サンドボックス解析結果に基づく URLブラックリスト生成についての一検討

畑田 充弘、田中 恭之、稲積 孝紀
先端IPアーキテクチャセンタ セキュリティTU
NTTコミュニケーションズ株式会社

もくじ

もくじ

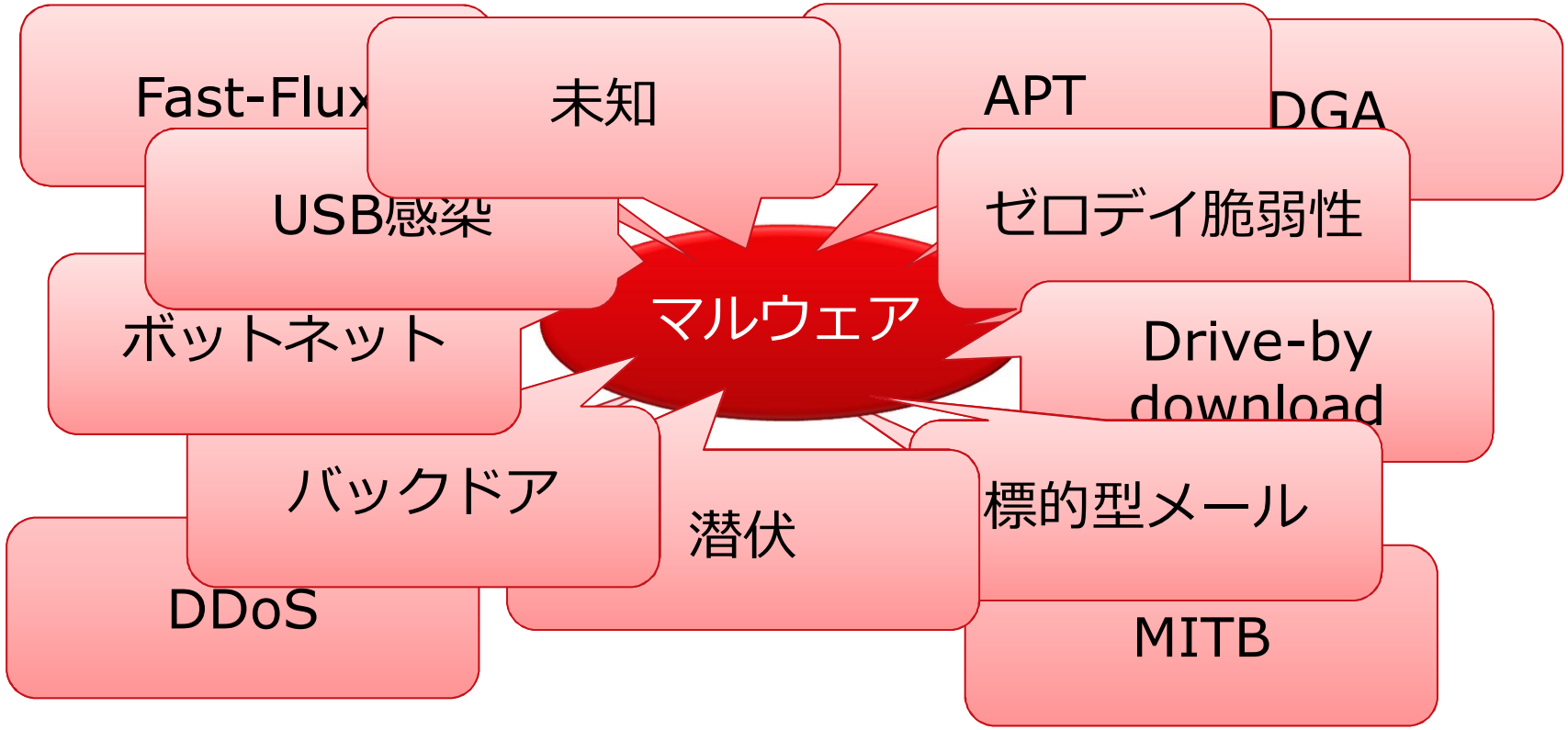
本当にそんなに精度高いの？どんなURL?

- AV Comparatives[1]
  - 主要なマルウェア対策ソフトが90%以上の検知率で悪意のあるWebサイトへのアクセスを検知・防御
  - テストデータは独自に収集したURL（実行ファイルへの直接リンクやDrive-by download含む）や手動で検索したURL，研究者から提供されたURL等

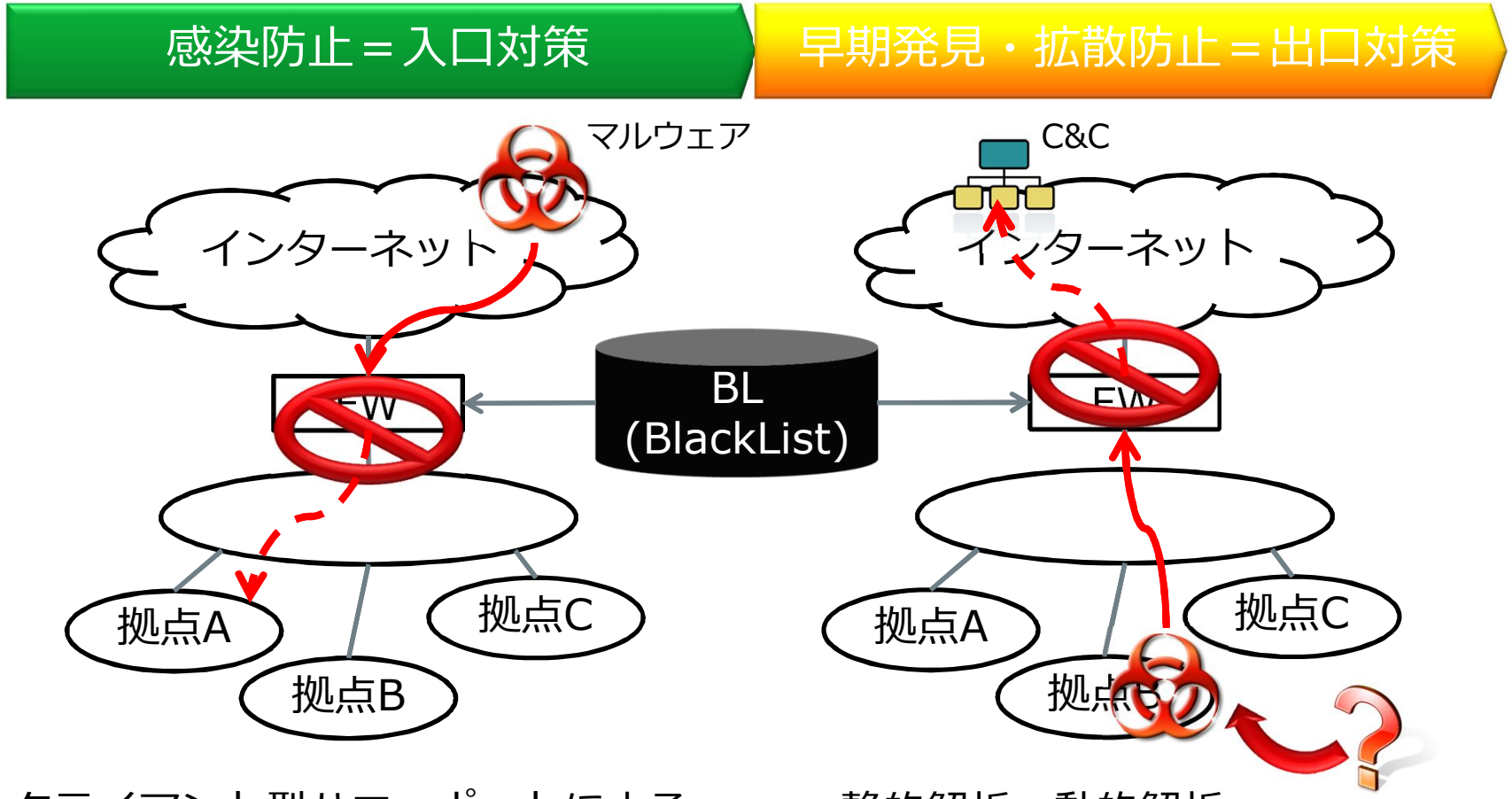
> We use our own crawling system to search continuously for malicious sites and extract malicious URLs (including spammed malicious links). We also search manually for malicious URLs. If our in-house crawler does not find enough valid malicious URLs on one day, we have contracted some external researchers to provide additional malicious URLs (initially for the exclusive use of AV-Comparatives) and look for additional (re)sources.

AV Comparatives: Whole Product Dynamic “Real-World” Protection Test – (March-June 2013), http://www.av-comparatives.org/wp-content/uploads/2013/07/avc_prot_2013a_en.pdfより引用

マルウェアを取り巻く環境



（一案）入口対策、出口対策のブラックリスト



クライアント型ハニーポットによる
悪性サイトの検知
→入口対策用BL

静的解析、動的解析
→出口対策用BL

研究の動機

マルウェアがアクセスするURL（出口対策用BL）

- インターネットの接続確認
- グローバルIPアドレスの確認
- C&C通信
- マルウェアのダウンロード
- 外部への情報送信

など

不正なアクセス先URLを区別する必要性

もくじ

関連研究

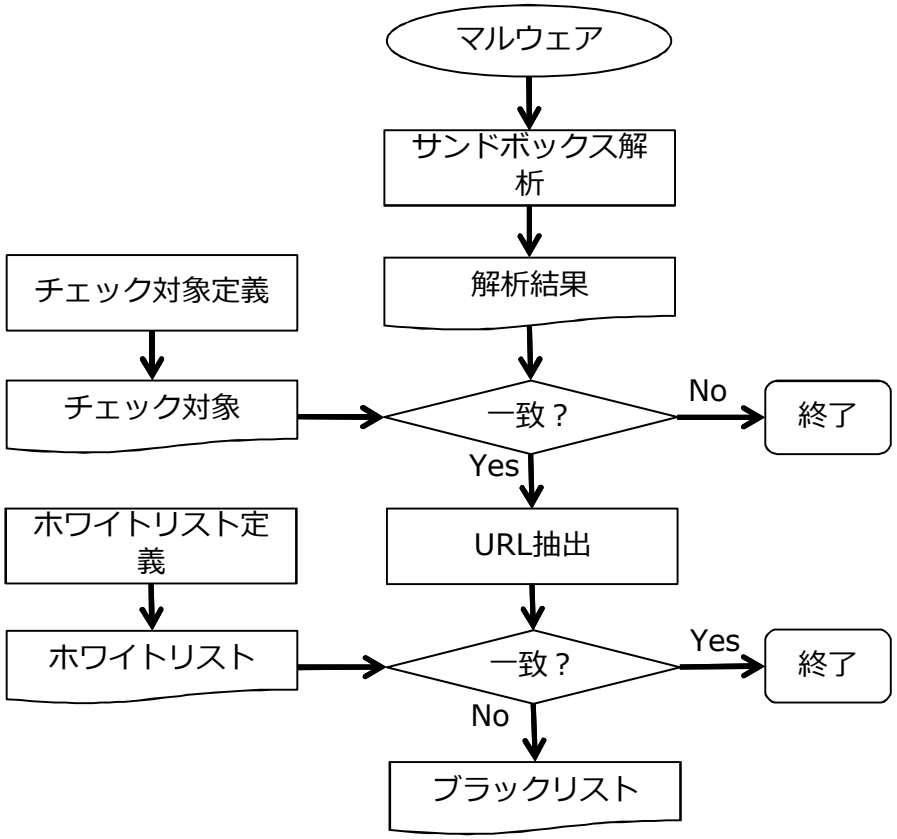
- トラフィックの特徴に基づく感染検知手法としてヘッダによるもの[6,7]、ペイロードによるもの[8]
  - 課題：厳密にURLを特定するものではない
- システムコールのビヘイビアグラフとデータフロー解析によって，外部への送信データや受信データに基づいてC&C通信を判定するもの[9]
- テイント解析により外部へ感染ホストのシステムやユーザ固有の情報などの送信先を判定するもの[10, 11]
  - 課題：ビヘイビアグラフの網羅性や、テイントの伝搬漏れ・誤伝搬などとともに、解析コストが大きい

サンドボックスを利用した簡易な方式

もくじ

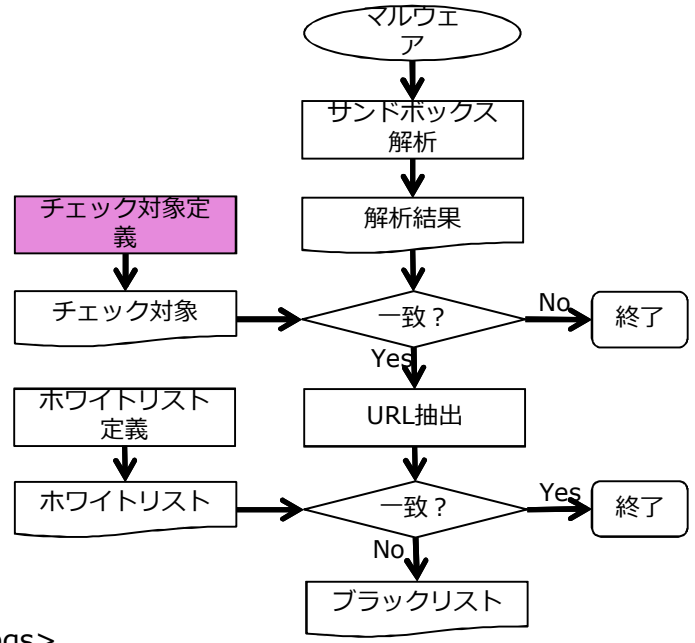
処理概要

- チェック対象定義
- サンドボックス解析
- URL抽出
- ホワイトリスト定義
- ブラックリスト



チェック対象定義

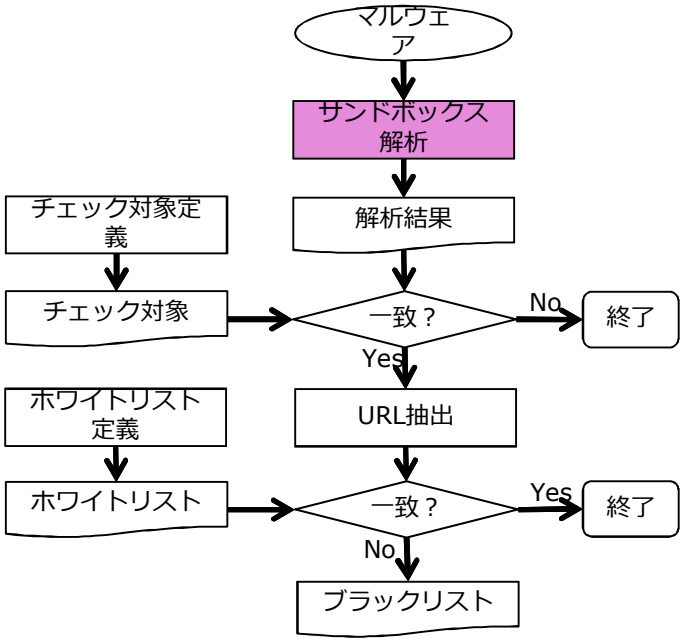
- 感染ホストの識別情報や、感染ホストを使用しているユーザやアプリケーションの情報を、C&Cサーバや他の外部サーバに送信するといったマルウェアの挙動に着目
  - OSのライセンス情報
  - ユーザ名
  - ブラウザのCookie
  - ブラウザの閲覧履歴
  - メールサーバ設定(FQDN, IPアドレ
    , ID/パスワード)
  - FTPサーバ設定（FQDN, IPアドレ
    , ID/パスワード）




```
<Regs>
<item>HKLM¥SOFTWARE¥MICROSOFT¥WINDOWS
NT¥CURRENTVERSION</item>
  <item>HKCU¥Software¥FTP
Explorer¥Profiles</item>
   :
</Regs>
<Files>

<item>C:¥Users¥user¥AppData¥Roaming¥Mozilla¥
Firefox¥Profiles¥</item>

<item>C:¥Users¥user¥AppData¥Roaming¥Microsoft
¥Windows¥Cookies¥Low¥user@yahoo[2].txt</item
>
   :
</Files>
```

サンドボックス解析

- MWS Datasets 2013のFFRI Dataset 2013[13]
  - Cuckoo Sandbox
  - マルウェア2,644検体の動的解析ログ
  - 解析対象1検体につき1ログファイル（JSON形式）
  - 解析対象ファイルや解析環境のメタ情報
  - APIコール
  - 挙動のサマリとしてのアクセスしたレジストリやファイル
  - ネットワークアクティビティ等



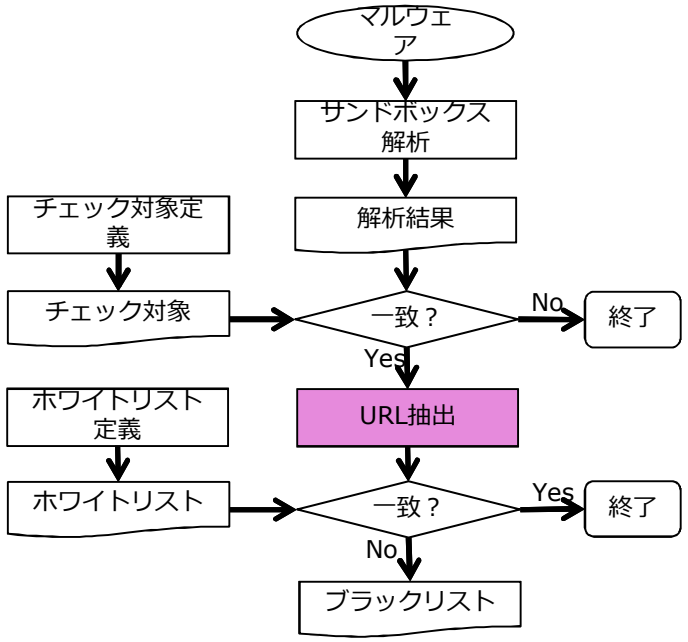
FFRI Dataset 2013の例（id=560）


```
{
   "info": {
      "category": "file",
      "started": "2013-04-09 04:44:53",
      "ended": "2013-04-09 04:47:15",
      "version": "0.5",
  : （省略）
 "category": "filesystem",
                  "status": "FAILURE",
                  "return": "0xc0000034",
                  "timestamp": "2013-03-28
11:05:10,828",
                  "thread_id": "1396",
                  "repeated": 0,
                  "api": "NtOpenFile",
                  "arguments": [
                     {
                        "name": "FileHandle",
                        "value": "0x00000000"
                     },
                     {
                        "name": "DesiredAccess",
                        "value": "0x001200a9”
  : （省略）
 "summary": {
         "files": [
            "C:¥¥WINDOWS¥¥system32¥¥msctfime.ime",

"C:¥¥DOCUME~1¥¥cuckoo1¥¥LOCALS~1¥¥Temp¥¥2632
645C17E1985396F0033D15EE253F.bin.2.Manifest",

"C:¥¥DOCUME~1¥¥cuckoo1¥¥LOCALS~1¥¥Temp¥¥2632
645C17E1985396F0033D15EE253F.bin.3.Manifest",
  : （省略）
```

```
:
 "http": [
         {
            "body": "",
            "uri":
"http://imp.xxxxxxxxxxxxxxxx.com/impression.do/?us
er_id=5bf96358-336b-4339-b511-
47279b470712&event=setup_run&spsource=engageBD
R_downloadmanager-US-
direct&subid=(null)&traffic_source=engageBDR&offer_id
=downloadmanager",
            "user-agent": "||Mozilla/5.0 (Windows NT
5.1) AppleWebKit/537.11 (KHTML, like Gecko)
Chrome/23.0.1271.97 Safari/537.11",
            "method": "GET",
            "host": "imp. xxxxxxxxxxxxxxxx.com",
            "version": "1.1",
            "path": "/impression.do/?user_id=5bf96358-
336b-4339-b511-
47279b470712&event=setup_run&spsource=engageBD
R_downloadmanager-US-
direct&subid=(null)&traffic_source=engageBDR&offer_id
=downloadmanager",
            "data": "GET
/impression.do/?user_id=5bf96358-336b-4339-b511-
47279b470712&event=setup_run&spsource=engageBD
R_downloadmanager-US-
direct&subid=(null)&traffic_source=engageBDR&offer_id
=do:
  : （省略）
```

URL抽出

- サンドボックス解析結果から
  - network ＞ http ＞ uri　を抽出
  - 条件
    - サンドボックス解析結果とチェック対象が一致（あるいは一致した件数がN件以上）
    - チェック対象の読み出し時刻以降のアクセスであること



URL抽出の注意事項

■hostとpathをそのままつなげたURL

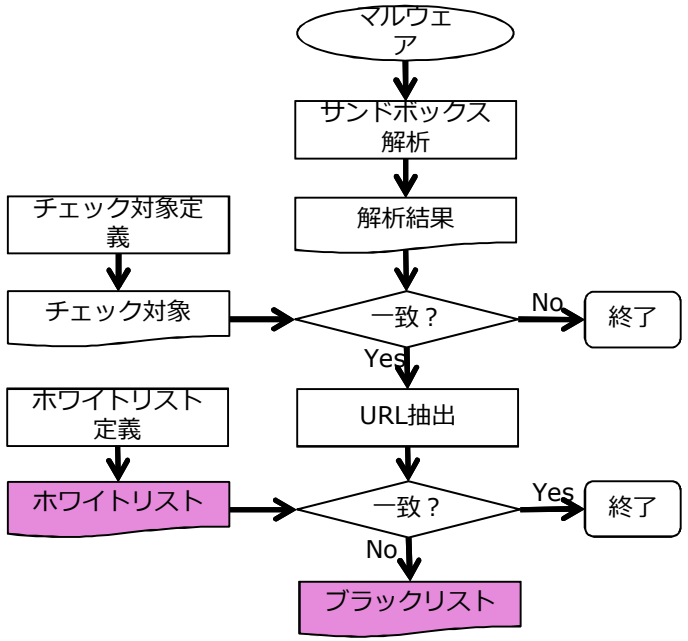
```
# id=1013より抜粋
"http": [
        {
           "body": "",
           "uri":
"http://app2.xxxxxxxx.com/http://app2.
xxxxxxxx.com/app.asp?prj=5&pid=wsk1&logdata=MacT
ryCnt:0&code=&ver=1.0.0.100&appcheck=1",
           "method": "GET",
           "host": "app2.winsoft3.com",
           "version": "1.1",
           "path": "http://app2.
xxxxxxxx.com/app.asp?prj=5&pid=wsk1&logdata=MacT
ryCnt:0&code=&ver=1.0.0.100&appcheck=1",
           "data": "GET http://app2.
xxxxxxxx.com/app.asp?prj=5&pid=wsk1&logdata=MacT
ryCnt:0&code=&ver=1.0.0.100&appcheck=1
HTTP/1.1¥r¥nHost: app2. xxxxxxxx.com¥r¥n¥r¥n",
           "port": 80
        },
```

■DNSクエリなし

```
# id=1027より抜粋
 "http": [
        {
           "body": "",
           "uri":
"http://xxxxxx.1ywsk79gm7g3iq9w.com/?cEI17q20=%9
6%CB%D2%D3%D6%D5%8F%94%AE%A9%D9%9Fgi
%9D%D3%98%A0f%CF%AA%9A%DD%94%98%A7%A
3%93u%82%94%9D%D3X%A7%E8%A2%E7%E5%CA
%C4T%A6%E2%DB%B0%A0nj%9D%93%A3%D5%9A
%A6%DB%9A%A4x%B3%95%DA%CC%A9%86hl%AAi
d%AB%AB%A3%A8%AB%B7_u%B2%A8%A6%A0xj%A
B%A3y%9Bk%AD%A6k%BA%60%9A%B5% (省略)
```

ホワイトリスト定義、ブラックリスト

- ホワイトリスト
  - インターネット接続確認
    - メジャーなWebサイト
    - http://www.google.com　等
  - グローバルIPアドレス確認のためのWebサイト
    - http://checkip.dyndns.org/ 等
- ブラックリスト
  - サンドボックス解析結果がチェック対象の情報と一致するものがあった場合に、当該マルウェアの通信先URLから、ホワイトリストに一致したURLを除外してブラックリストとする



もくじ

FFRI Dataset 2013

- 4件（id=1199, 1496, 1578, 2567）についてはJSON形式のファイルが途中で切れているため集計からは除外
- 2,640検体について

項目	件数
HTTP通信ありの検体	256
延べURL	1,628
ユニークURL	628
ユニークFQDN	147

  - 約9%の検体のみでHTTP通信
  - 延べURL数に対するユニークURL数は約38%
    - 同一のURLに複数回アクセス
    - 同一あるいはURLクエリパラメータが異なるURLへアクセス

FFRI Dataset 2013

- 課題
  - 1検体あたり90秒間の解析時間では不十分（短い?）
  - 解析環境を検知して動作を停止/変更する検体
  - 検体を取得してから動的解析までの時間経過によってアクセス可能な通信先が減少

⇒これらを解決することにより抽出可能なURL数を増やすことができる可能性

ブラックリスト生成例1

- チェック対象
  - ブラウザのCookieや履歴の情報をファイル読み出し
- URL抽出対象
  - id=1565など
- パス部が{3桁の数字}/{3桁の数字}.htmlというパターン
- 当該検体以外でもいくつかの検体で同様のパターンが確認

<チェック対象>
C:¥¥Documents and Settings¥¥cuckoo1¥¥Cookies¥¥index.dat
C:¥¥Documents and Settings¥¥cuckoo1¥¥Local Settings¥¥History¥¥History.IE5¥¥index.dat

<ブラックリスト>
http://xxxxxxx.net/826/759.html
http://sp3.xxxxxx.com/?dm=elacyts.net&acc=3F254F8E-C939-4DF2-84B2-CA2A97E466E5
http://xxxxxxx.net/501/751.html
http://xxxxxxxxxxx.com/497/873.html

ブラックリスト生成例2

- チェック対象
  - インストール済ソフトウェアに関するレジストリ読み出し
- URL抽出対象
  - id=1690
- ホスト名がIPアドレスのURL
- 80番ポートを使用せず、他のポートを指定したURL
- プロダクトキーの読み出しはFFRI Dataset 2013では見られなかった

<チェック対象>
HKEY_LOCAL_MACHINE¥¥SOFTWARE¥¥Microsoft¥¥Windows¥¥CurrentVersion¥¥Uninstall

<ブラックリスト>
http://x.xxx.175.164/content/offers/default.aspx

もくじ

課題（1/2）

- チェック対象のメンテナンス
  - OSのバージョンやインストールされているアプリケーションの種類によって、チェック対象とすべきファイルやレジストリが変わってくる
  - プロダクトキーや設定情報など何をチェック対象とすべきか
  - その利用価値や実際のマルウェアの挙動を分析しながらチェック対象をメンテナンスしていく必要性あり
- ホワイトリストのメンテナンス
  - 正常通信に紛れてマルウェアが通信を行うことを考えると、メジャーサイトであってもマルウェアが不正利用
  - マルウェアがインターネット接続確認を行う場合、メジャーサイトであるとは限らない
  - チェック対象のメンテナンスと同様に実際のマルウェアの挙動を分析しながらメンテナンスしていく必要性あり

課題（2/2）

- サンドボックス解析の詳細度
  - APIコールの時刻やプロセスツリーなど詳細なデータが利用できるが、その詳細度はサンドボックスに依存
  - サンドボックスのインターネットへの接続条件や、マルウェアが備える解析環境の回避技術に対してシステム固有の情報をランダム化する等の実現レベルの違い
- ブラックリストとしての確からしさ
  - 提案方式が、テイント解析等による従来手法と比較してどの程度悪性サイトを判定できているのか
  - ドメインやIPアドレスのレピュテーション情報などとの比較も一つの評価方法
- 攻撃耐性
  - マルウェア検体の役割分担
  - ある検体で情報収集、他の検体がそのファイルを読み出して外部へ送信

もくじ

まとめ

- マルウェアのサンドボックス解析結果をもとに、システム情報やユーザ情報に関するファイルやレジストリの読み取りを条件とした、簡易なURLブラックリスト生成方式を提案
- FFRI Dataset 2013を用いた事例調査の結果を示すとともに、提案方式およびデータセットの課題を考察
- 精査したチェック対象リストをもとに、テイント解析結果との比較など有効性の評価

参考文献

[1] AV Comparatives: Whole Product Dynamic “Real-World” Protection Test – (March-June 2013), http://www.av-comparatives.org/wp-content/uploads/2013/07/avc_prot_2013a_en.pdf (参照2013/08/19)

[2] IPA：「新しいタイプの攻撃」の対策に向けた設計・運用ガイド改訂第2版, http://www.ipa.go.jp/security/vuln/newattack.html (参照2013/08/19)

[3] Palo Alto Networks: URL Filtering, https://www.paloaltonetworks.com/products/features/url-filtering.html (参照2013/08/19)

[4] DigitalArts：i-Filter，http://www.daj.jp/bs/i-filter/ (参照2013/08/19)

[5] Akiyama, M., et al.: Design and Implementation of High Interaction Client Honeypot for Drive-by-download Attacks, IEICE Transactions on Communication, Vol.E93-B No.5 pp.1131-1139, May 2010.

[6] 川元，他：マルウェア感染検知のためのトラヒックデータにおけるペイロード情報の特徴量評価, MWS2011(2011年10月)

[7] 市野，他：トラヒックの時系列データを考慮したAdaBoostに基づくマルウェア感染検知手法，情報処理学会論文誌 53(9) 2012年

[8] 大月，他：マルウェア感染検知のためのトラヒックデータにおけるペイロード情報の特徴量評価，MWS2012 (2012年10月)

[9] Jacob, G., et al.: Jackstraws: Picking Command and Control Connections from Bot Traffic, 20th Usenix Security Symposium. USA, August 2011.

[10] Kang, G. M., et al.: DTA++: Dynamic Taint Analysis with Targeted Control-Flow Propagation, Proceedings of the 18th Annual Network and Distributed System Security Symposium, Feb. 2011

[11] 川古谷，他：テイント伝搬に基づく解析対象コードの追跡方法，MWS2012 (2012年10月)

[12] Egele, M., et al.: A survey on automated dynamic malware-analysis techniques and tools. ACM Comput. Surv. 44, 2, Mar. 2008

[13] 神薗，他：マルウェア対策のための研究用データセット ～MWS Datasets 2013～，MWS2013 (2013年10月)

[14] TrendLabs SECURITY BLOG, マルウェア解析の現場から-06 「DGA」, http://blog.trendmicro.co.jp/archives/3799 (参照2013/08/19)