ポップアップによるリダイレクトがあり，これらの分析も今後の課題としてあると考えている．

7. まとめ

本稿では，WarpDrive 実証実験により収集された Web ブラウジングログのデータセットからリダイレクトチェーンを抽出し構造分析することによって悪性リダイレクトチェ

表 10　実験結果

評価項目＼日付	2/3	2/4	2/5	2/6	2/7	2/8	2/9	2/10
①導出ドメイン名数 Apriori/SPADE/次数+媒介中心性	14/12/11	19/15/9	33/30/6	27/27/5	21/20/3	11/11/0	16/15/6	–
②累積ユニーク導出ドメイン名数 Apriori/SPADE/次数+媒介中心性	14/12/11	24/21/15	48/44/17	62/59/19	73/71/20	80/78/20	88/86/21	–
③評価用データ(リダイレクトチェーン)数	–	912	1043	992	854	751	857	945
④Apriori 悪性リダイレクトチェーン数 /検出リダイレクトチェーン数(VirusTotal)	–	35/35 (100)*	51/52 (98)	37/46 (80)	41/53 (77)	22/29 (75)	32/47 (68)	78/87 (89)
⑤SPADE 悪性リダイレクトチェーン数 /検出リダイレクトチェーン数(VirusTotal)	–	35/35 (100)	48/48 (100)	37/43 (86)	39/49 (79)	22/29 (75)	31/46 (67)	78/87 (89)
⑥中心性 悪性リダイレクトチェーン数 /検出リダイレクトチェーン数(VirusTotal)	–	37/37 (100)	51/52 (98)	37/39 (94)	39/45 (86)	14/15 (93)	18/23 (78)	71/72 (98)

*④-⑥の括弧内は検出率%（小数点以下切り捨て）.

表 11 検出ドメイン名の公開レポートでの報告有無

検出ドメイン名	報告サイト	報告概要
e**srv.com	MALWEARETIPS 他	redirect adware
i**oxpush.com	MALWEARETIPS 他	pop-up ads
p**ads.net	MALWAREBYTES[19]他	PUPs*
p**u5l.com	MALWEARETIPS 他	redirect adware
t**cking.marketing	ANY.RUN	malicious activity
v**blife-4.co	MALWAREBYTES 他	hijacks

* Potentially Unwanted Programs

ーンに固有の特徴を明らかにした．そして，悪性リダイレクトチェーンを特定するための新しい方式を提案した．評価用データセットを用いて提案方式を評価した結果，悪性サイトへのリダイレクトを正確に特定する上で非常に効果的であることを実証した．

謝辞　本研究成果は，国立研究開発法人情報通信研究機構（NICT）の委託研究「Web 媒介型攻撃対策技術の実用化に向けた研究開発」により得られたものです．

参考文献

[1] Nelms, T., Perdisci, R., Antonakakis, M., Ahamad, M.. “WebWitness: investigating, categorizing, and mitigating malware download paths,” USENIX Security Symposium. 2015, p. 1025-1040.

[2] 嶌田一郎, 太田敏史, 白石訓裕, 中嶋淳, 田中翔真, 山田明, 高橋健志.“リダイレクトの追跡による悪性 Web ページアクセス事例分析,” 情報処理学会研究報告, March 2020.

[3] Stringhini, G., Kruegel, C., and Vigna, G.. “Shady Paths: Leveraging Surfing Crowds to Detect Malicious Web Pages Categories and Subject Descriptors,” ACM Conference on Computer and Communications Security (CCS), November 2013.

[4] Mekky, H., Torres, R., Zhang, Z.L., Saha, S. and Nucci, A.. “Detecting malicious http redirections using trees of user browsing activity,” IEEE International Conference on Computer Communications (INFOCOM), April 2014.

[5] Zhang, J., Seifert, C., Stokes, J.W., and Lee, W.. “Arrow: Generating signatures to detect drive-by downloads,” World Wide Web Conference (WWW), April 2011.

[6] Lu, L., Perdisci, R., and Lee, W.. “Surf: Detecting and measuring search poisoning categories and subject descriptors,” ACM SIGSAC Conference on Computer and Communications Security (CCS), October 2011.

[7] Lee, S. and Kim, J.. “Warningbird: Detecting suspicious urls in twitter stream,” IEEE Transactions on Dependable and Secure Computing, vol.10, no.3, January 2013.

[8] 山田明, 笠間貴弘, 井上大介. Web 媒介型攻撃対策「WarpDrive」の取組 ユーザ参加型による Web 攻撃対策の実現に向けて. NICT NEWS. 2018, Vol. 472, No.6, p. 10-11.

[9] “Top site - Alexa”. https://www.alexa.com/topsites, (参照 2020-08-06).

[10] Agrawal, R. and Srikant, R.. "Fast Algorithms for Mining Association Rules", Proceedings of the 20th VLDB Conference Santiago, 1994.

[11] Srikant, R. and Agrawal, R.. “Mining Sequential Patterns:Generalizations and Performance Improvements”, Proceedings of the 5th International Conference on Extending Database Technology, 1996

[12] “VirusTotal: VirusTotal”. https://www.virustotal.com/, (参照及び調査 2020-08-06).

[13] “MalwareTips Community”, https://malwaretips.com/, (参照 2020-08-07).

[14] “ANY.RUN – interactive malware hunting service”, https://any.run/, (参照 2020-08-07).

[15] “TROJAN KILLER”, https://trojan-killer.net/, (参照 2020-08-07).

[16] Anthonisse, J. M.. “The Rush In A Directed Graph,” Technical Report BN 9/71, Stichting Mathematisch Centrum, 1971.

[17] Freeman, L. C.. “A Set of Measures of Centrality Based on Betweenness,” Sociometry, 1977.

[18] Hayashi, T., Akiba. T., Yoshida, Y.. “Fully Dynamic Betweenness Centrality Maintenance on Massive Networks,” International Conference on Very Large Data Bases, 2016

[19] “MALWAREBYTES”, https://blog.malwarebytes.com/, (参照 2020-08-07).