研究用データの解説

ベンチマークデータとマルウェア情報の２種類を用意しております。

ベンチマークデータ

NII-SOCS参加機関のトラフィックデータの一部を抽出してランダム化したベンチマークと、NII-SOCS攻撃検知システムの警報データのセット。

トラフィックデータは、通信の内容であるペイロードを含まないKyotoData2016フォーマット[1][2]に準拠した形に整形されております。

ベンチマークデータは、以下のファイルで構成されます

ファイル名	説明	更新
KYOTODATA-X_yyyyMMdd.txt	ベンチマークデータ本体。通信の内容であるペイロードを含まないKyotoData2016フォーマット[1][2]に準拠した形に整形し、元データに関連した参加機関側の機器が特定され難いよう、ランダム化したファイル。	翌日１回のみ
DETECTION-SUMMARY-X_yyyyMMdd.txt	検出サマリファイル。 NII-SOCSの検知システムで検知した警報(PaloaltoのThreat ID/Cisco FirepowerのGID_SID_REV)と括弧書きでセッションごとの検知回数を表記。	
AUDITDATA-SNORT-X_yyyyMMdd.txt	Snort事後検証ファイル。 SnortのGID-SID-REV、および、1セッションで2回以上の検知があった場合はその回数を括弧書きで表記。	翌日～50日後まで１週間おき

ZERODAY-SNORT-X_yyyyMMdd.txt	Snortゼロデイファイル。 ベンチマーク全体で初めて検知した検知ルール：GID-SID-REV　0 翌日の検査では未検知かつ2または5週目に検知した検知ルール：GID-SID-REV　1 5週目の検査までは未検知かつ6週目以降の検査で検知ルール：GID-SID-REV 2 と表記。(無償版snortでは検知ルールの提供が30日程度遅れることがあるため。)	該当レコードがあれば更新
ZERODAY_COUNT-SNORT-X_yyyyMMdd.txt	Snort件数ファイル。 ベンチマーク全体で初めて検知した件数(snort_new)、翌日の検査では未検知かつ2または5週目の検査で検知した件数(snort_2_5)、5週目の検査までは未検知かつ6週目以降の検査で検知した件数(snort_6_8)を表記。	該当レコードがあれば更新
AUDITDATA-CLAMAV-X_yyyyMMdd.txt	ClamAVの事後検証ファイル。 ClamAVで検知したマルウェアの名称、および、1セッションで2回以上の検知があった場合はその回数を括弧書きで表記。	翌日～50日後まで１週間おき
ZERODAY-CLAMAV-X_yyyyMMdd.txt	ClamAVゼロデイファイル。 ベンチマーク全体で初めて検知したマルウェア：マルウェア名称　0 翌日の検査では未検知かつ2週目以降に検知したマルウェア：マルウェア名称　1	該当レコードがあれば更新
ZERODAY_COUNT-CLAMAV-X_yyyyMMdd.txt	ClamAV件数ファイル。 ベンチマーク全体で初めて検知した件数(ClamAV_new)、翌日の検査では未検知かつ2週目以降の検査で検知した件数(ClamAV_2_8)を表記。	該当レコードがあれば更新
AUDITDATA-SHELLCODE-X_yyyyMMdd.txt	Shellcode事後検証ファイル。ShellocdeのID、および、1セッションで2回以上の検知があった場合はその回数を括弧書きで表記。	翌日１回のみ

※　X：元データを取得した拠点のコード　yyyyMMdd：元データの通信日

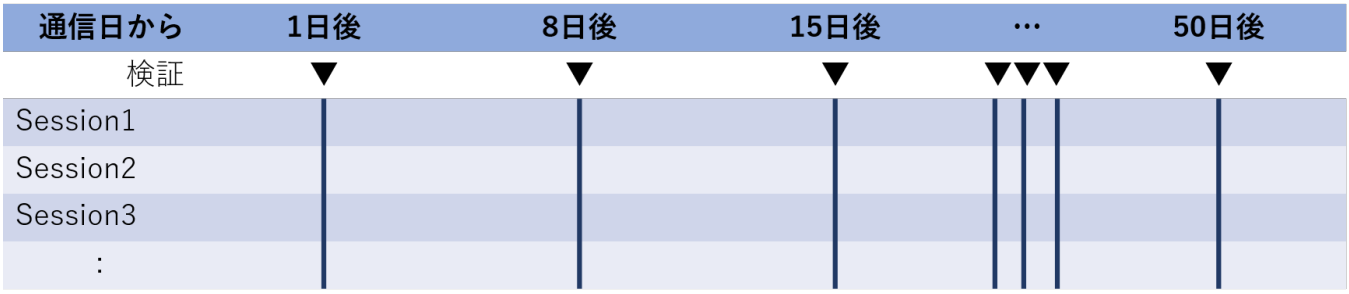
ランダム化処理方法

1. 毎日、以下の条件でトラフィックデータを抽出(pcapファイルの生成)
    a. 参加機関のIPアドレス領域から/24(IPv4)または/64(IPv6)のブロックをランダムに複数選択。(IPv4、IPv6ともに10ブロック程度、このセットを7日間使用する。)
    b. 00時00分00秒から23時00分00秒の間からランダムに30分間の枠を二つ選択。
2. 観測対象時刻のトラフィックデータのタイムスタンプを当該日の0時0分0秒から0時29分59秒と12時0分0秒から12時29分59秒に振り直す。
3. 送信元IPアドレス/受信先IPアドレスをIPv6形式のランダムなIPアドレス領域に振り直す。
    a. IPv4は/24、IPv6は/32の範囲内のIPアドレスの連続性を維持する。
    b. ランダム化処理に使用するseedは定期的に変更する。
4. ポート番号についてはwell-known port(1024未満)はそのままとし、それ以外をランダムな値に振り直す。
    a. ランダム化処理に使用するseedは7日おきに変更する。
5. 他はKyotoData2016[1][2]に準拠した統計データとし、ペイロードやDNS名は含まない。

更新時系列について

通信データ（KYOTODATA-X_yyyyMMdd.txt）と、その通信を各種検知システムで検証した結果ファイルを１セットとして提供します。

Shellcodeは検知後１回のみ、ClamAVとSnortは、通信日から50日後まで１週間おきに繰り返し検証を行い、検知結果を経過も含めて提供します。



kyoto_data

KYOTODATA-X_yyyyMMdd.txt

項目の説明

Session ID	セッションID
Duration	セッションの継続時間
Service	サービスの種類
Source_Bytes	送信バイト数
Destionation_Bytes	受信バイト数
Count	過去2 秒間のセッションのうち現在のセッションと宛先IPアドレスが同じ数
Same_srv_rate	Count特徴で該当したセッションのうち現在のセッションとサービスの種類が同じ割合
Serror_rate	Count特徴で該当したセッションのうちSYNエラーが起こった割合
Srv_serror_rate	過去2 秒間のセッションで現在のセッションとサービス種類が同じセッションのうち，”SYN"エラーが起こった割合
Dst_host_count	宛先ポートが同じ過去の100 セッションのうち，現在のセッションと送信元IP アドレスと宛先IP アドレスが同じ数
Dst_host_srv_count	宛先ポートが同じ過去の100 セッションのうち，現在のセッションと宛先IP アドレスとサービス種類が同じ数
Dst_host_same_src_port_rate	Dst_host_count特徴で該当したセッションのうち現在のセッションと送信元ポートが同じ割合:
Dst_host_serror_rate	Dst_host_count特徴で該当したセッションのうち”SYN"エラーが起こった割合
Dst_host_srv_serror_rate	Dst_host_srv_count特徴で該当したセッションのうち”SYN"エラーが起こった割合
Flag	セッション終了時の接続の状態
Source_IP_Address	送信元IPアドレス※ランダム化処理済
Source_Port_Number	送信元ポート番号　※well-known port(1024未満)はそのままとし、それ以外はランダム化処理済
Destination_IP_Address	宛先IPアドレス　※ランダム化処理済
Destination_Port_Number	宛先ポート番号　※well-known port(1024未満)はそのままとし、それ以外はランダム化処理済
Start_Time	セッション開始時刻　※開始時間変更済
Protocol	プロトコル種別

データ例

Session ID	Duration	Service	Source_Bytes	Destionation_Bytes	Count	Same_srv_rate	Serror_rate	Srv_serror_rate	Dst_host_count	Dst_host_srv_count
18922621717	0	-	0	0	0	0	0	0.980403	0	0
18922621718	0	-	0	0	0	0	0	0.980403	0	0
18922621719	0	-	0	0	0	0	0	0.980403	0	0
18922621720	0	-	0	0	0	0	0	0.980402	0	0
18922621721	0	-	0	0	1	1	1	0.980402	0	0
18922621722	0	-	0	0	0	0	0	0.980402	0	0

18922621723	0	-	0	0	0	0	0	0.980419	0	0
18922621724	0	-	0	0	0	0	0	0.980436	0	0
18922621725	0	-	0	0	0	0	0	0.980436	0	0
18922621726	0	-	0	0	0	0	0	0.980437	0	0
18922621727	0	-	0	0	0	0	0	0.980436	0	0
18922621728	0	-	0	0	0	0	0	0.980437	0	0
18922621729	0	-	0	0	0	0	0	0.980435	0	0
18922621730	0	-	0	0	0	0	0	0.980436	0	0
18922621731	0	-	0	0	0	0	0	0.980436	0	0
18922621732	0	-	0	0	0	0	0	0.980435	0	0
18922621733	0	-	0	0	0	0	0	0.980453	0	0
18922621734	0	-	0	0	0	0	0	0.980453	0	0
18922621735	0	-	0	0	0	0	0	0.980453	0	0
18922621736	3.002422	-	0	0	60	1	1	0.980453	0	93

検出サマリ

DETECTION-SUMMARY-X_yyyyMMdd.txt

項目の説明

Session ID	検証結果
セッションID セッションの詳細はkyoto_dataをご確認下さい。	NII-SOCSの検知システムで検知した警報。 PaloaltoのThreat ID / Cisco FirepowerのGID_SID_REV)　と括弧書きでセッションごとの検知回数を表記。

データ例

Session ID	検証結果
18922621726	58483(1)
18922621734	1_54794_2(1),58706(1)

Snort

無償版IDSであるsnortを用いた検知結果

AUDITDATA-SNORT-X_yyyyMMdd.txt

項目の説明

Session ID	1日後	2週目（8日後）	…
セッションID セッションの詳細はkyoto_dataをご確認下さい。	セッションの１日後にSnortで検証した結果。 SnortのGID-SID-REV、および、1セッションで2回以上の検知があった場合はその回数を括弧書きで表記。	セッションの８日後にSnortで検証した結果。	セッションのｎ日後にSnortで検証した結果。

データ例

Session ID \ 検証	1日後	2週目（8日後）	3週目（15日後）	4週目（22日後）	5週目（29日後）	6週目（36日後）	7週目（43日後）	8週目（50日後）
18922621720	1-31136-2,1-23493-6	1-31136-2,1-23493-6	1-31136-2,1-23493-6	1-31136-2,1-23493-6	1-31136-2,1-23493-6	1-31136-2,1-23493-6	1-31136-2,1-23493-6	1-31136-2,1-23493-6
18922621726	1-42016-2	1-42016-2	1-42016-2	1-42016-2	1-42016-2	1-42016-2	1-42016-2	1-42016-2
18922621734	129-12-1(2)	129-12-1(2)	129-12-1(2)	129-12-1(2)	129-12-1(2)	129-12-1(2)	129-12-1(2)	129-12-1(2)

ZERODAY-SNORT-X_yyyyMMdd.txt

項目の説明

合致したルール名	フラグ情報
GID-SID-REV	ベンチマーク全体で初めて検知した検知ルール：0 翌日の検査では未検知かつ2または5週目に検知した検知ルール： 1 5週目の検査までは未検知かつ6週目以降の検査で検知ルール： 2

データ例

合致したルール名	フラグ情報
1-42016-1	0
1-53026-1	1
1-79018-1	2

ZERODAY_COUNT-SNORT-X_yyyyMMdd.txt

項目の説明

snort_new	ベンチマーク全体で初めて検知した件数
snort_2_5	翌日の検査では未検知かつ2～5週目の検査で検知した件数
snort_6_8	5週目の検査までは未検知かつ6週目以降の検査で検知した件数 NII-SOCSではsnortの無償版を利用している。 Snortの無償版は、有償版よりルールセットの提供が30日遅れる（Snort＞Snort FAQ＞What are the differences in the rule sets? ）ため、 6週目で検知できず、7週目以降に検知できた場合は、有償版でも当初は検知できなかったセロデイ攻撃とみなすことができる。

データ例

	件数
snort_new	0
snort_2_5	0
snort_6_8	0

ClamAV

無償版アンチウィルスであるClamAVを用いた検知結果

AUDITDATA-CLAMAV-X_yyyyMMdd.txt

項目の説明

Session ID	1日後	2週目（8日後）	…
セッションID セッションの詳細は kyoto_dataをご確認下さい。	セッションの１日後にClamAVで検証した結果。 ClamAVで検知したマルウェアの名称、および、1セッションで2回以上の検知があった場合はその回数を括弧書きで表記。	セッションの８日後にClamAVで検証した結果。	セッションのｎ日後にClamAVで検証した結果。

データ例

Session ID \ 検証	1日後	2週目（8日後）	3週目（15日後）	4週目（22日後）	5週目（29日後）	6週目（36日後）	7週目（43日後）	８週目（50日後）
18922621721	Email.Exploit.Efail-6543463-0 (2)	Email.Exploit.Efail-6543463-0 (2)	Email.Exploit.Efail-6543463-0 (2)	Email.Exploit.Efail-6543463-0 (2)	Email.Exploit.Efail-6543463-0 (2)	Email.Exploit.Efail-6543463-0 (2)	Email.Exploit.Efail-6543463-0 (2)	Email.Exploit.Efail-6543463-0 (2)
18922621727	Email.Exploit.Efail-6543463-0	Email.Exploit.Efail-6543463-0	Email.Exploit.Efail-6543463-0	Email.Exploit.Efail-6543463-0	Email.Exploit.Efail-6543463-0	Email.Exploit.Efail-6543463-0	Email.Exploit.Efail-6543463-0	Email.Exploit.Efail-6543463-0
18922621732	-	-	Win.Trojan.Hide-1	Win.Trojan.Hide-1	Win.Trojan.Hide-1	Win.Trojan.Hide-1	Win.Trojan.Hide-1	Win.Trojan.Hide-1

ZERODAY-CLAMAV-X_yyyyMMdd.txt

項目の説明

マルウェア名	フラグ情報
合致したマルウェアの名称	ベンチマーク全体で初めて検知したマルウェア：マルウェア名称　0 翌日の検査では未検知かつ2週目以降に検知したマルウェア：マルウェア名称　1

データ例

マルウェア名	フラグ情報
Doc.Dropper.EmotetUpdate1020-9778523-0	0

ZERODAY_COUNT-CLAMAV-X_yyyyMMdd.txt

項目の説明

ClamAV_new	ベンチマーク全体で初めて検知した件数
ClamAV_2_8	翌日の検査では未検知かつ2週目以降の検査で検知した件数 初回で検知できず２週目以降に検知できた場合は、当初は検知できなかったセロデイ攻撃とみなすことができる。

データ例

	件数
ClamAV_new	0
ClamAV_2_8	1

Shellcode

外部権限取得(remote exploit)プログラム[3]の有無判定の検知結果

AUDITDATA-SHELLCODE-X_yyyyMMdd.txt

項目の説明

Session ID	ShellcodeID
セッションID セッションの詳細はkyoto_dataをご確認下さい。	セッションの１日後にShellcodeで検証した結果。 ShellcodeのID(確認順に採番)

データ例

Session ID	ShellcodeID
18922621718	11
18922621729	12

マルウェア情報

NII-SOCSで観測したマルウェア検体、および、その検体のNII-SOCSのサンドボックスの解析結果。

提供するマルウェア情報の選定条件は以下のものとなります。

- exe、dll、dmg、apk、javascript、swfなどからなる実行形式ファイル （※pdf, office系など文書系ファイルは提供しない）
- ５機関以上の通信で検知があったもの

マルウェア情報は、以下のファイルで構成されます

フォルダ名	説明
malware/	マルウェアファイル（検体）。 マルウェアファイルをパスワード付zip形式で圧縮したファイルです。解凍時のパスワードはreadme内にあります。
analysis/	マルウェアファイル挙動解析ファイル。 挙動解析結果ファイルをzip形式で圧縮したファイルです。
readme/	Shellcode事後検証ファイル。

※　最上位フォルダの名前は該当マルウェアのハッシュ値となっています

---

[1] Jungsuk Song, Hiroki Takakura and Yasuo Okabe, “Cooperation of Intelligent Honeypots to Detect Unknown Malicious Codes,” WOMBAT Workshop on Information Security Threat Data Exchange (WISTDE 2008), Amsterdam, Netherlands, 21-22 April 2008.

[2] 多田竜之介, 小林良太郎, 嶋田創, 高倉弘喜, NIDS評価用データセット： Kyoto 2016 Datasetの作成, 情報処理学会論文誌, Vol. 58, No.9, pp.1450-1463, 2017年9月.

[3] 野川裕記, 足立史宜, 辻野泰充, 守屋誠司, 齋藤和典, エクスプロイトコードの中のシェルコード検知：構造分析に基づいた検知手法, 電子情報通信学会技術研究報告. IA, インターネットアーキテクチャ 109(85), pp.7-12, 2009年6月.