Drive-by-Download 攻撃における悪性 PDF の特徴に関する考察

今野 由也† 角田 裕‡

†東北工業大学大学院 工学研究科 通信工学専攻
‡東北工業大学 工学部 情報通信学科
982-8577 宮城県仙台市太白区香澄町 35-1
m142803@tohtech.ac.jp tsuno@m.ieice.org

あらまし Drive-by-Download (DbD) 攻撃は，Web サイトを改ざんし，そのサイトにアクセスしたユーザに対し不正なコードを埋め込んだ悪性 PDF ファイルなどを自動的にダウンロードさせる攻撃である．悪性 PDF ファイルについては内部の JavaScript やファイル構造などの解析が行われているが，ファイルサイズやファイルの構成要素数などに着目した量的な解析は行われていない．そこで本研究では D3M データセットを利用し，DbD 攻撃の悪性 PDF ファイルに関する量的な特徴を調査した．その結果，DbD 攻撃において利用される PDF ファイルの内部には表示用のデータが極端に少ないなど，通常の PDF ファイルや標的型攻撃などで利用される PDF ファイルとは異なる特徴を持つことがわかった．

A Study on Characteristics of PDF Files in Drive-by-Download Attacks

Yuya Konno† Hiroshi Tsunoda‡

†Graduate School of communication Engineering, Tohoku Institute of Technology
‡Faculty of Engineering, Tohoku Institute of Technology
35-1, Yagiyama Kasumi-cho, Taihaku-ku, Sendai, Miyagi, 982-8577, JAPAN
m142803@tohtec.ac.jp tsuno@m.ieice.org

Abstract In Drive-by-Download (DbD) attacks, the attacker falsifies a web site. Users who access the falsified site are forced to automatically download malicious contents such as PDF files including malicious codes. For such malicious PDF files, analysis focused on JavaScript included in files and PDF structure have been conducted. However, to the best of authors' knowledge, any analysis regarding the quantitative features of PDF files, such as file size and the number of objects in a file, has not been made. In this research, we analyze quantitative features of malicious PDF files using D3M dataset and compare it with that of normal PDF files.

1 はじめに

Drive-by-Download (DbD) 攻撃は，Web サイトを改ざんすることでそのサイトにアクセスしたユーザを悪性サイトにリダイレクトし，自動的にマルウェアをダウンロードさせる攻撃である．近年では，海外の Web サイトのみならず日本の Web サイトも悪性サイトとなっていることが確認

されている. Tokyo SOC 分析レポート [1] によれば, 2013 年下半期 DbD 攻撃の検知数は 2012 年下半期の 956 件と比較して約 2 倍の 1922 件と増加している. 被害者は Web サイトを閲覧するだけでマルウェアに感染するため, 被害を受けたことに気づきにくく, 多くのユーザが閲覧する Web サイトが改ざんされた場合には被害の広がりが大きい. 従って, DbD 攻撃の手口の分析や早期検知などの対策が急務であり DbD 攻撃に関する様々な研究が行われている [2] [3] [6] [7] .

ユーザが DbD 攻撃でダウンロードさせられるマルウェアは, ブラウザ自身, Adobe Reader, Adobe Flash Player , Java Runtime Environment (JRE)などの脆弱性を利用する. 2010~2013 年の Tokyo SOC 分析レポート [1] より, これらの中でも, JRE の脆弱性を狙う JAR ファイル, Adobe Reader の脆弱性を狙う Portable Document Format (PDF) ファイルが特に利用される割合が高いことが知られている. また, 特に PDF ファイル[5] は動画や音楽などを内部に埋め込むことができマルチメディア性に長ける一方, 埋め込んだデータを外部プログラムで実行する機能を有していることから, マルウェアの感染経路として広く利用されている.

本研究では, DbD 攻撃に利用される悪性 PDF ファイルの特徴を分析する. これまで, 悪性 PDF に関しては内部に埋め込まれた JavaScript やシェルコードの解析が行われている. また, PDF のファイル構造に着目した解析が行われている. しかし, PDF ファイルのサイズや内部の構成要素数など最も単純な量的特徴に関する解析は行われていない. そこで本稿では, DbD 攻撃の通信パケットからなる D3M データセットを利用し悪性 PDF の量的特徴を分析する.

以下, 第２章では関連研究と本研究の目的について述べ, 第３章では PDF のファイル構造と DbD 攻撃で利用される悪性 PDF の特徴について述べる. 第４章では量的特徴について調査し, 第５章で調査より得られた結果を考察する. 第６章は本研究で得られた特徴についてまとめ, 今後の課題について述べる.

2 悪性 PDF の特徴解析

本章では, PDF ファイルの構造について説明した後, 悪性 PDF の解析に関する関連研究について述べる.

2.1 PDF ファイルの構造[8]

PDF ファイルは図 1に示すようにヘッダ, 本体, 相互参照テーブル, トレーラの4つのセクションで構成される.

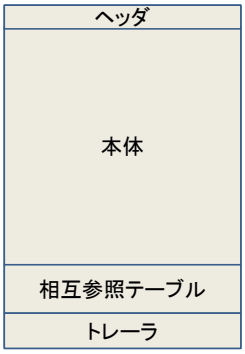

図 1 PDF ファイルの構造

ヘッダはファイルが PDF であることを示すとともに PDF のバージョン番号を指定するセクションである.

本体はオブジェクトの集合である. オブジェクトはドキュメントの構成要素であり, 整数や実数および文字列などの基本オブジェクトと複合オブジェクトに分類される. 複合オブジェクトは配列や辞書など他のオブジェクト格納するためのものと, バイナリデータを格納するストリームがサポートされている. ストリームには PDF ファイルの各ページのコンテンツや画像データなどが格納される. PDF 内部で使用されるフォント情報や JavaScript のコードは辞書オブジェクトとして格納されている.

相互参照テーブルは本体にある各オブジェクトのバイトオフセットのテーブルである. このテーブルを調査することで, ファイル内に存在する

オブジェクトの総数や位置を知ることができる.

トレーラには相互参照テーブルのバイトオフセットとドキュメントカタログへの参照情報が記載されている. ドキュメントカタログはオブジェクトのツリー構造の頂点のオブジェクトであり, すべてのオブジェクトはドキュメントカタログからの参照を辿ってアクセスできるようになっている.

2.2 内部の JavaScript の動的解析

悪性 PDF の内部に JavaScript が埋め込まれ利用されている場合, その JavaScript の多くは難読化されている. そのため, 悪性 PDF から JavaScript のコードを抽出しても処理内容や悪意の有無を判別することが難しい. また, 多様な難読化手法の存在により, 単純なシグネチャマッチングによる悪性 PDF の検出が困難になっている.

この問題に対応するため, 文献[9] では JavaScript をエミュレートして動的解析を行うシステムを提案している. 提案システムでは, 動的解析により可読化した JavaScript を導出し, 利用している脆弱性の分析などの悪性 PDF の挙動の解析を実現している. また, 可読化された JavaScript 内部にあるシェルコードの自動解析手法についても提案している.

2.3 ファイル構造の解析

文献[10] では, 標的型攻撃で利用される悪性 PDF ファイルの構造が通常の PDF ファイルとは異なる場合があることを指摘している. これは, PDF ファイルを閲覧ソフトで表示した場合に誤動作や文字化けを起こさぬよう, 閲覧ソフトが通常読み込まない場所に実行ファイルが格納されるためである. [10] では, 分類不能なセクションの存在や, 参照されないオブジェクトの存在などに着目してファイルの構造を検査することで, 悪性 PDF ファイルを検知する手法を提案している. この検知を回避するために攻撃者が悪性 PDF のファイル構造を変更すると, 閲覧ソフトが誤動作する可能性があることから, ファイル構造は有用な特徴であると言える.

3 PDF ファイルの量的な特徴に着目した解析

本節では DbD 攻撃と標的型攻撃で利用される悪性 PDF ファイルの特徴について, 攻撃の仕組みの違いから考察し, 本研究で実施する PDF ファイルの量的な特徴に着目した解析の概要を説明する.

標的型攻撃では, 被害者自身が閲覧ソフトによって悪性 PDF ファイルを開いて閲覧することで閲覧ソフトの脆弱性を攻撃する内部の不正なコードが実行される. このとき, 閲覧した被害者が不審感を抱かないよう, ダミー表示用の文書データを含めたり, 実行ファイルを閲覧ソフトが読み込まない部分に埋め込んだりするなどの対策が行われる[10] .

しかし, DbD 攻撃の場合, ダウンロードさせられた悪性 PDF ファイルはウェブブラウザの機能により被害者の意図しないところで自動的に閲覧ソフトにより開かれ, 内部の不正なコードが実行される. すなわち, 被害者の目を意識する必要がないため, 標的型攻撃の悪性 PDF ファイルや一般の PDF ファイルとは異なる特徴を持つと考えられる.

例えば, 文献[10] では悪性 PDF ファイルの平均サイズは 351.2KB とされているが, 本研究で D3M データセットを利用して DbD 攻撃の悪性 PDF ファイルを調査したところ, その平均サイズは 28.1KB しかなかった.

これは, DbD 攻撃の悪性 PDF ファイルは PDF の仕様を満足していれば良く, 表示用のデータは必ずしも重要ではないことに起因していると考えられる. また, サイズだけでなく, ファイル内部の構成要素数などにも違いがあると予想される. そこで本研究では, このような PDF ファイルの量的な特徴に着目して DbD 攻撃の PDF ファイルを分析した. 具体的な以下の項目について PDF ファイルを調査した.

- ページ数
- ファイルサイズ
- オブジェクト数
- フォント情報が記載されたオブジェクト数

次章では PDF ファイルのサイズや内部の構成要素数に着目し必要最小限の構成で動作している PDF が存在するか調査し，構成要素数に与える影響について分析を行う.

4 量的な特徴の調査

調査で利用したデータについて述べ，前章で挙げた PDF ファイルサイズや内部の構成要素数について述べる.

4.1 調査対象データ

調査で利用する悪性 PDF はマルウェア対策研究人材育成ワークショップ[11] 提供の D3M データセットに含まれる PDF を利用した. D3M データセットは高対話型の Web クライアント型ハニーポットによって悪性 URL を巡回して採取した DbD 攻撃に関する通信データであり，次のデータが含まれている.

- 攻撃通信データ
- マルウェア
- マルウェア通信データ

本研究では D3M データセット(2010~2014)の攻撃通信データに含まれる 98 個の PDF ファイルを悪性 PDF として調査対象とした.

また， 比較対象の通常 PDF には google 検索により得られた 510 個の PDF ファイルを使用した.

4.2 ページ数

悪性 PDF が必要最小限の構成となっている場合ページ数は少なくなると予想される.

悪性と通常それぞれの PDF ページ数について調査し，結果を図 2, 図 3に示す.

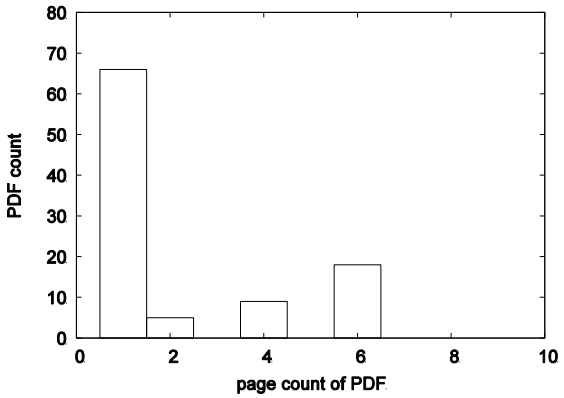

図 2 悪性 PDF のページ数

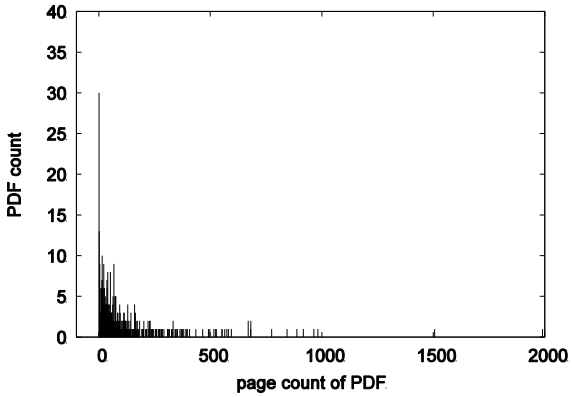

図 3 通常 PDF のページ数

通常の PDF ファイルの場合，ページ数は様々であるが，悪性 PDF ファイルのページ数は 1, 2, 4, 6 のいずれかのみであり，全体の約 67%の 66 個のファイルでページ数が1ページであった.

実際に PDF ファイルを閲覧ソフトで表示させ，ファイルのプロパティと表示内容を目視で確認したところ，表 1に示す特徴が見られた.

表 1 悪性 PDF のページ数による特徴の違い

ページ数	特徴
1	・図もしくは文字のみで構成 ・文字の分量は多くても 3/4 ページ
2,4,6	・Version が 1.3.3 もしくは 1.7.3 ・文字のみで構成

ページ数1枚文字のみで構成されている悪性

PDF の中には，”HelloWorld!”のように明らかに必要最小限のダミーデータが表示されるものも含まれていた．

また，ページ数が複数の22個の悪性PDFファイルは，全てのページが文字のみで構成されページ左上にカラーの四角形が描画されているという共通した特徴が見られた．このことから，これらの悪性PDFは同一の攻撃者や同一のツールによって作成された可能性があると考えられる．

4.3 ファイルサイズ

3. で述べた通り，DbD 攻撃の悪性 PDF ファイルは標的型攻撃のそれと比較してファイルサイズが小さい．ここでは，通常の PDF ファイルと比較してもファイルサイズが小さいことを示す．DbD 攻撃の悪性 PDF ファイルと通常 PDF ファイルのサイズのヒストグラムを図 4，図 5にそれぞれ示す．

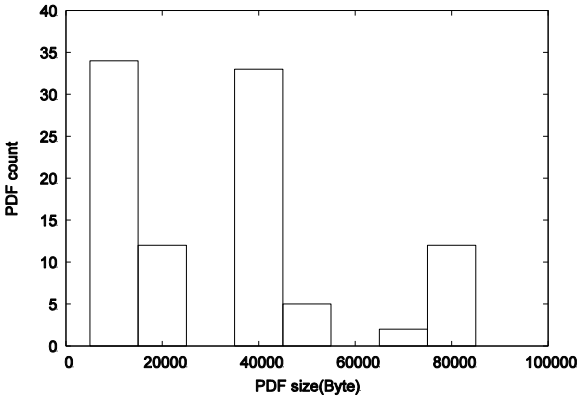

図 4 悪性 PDF のファイルサイズ

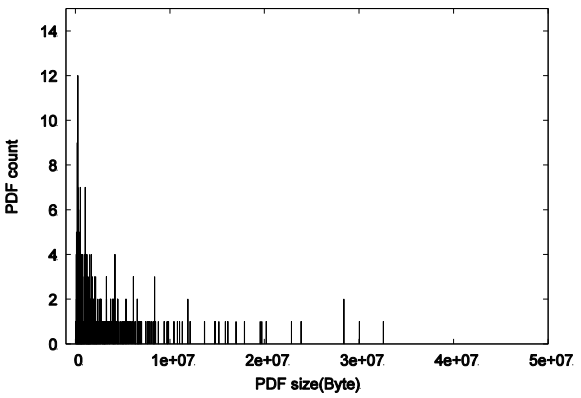

図 5 通常 PDF のファイルサイズ

図 4，図 5から，悪性 PDF ファイルは通常 PDF ファイルと比較して明らかに小さいことが確認できた．

4.4 オブジェクト数

オブジェクトは PDF ファイルの構成要素であるため，オブジェクト数が PDF ファイルの特徴を表すと考えられる．そこで，PDF に含まれているオブジェクト数を調査した．図 6に悪性 PDF ファイルと通常 PDF ファイルにおける，オブジェクト数とファイルサイズの関係を示す．

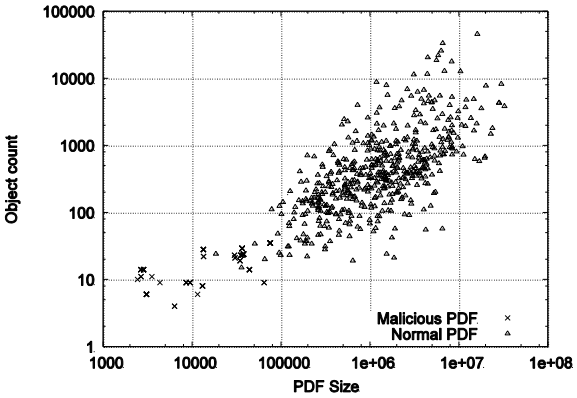

図 6 PDF に含まれるオブジェクト数

図 6より，悪性 PDF，通常 PDF 共にファイルサイズとオブジェクト数に相関は見られるものの明らかな違いは発見されなかった．

4.5 フォントオブジェクト数

DbD 攻撃の悪性 PDF ファイルが PDF のフォーマットを満たす必要最小限な要素で構成されている場合，画面に表示する文字数は少なくなりフォントに関連するオブジェクト数も少ないと考えられる．そこで，全オブジェクト数に対してフォントに関連する数の関係を調査し 図 7，図 8に示した．ただし，図 8はフォントに関連するオブジェクトの数が極端に多く 500 以上ある PDF ファイルについては省いている．

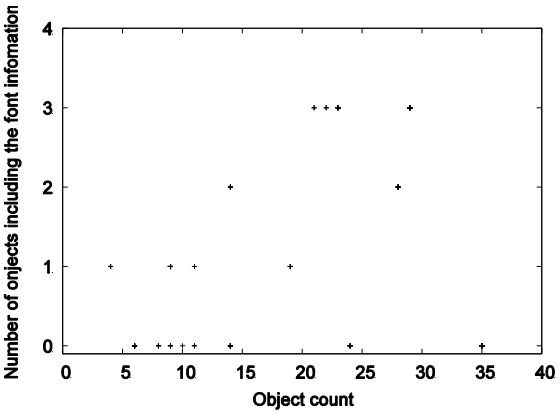

図 7 悪性 PDF のフォントオブジェクトとオブジェクト数

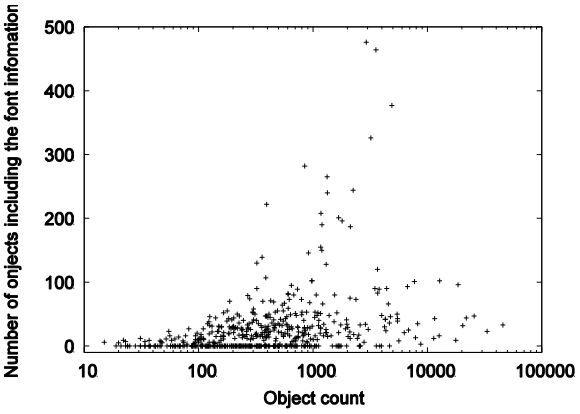

図 8 通常 PDF のフォントオブジェクトとオブジェクト数

悪性 PDF, 通常 PDF の双方においてフォント関連のオブジェクトの数が 0～3 個しかないファイルがあることがわかる. これは悪性 PDF ファイルについては表示用のデータがほとんど存在しないことによるものと言える. 一方で, 通常 PDF については図が大半を占めているファイルが存在したことによるものと考えられる.

5 考察

3.と4.の調査より DbD 攻撃の悪性 PDF のファイルサイズは小さくページ数が少ないという特徴が得られた. また, オブジェクトの総数やフォント関連のオブジェクト数も少なく, DbD 攻撃の悪性 PDF は表示するデータは必要最小限となっていることがわかった.

悪性 PDF のファイルサイズが小さい理由として, ダウンロードにかかる時間を短縮したいという攻撃者の意図があると考えられる. これは, 不審なダウンロードが行われたことに気づきにくくするためと考えられる.

ただし, 攻撃者は悪性 PDF ファイルのサイズに任意のデータを加えることで, 通常の PDF と同程度のファイルサイズにできるため, 量的な特徴だけでは悪性 PDF ファイルの判別は困難である. ただし, この場合 PDF のダウンロード完了までの時間が長くなると考えられるため, 意図しない PDF のダウンロードが完了するまでの時間を長引かせることにつながるといえる.

6 まとめと今後の課題

本稿では Drive-by-Download 攻撃で利用される悪性 PDF に関して量的な特徴を調査し PDF のファイルサイズやページ数などに関して特徴が現れることが分かった.

今後は D3M データセット以外のデータにも適応し, Drive-by-Download 攻撃のみで得られる特徴であるか再確認が必要である. また, 今回調査を行った量的な特徴は基本的にダウンロード後にのみ知りえる情報である. よって, 量的な特徴を検知に利用する場合ダウンロードしたファイルを実行しないように保存する仕組みが必要である.

参考文献

[1] IBM , Tokyo SOC 分析レポート, https://www-304.ibm.com/connections/blogs/tokyo-soc/?lang=ja

[2] 金子博一, 松木隆宏, 新井悠, “通信トラヒックの分析による Gumblar 感染 PC の可視化”, 電子情報通信学会研究会報告, ICSS-79, pp1~6,Jun, 2010

[3] M.Cova , C.Kruegel , G.Vigna , “Detection and Analysis of Drive-by-Download Attacks and Malicious JavaScript Code” , In International Conference on World Wide Web(WWW), pp.281~290, 2010

[4] Adobe , Document management –

Portable document format - part1 : PDF 1.7 first editon , http://wwwimages.adobe.com/content/dam/Adobe/en/devnet/pdf/pdfs/PDF32000_2008.pdf

[5] Nedim,S. Pavel,L. , "Detection of Malicious PDF Files Based on Hierarchical Document Structure" , Network and Distributed System Security Symposium, 2013

[6] Z.Shafiq , S.Khayam , M.Farooq , "Embedded Malware Detection using Markov n-grams", In Detection of Instrusions and Malware & Vulnerability Assessment(DIMVA), pp.88~107, 2007

[7] D.Cacali , M.Cova , G.Vigna , C.Kruegel, "Prophiler:A Fast Filter for the Large-Scale Detection of Malicious Web Pages" , In International Conference on World Wide Web(WWW), pp.197~206, 201

[8] John Whitington 著 , 村 上 雅 章 訳,"PDF 構造解説", オライリージャパン

[9] 神薗雅紀, 西田雅太, 星澤裕二, "動的解析を利用した PDF マルウェア解析システムの実装と評価" , 電子情報通信学会研究 会 報 告 , ICSS-Vol.110 , No.475 , pp.47-52, 18-3, 2011.

[10] 大坪雄平, 三村守, 田中英彦, "PDF の構造検査による悪性 PDF ファイルの検知", コンピュータセキュリティシンポジウム 2013 論 文 集 , Vol. 2013 , No.4 , pp.649-656, 14-10, 2013.

[11] 秋山満昭, 神薗雅紀, 松木隆宏, 畑田充弘, "マルウェア対策のための研究用データセット～MWS Datasets 2014～," 情報処理学会 研究報告コンピュータセキュリティ（CSEC） Vol. 2014-CSEC-66, No. 19, pp. 1 - 7, 2014.

[12] AUN CONSULTING, INC. , "世界 40 の国と地域の検索エンジンシェア", https://www.auncon.co.jp/corporate/2012/1120.html