2013-10-21(Mon)
マルウェア対策研究人材育成ワークショップ(MWS2013)
特別セッション

マルウェア対策のための研究用データセット ～ MWS Datasets 2013 ～

神薗 雅紀 (NICT/セキュアブレイン)
畑田 充弘 (NTTコミュニケーションズ)
寺田 真敏 (日立製作所)
秋山 満昭 (NTT SC研)
笠間 貴弘 (NICT)
村上 純一 (ＦＦＲＩ)

2013/10/21

背景

- 高度かつ複雑化したサイバー攻撃が国際的な問題
  - APT攻撃
  - ボットネットを利用した国家および企業間レベルでのDDoS攻撃
  - Drive-by Download攻撃（DBD攻撃）
- 国家および企業レベルでの対策が急務

- マルウェア対策やそこから派生する様々な研究が盛んに行われているが，研究を進める上で様々な課題が浮き彫りに！

課題

- 共通の研究素材がない
  - 独自に収集した研究教材に対し、提案手法の有効性を評価
  - 研究を相互に比較することが困難

- 研究素材を収集すること自体が困難
  - 例）近年のDBD攻撃サイトは様々な解析および検知回避機能等を有しており，期待した情報を得ることが困難
  - 例）近年のC&Cサーバは短時間で活動を停止するため，期待した通信データや挙動を収集することが困難

目的

- 研究教材（MWS Datasets 2013）の提供
  - マルウェア対策を含むサイバー攻撃対策を研究する上で必要となる情報を可能な限り網羅的に，かつ攻撃の進化に合わせて適切に選択したデータセット
  - 研究用データセット自身が研究者間での共通言語としての役割を担う

- 研究開発した技術等の共有により，人材育成を含む本研究分野の発展に寄与する
  - マルウェア対策研究人材育成ワークショップ
  - MWS Cup

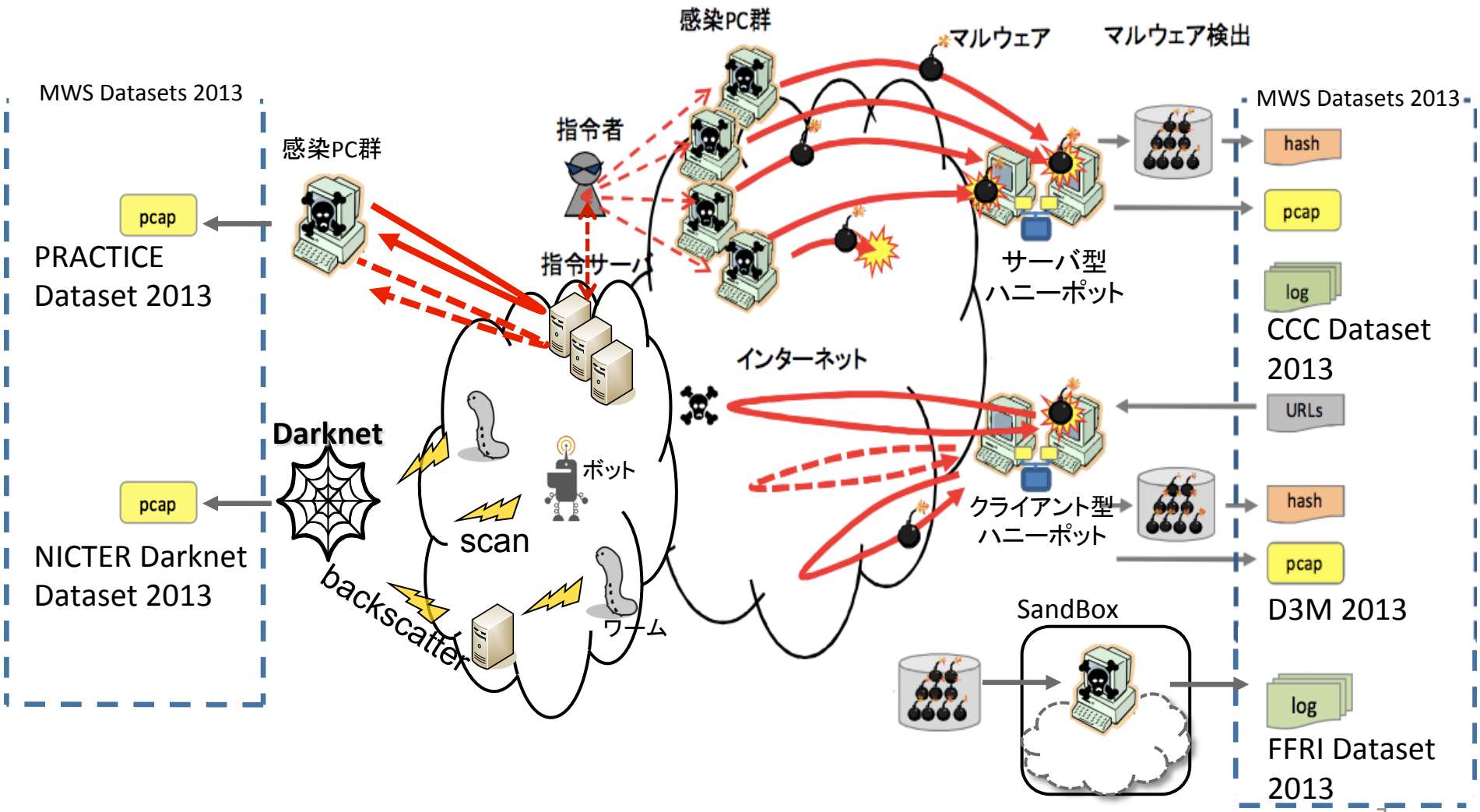
関連動向（関連研究）

- DARPA Intrusion Detection Evaluation Data Sets
  - 侵入検知システムの評価に用いられる
  - 2000年のデータセットが最新
- CDX Datasets
  - 2009 年、サイバー防御演習時のデータセット
  - マルウェアによる攻撃ではない
- BADGERS
  - 大規模セキュリティ関連データの収集と分析をもとに，より良いデータとナレッジの共有をするワークショップ
- PREDICT
  - コンピュータ・ネットワークの運用データをレポジトリとして蓄積し，インフラ防護と脅威評価に活用する
- CAIDA
  - 広域ネットワークの情報・状況を分析し，幾つかのタイプのデータセットを提供するプロジェクト

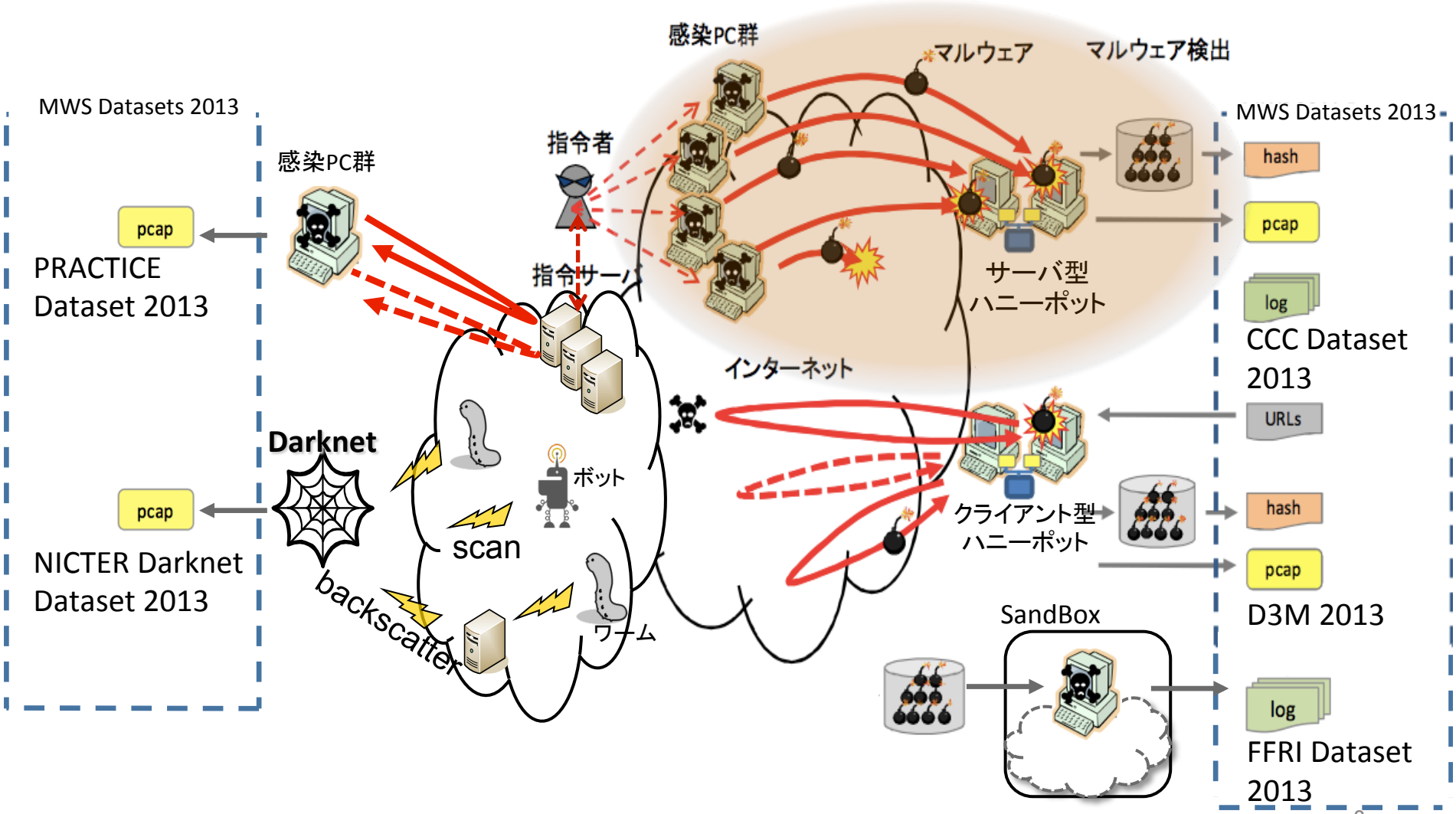
MWS 2013 Datasetsの概要

- CCC Dataset 2013（2008～2012含む）
  - CCC運営連絡会が運用する サーバ型ハニーポットで収集したマルウェア検体とウイルス対策ソフト6製品での検知名
- NICTER Darknet Dataset 2013
  - NICTが監視するダークネットトラフィック（過去、準リアルタイム）
- D3M 2013（2010～2012含む）
  - NTT SC研究所の高対話型のWebクライアント型ハニーポット(Marionette)で収集したマルウェア検体、攻撃通信データ
- FFRI Dataset 2013
  - ＦＦＲＩのマルウェア自動解析によるマルウェアの挙動ログ
- PRACTICE Dataset 2013
  - 総務省「国際連携によるサイバー攻撃予知・即応に関する実証実験」プロジェクトで得られたマルウェア長期観測データ

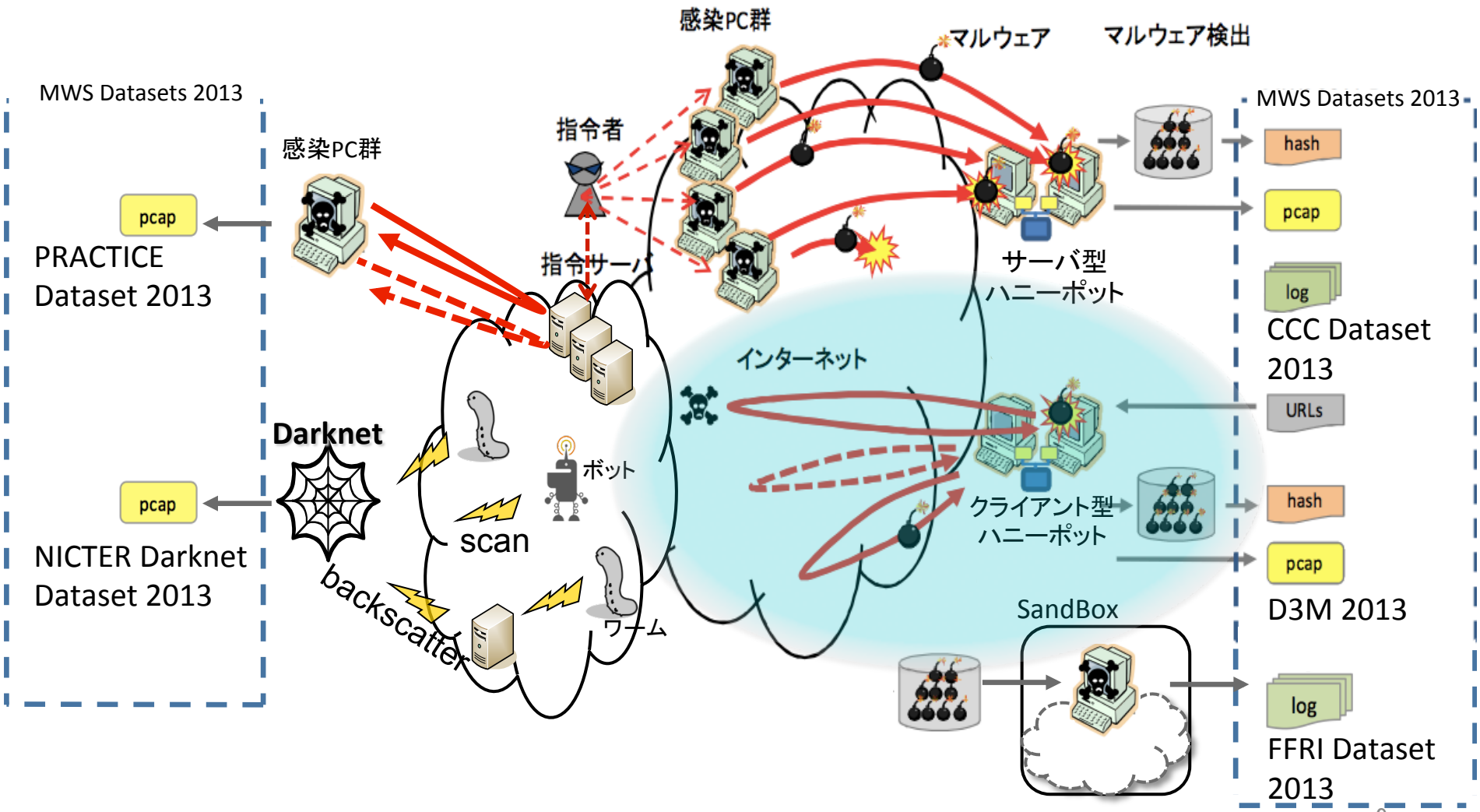
MWS 2013 Datasetsの概要



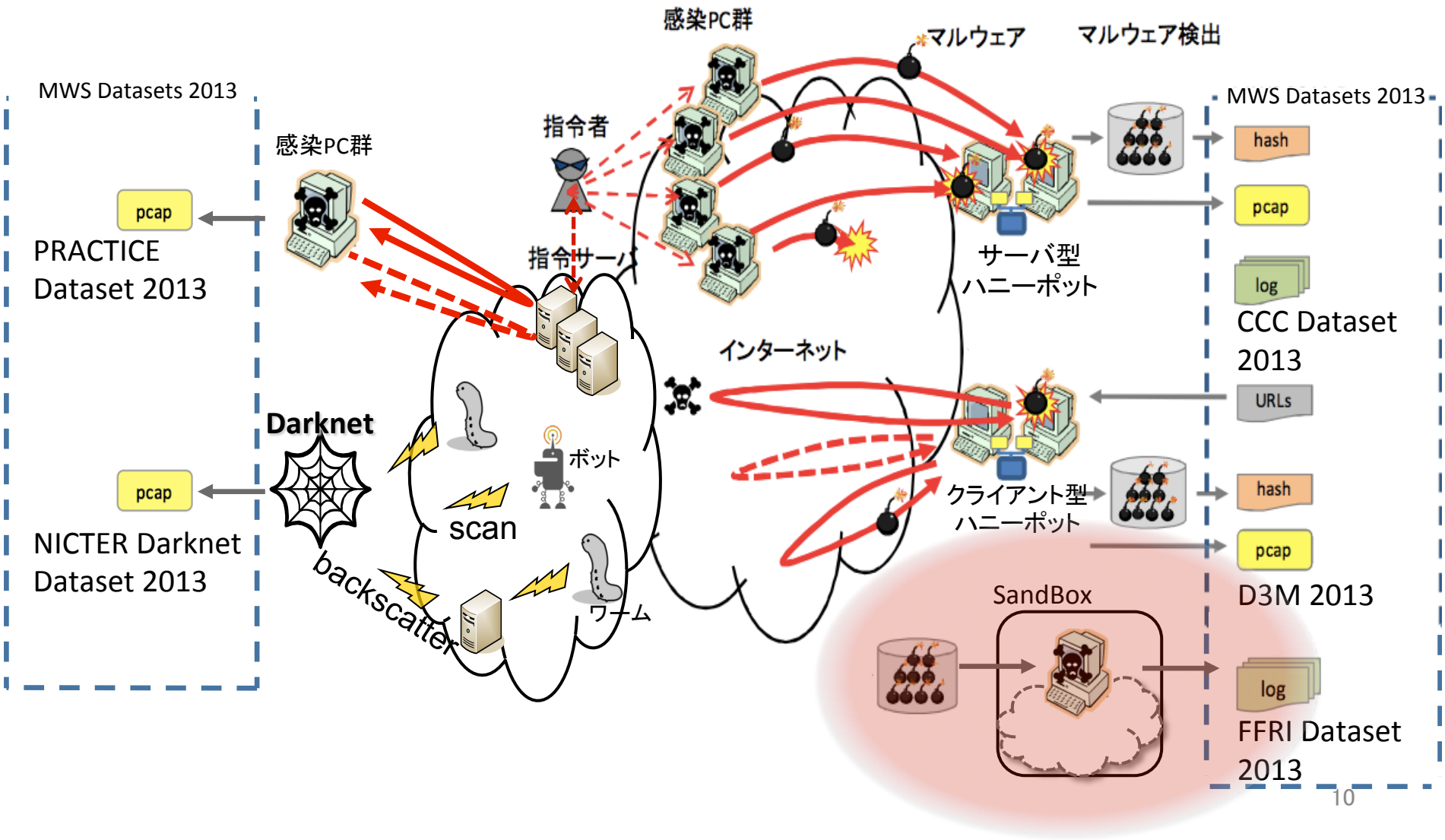
MWS 2013 Datasetsの概要



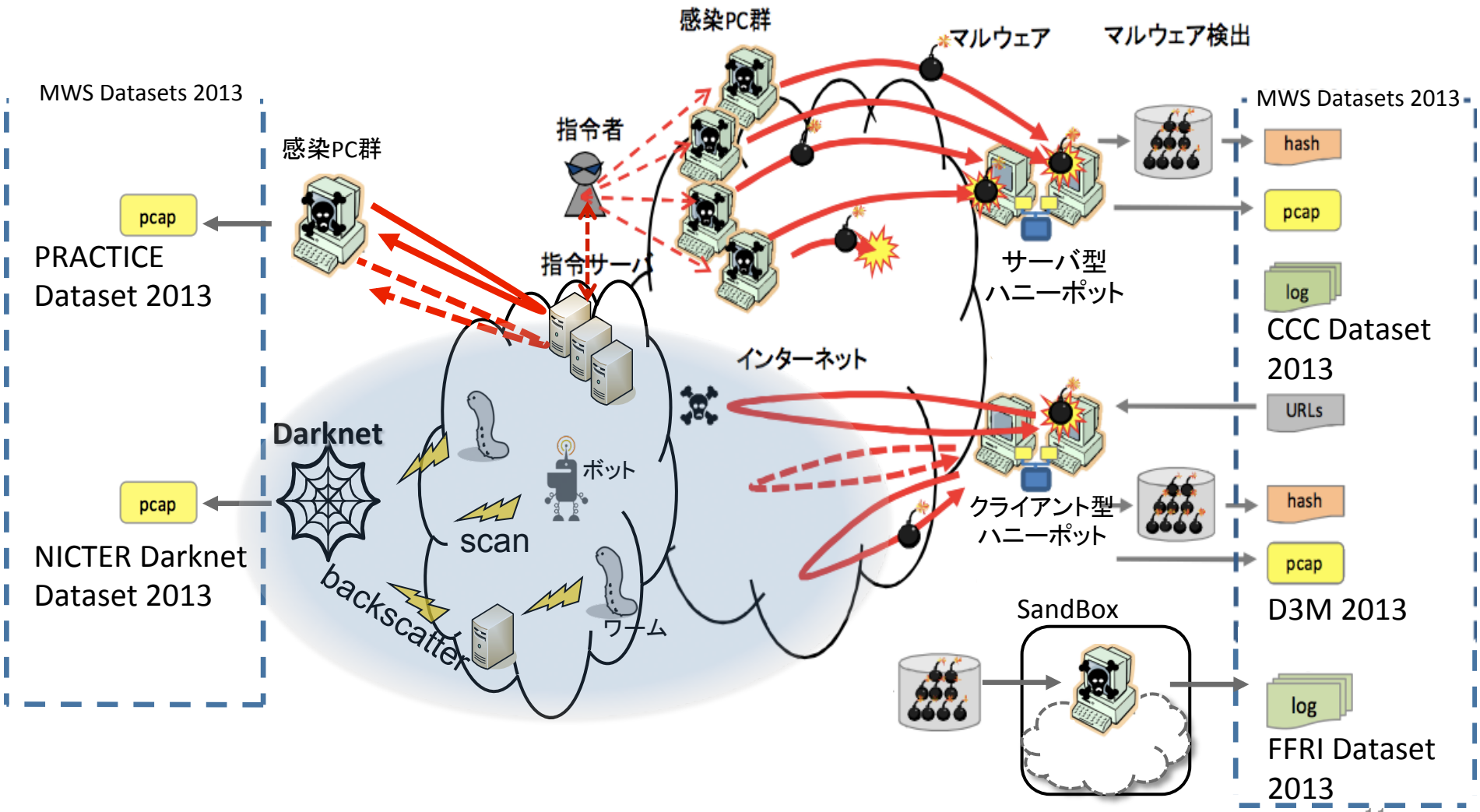
MWS 2013 Datasetsの概要



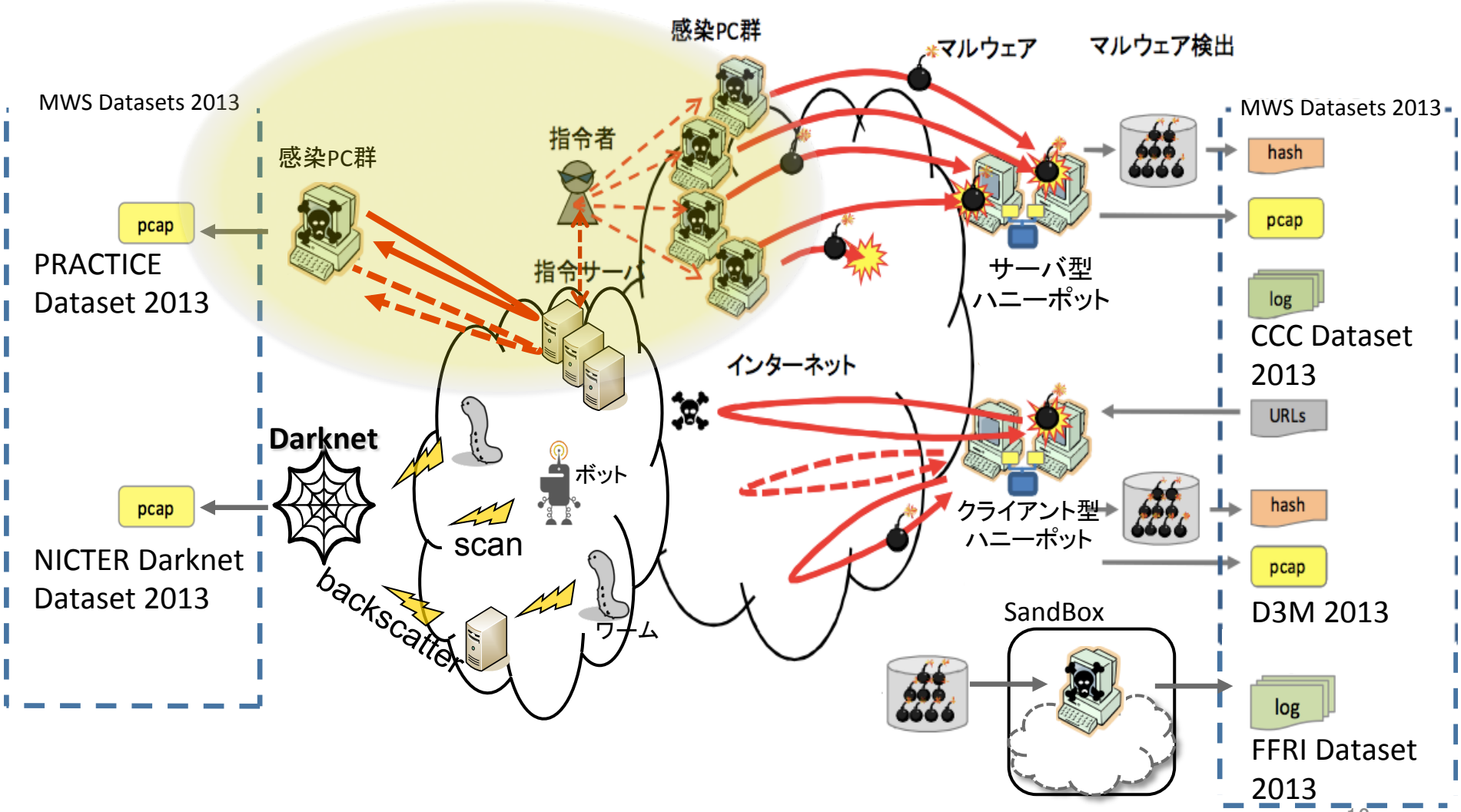
MWS 2013 Datasetsの概要



MWS 2013 Datasetsの概要



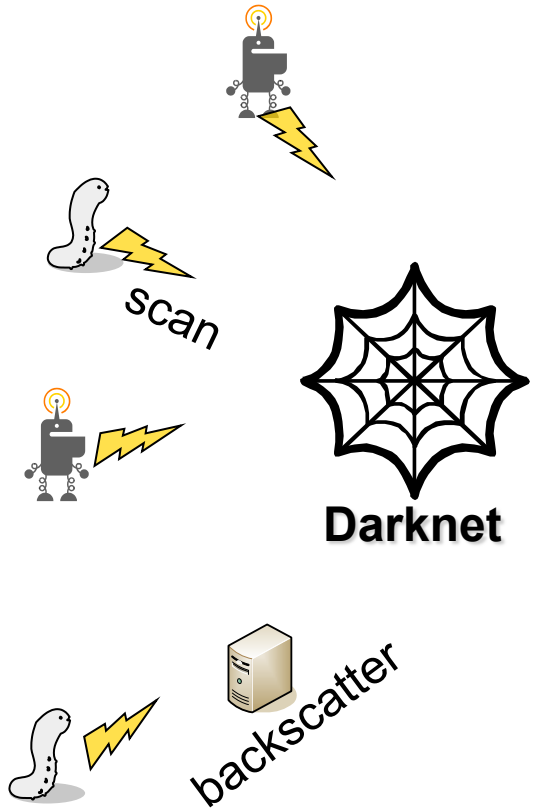
MWS 2013 Datasetsの概要



NICTER Darknet 2013

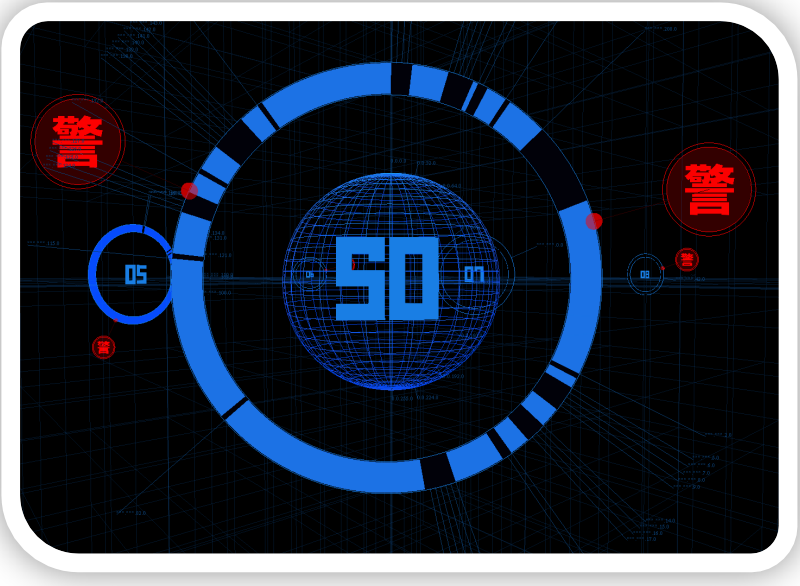
darknetとは？

- 実ホストが存在しない未使用IPアドレス（ブロック）
- ダークネットに届くパケットは
  - マルウェア（リモートエクスプロイト型）によるスキャン
  - マルウェア本体の感染行為（主にUDP）
  - DDoS攻撃の跳ね返り（バックスキャッタ）
  - 設定ミス　などが原因
- インターネット上で広範囲に影響を与える攻撃の把握に役立つ

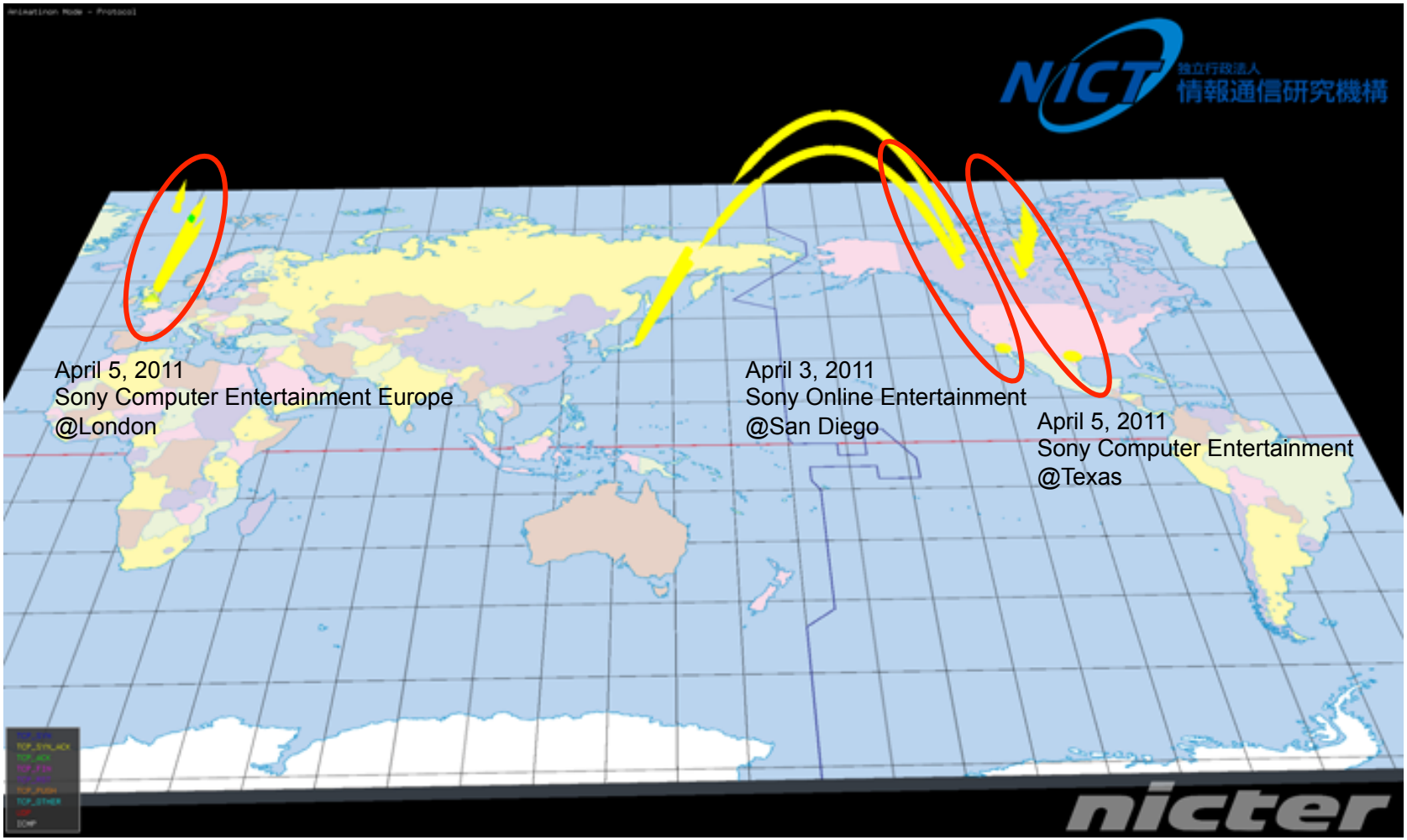


nicter darknet 2013 内容

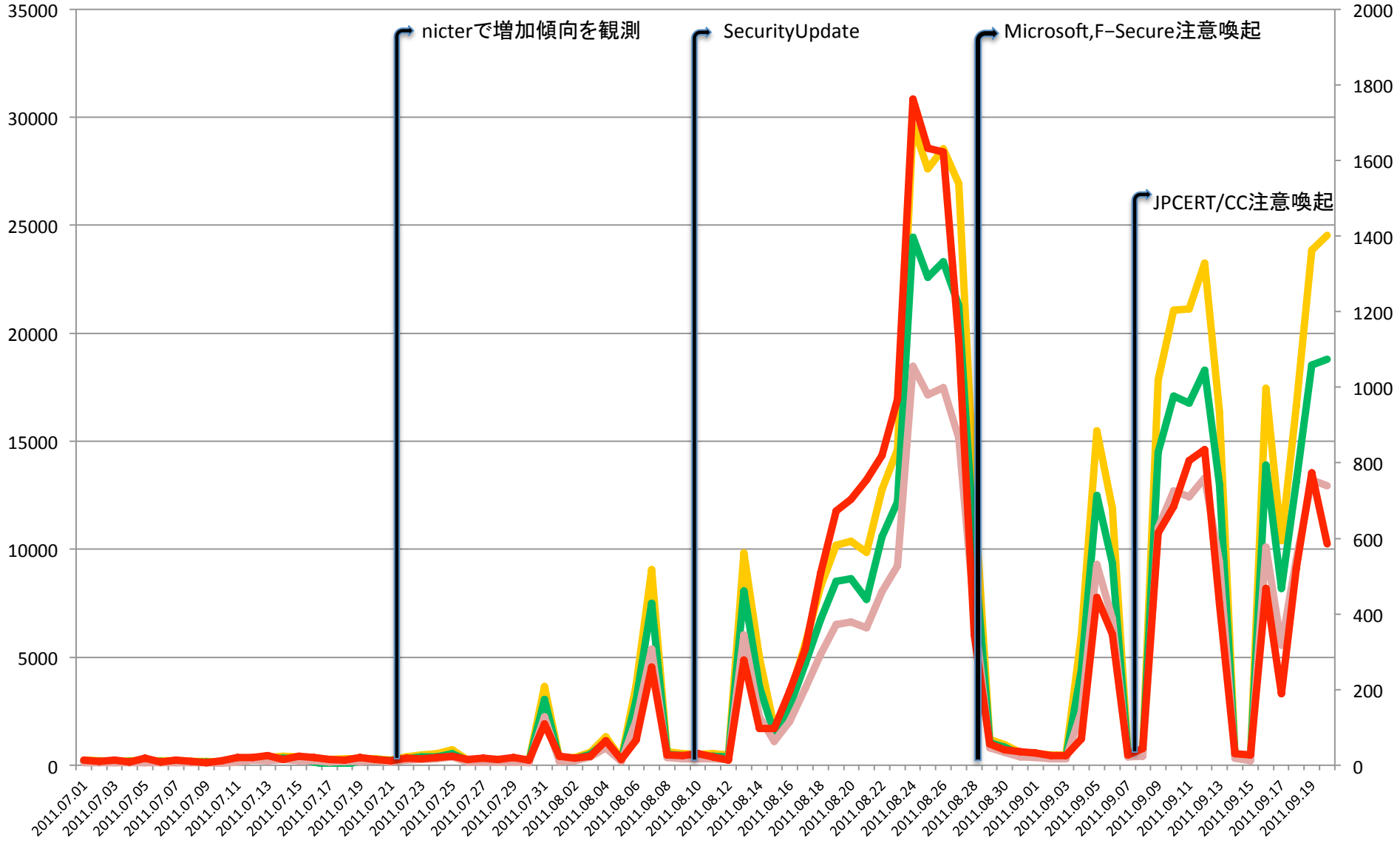
- ダークネットトラフィックデータ
  - ある1組織（/16）のダークネット（/20）で観測されたトラフィックデータ（Pcap）
  - 2011年4月1日～2013年3月31日の2年間分＋α
  - NONSTOPにてデータセットを提供
    - サイバーセキュリティ情報（ダークネットトラフィック,マルウェア検体,etc.）を遠隔から安全に利用するための分析基盤



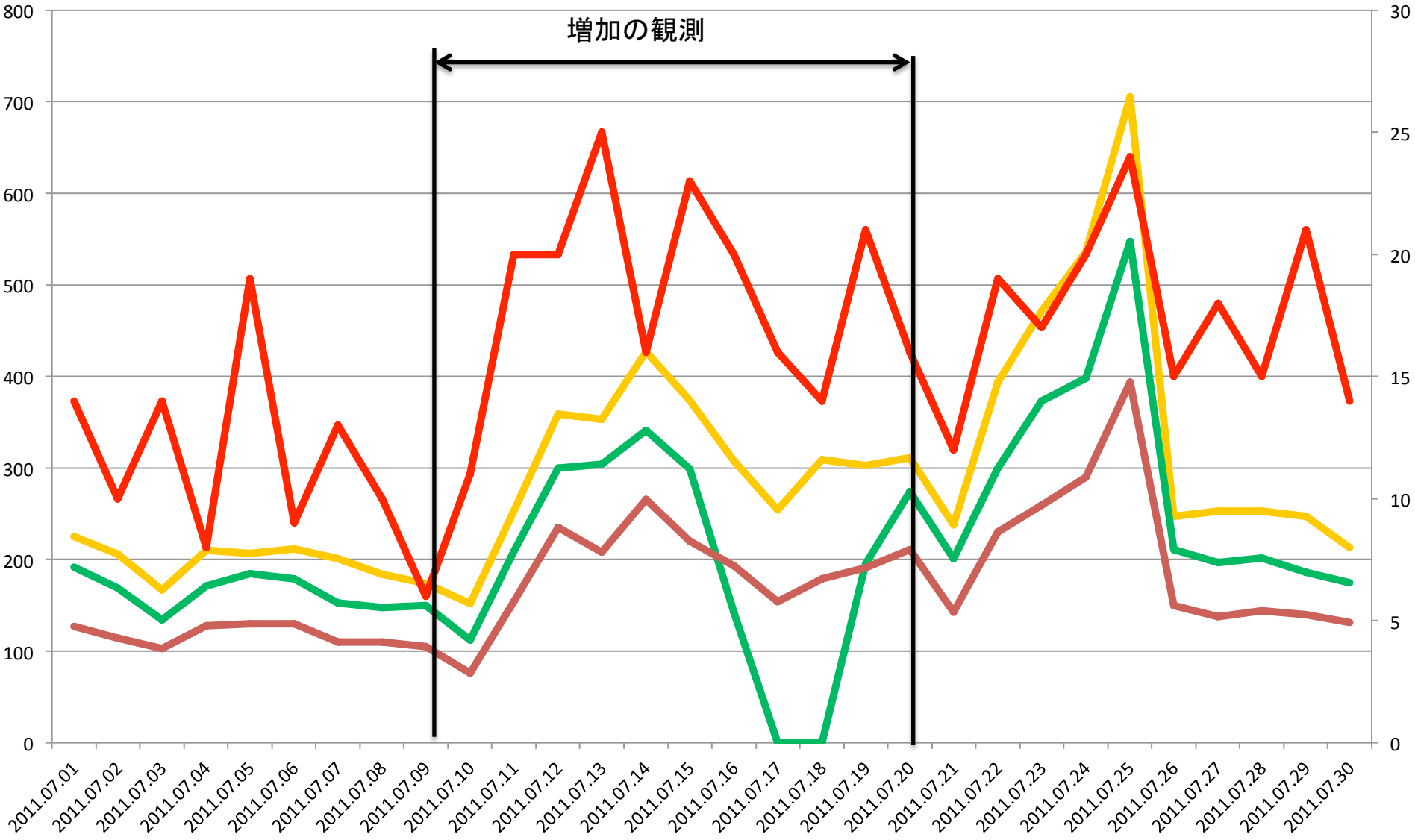
SONY DDoS攻撃時の観測状況



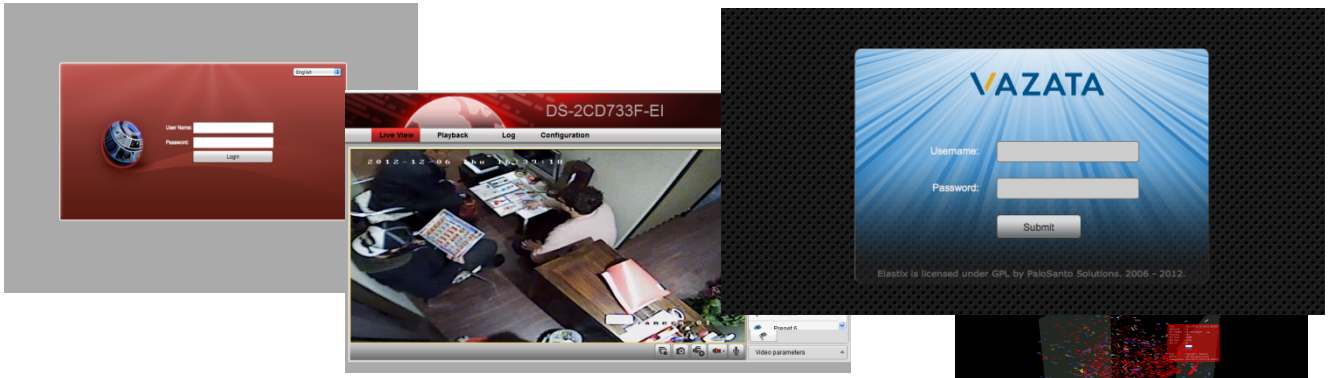
3389/TCPへの通信増加（Morto）



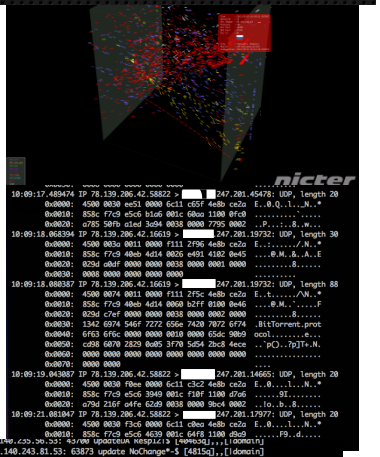
3389/TCPへの通信増加（Morto）

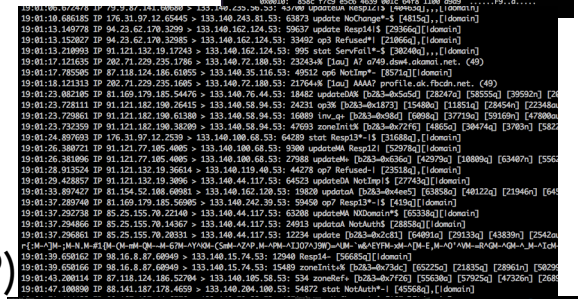


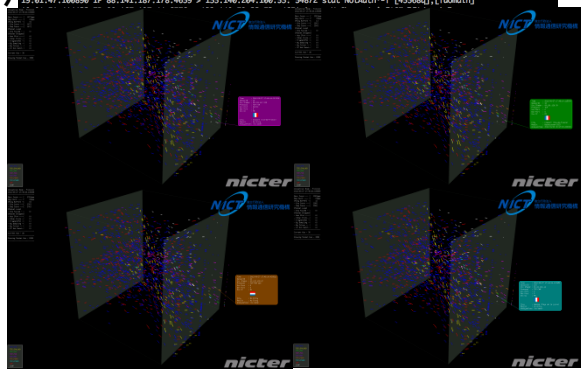
2012年度トピック

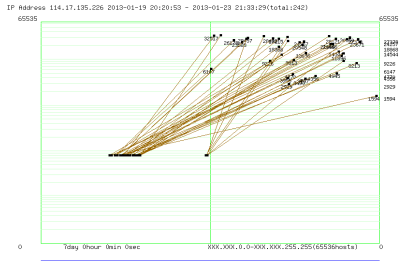


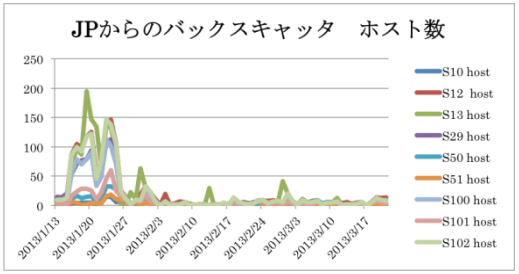
- 23/TCP　　idora？　ルータにIRCボットが感染
- 23・210/TCP　　Hikvision Digital Technology
- BitTorrentのポイゾニング？パケットの増加
  - ー　2012年7月から日本国内から急増(詐称？)
- 0/TCPへのパケット
- 53/UDP　毎週金曜日に増加するホスト
- 1点集中UDPパケット
  - ー　400万パケット/day(10320・10321/UDP)
- JPからのバックスキャッタ











D3M（Drive-by-Download Data by Marionette）2013

日々進化するDrive-by-Download攻撃

- マルウェア感染経路の変化
  - ドライブバイダウンロード攻撃（Webブラウザの脆弱性に対する攻撃）が主流
- 脆弱性の多様化
  - Webブラウザ（IE 6/7/8/9，FireFox，Opera）、プラグインアプリケーション（Acrobat 8/9，Flash　9/10/11，Java 6/7，．．．）
- 難読化手法の高度化による検知・解析妨害
  - HTML難読化，JavaScript難読化，PDF難読化，Java難読化
- 悪性サイトのクローキング
  - 自動転送（HTTPリダイレクト、iframeリダイレクト、JavaScriptリダイレクト、外部スクリプト読込、Traffic Direction System (TDS) ）
  - クライアントブラックリスト化によるアクセス拒否
  - Torからのアクセス拒否

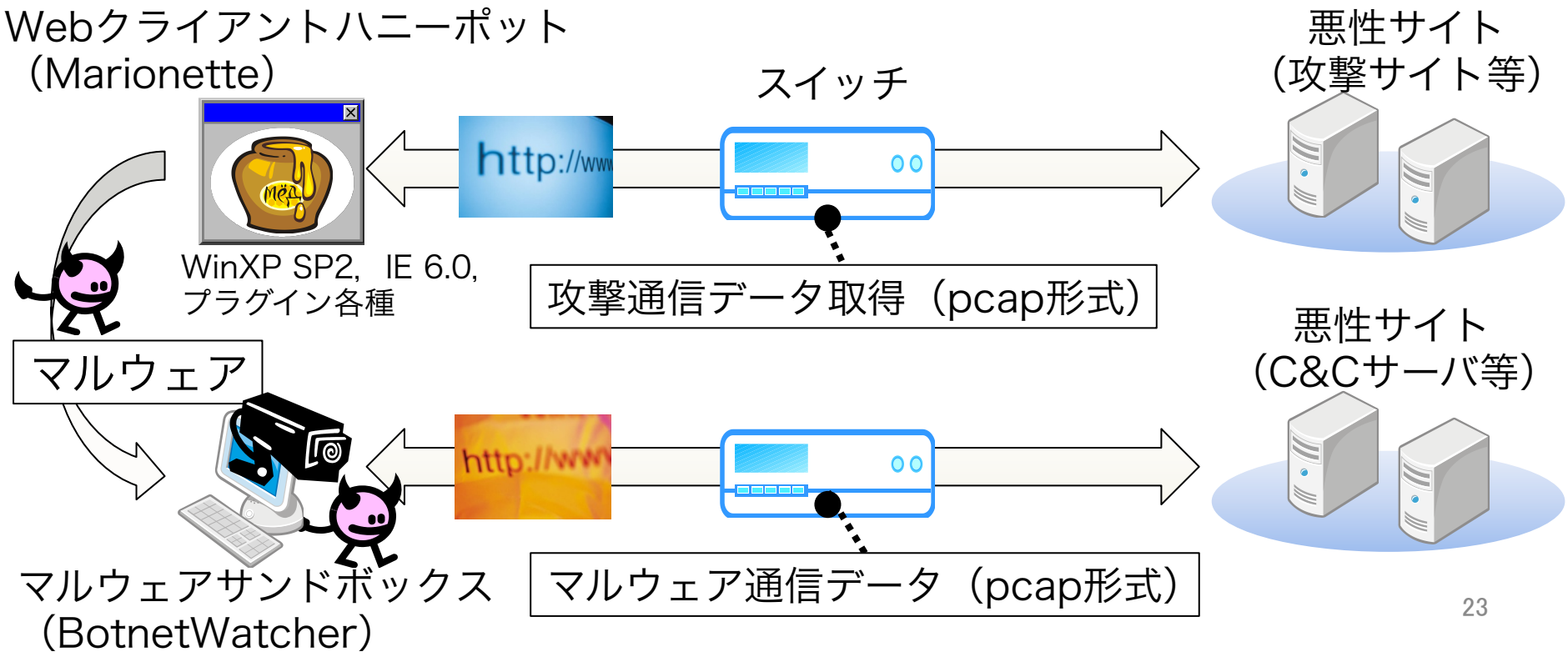
日々進化するDrive-by-Download攻撃

- マルウェア感染経路の変化
  - ドライブバイダウンロード攻撃（Webブラウザの脆弱性に対する攻撃）が主流
- 脆弱性の多様化
  - Webブラウザ（IE 6/7/8/9，FireFox，Opera）、プラグインアプリケーション（Acrobat 8/9，Flash

・高度化・多様化するドライブバイダウンロード攻撃は、データセットの収集自体も困難な状況になりつつある。
・D3Mでは一連の攻撃通信およびマルウェアの通信が記録されており、また、多様な攻撃手法やマルウェアが含まれている。

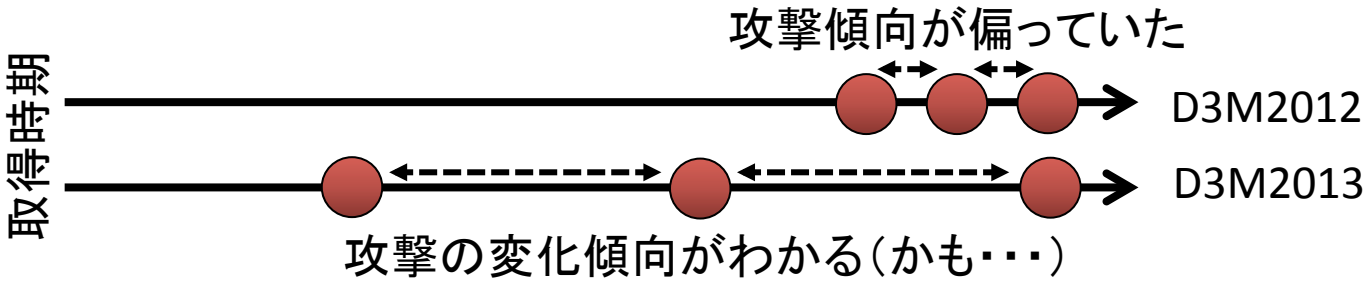
データセット概要

- ドライブバイダウンロード攻撃に関わるURLをブラウザに入力し、自動的に発生する一連のWeb通信、および感染するマルウェアの通信を記録
- 取得手順
  - １．公開ブラックリスト（※）をWebクライアントハニーポットで巡回
  - ２．検知したURLを直ちに再巡回し、その際の通信データを記録
  - ３．２で取得したマルウェア検体をマルウェアサンドボックスで解析し、その際の通信データを記録



データセット概要

- データセット内容
  - 攻撃通信データ
    - 悪性URLを巡回した際に得られたドライブバイダウンロード攻撃の通信データ
  - マルウェア
    - ドライブバイダウンロード攻撃によってホスト上にダウンロードされた実行形式のファイル
  - マルウェア通信データ
    - 取得して24時間以内にマルウェアサンドボックス上で実行した際の通信データ
    - マルウェアサンドボックスはインターネットに接続可能（攻撃通信は遮断）
- 取得時期
  - 期間を空けて合計3回分提供する予定



- D3M2013 には D3M2012 , D3M2011, D3M2010 が同梱されています

FFRI Dataset 2013

FFRI Dataset 2013の概要

- FFRIで収集したマルウェアの動的解析ログ
  - 2012/9～2013/3に収集された検体
  - PE形式かつ実行可能なもの
  - 約2600検体分（ログファイル計1.7GB)
- 下記の3検体の解析ログを同梱
  - 遠隔操作マルウェア/ MITBマルウェア / 韓国MBR破壊マルウェア

動的解析Cuckoo Sandbox

- オープンソースのマルウェア解析システム
  - 仮想環境内でマルウェアを実行
  - 実行時のふるまいをモニタリング
  - VirusTotal連携、yara連携等
- 社内のマルウェア解析用ネットワークにシステムを設置、実行
  - 1検体当り90秒実行
- 1検体（解析対象） 1ログファイル
  - ログファイルは、json形式

※ Cuckoo Sandbox - http://www.cuckoosandbox.org

具体的な解析項目

項目（大見出し）	内容
info	解析の開始、終了時刻、id等（idは1から順に採番）
yara	yara(OSSのマルウェア検知・分類エンジン）の標準ルールとの照合結果 - https://code.google.com/p/yara-project/
signatures	ユーザー定義シグニチャとの照合結果（今回は使用無）
virustotal	VirusTotalの検査履歴との照合結果（検体のMD5値に基づく）
static	検体のファイル情報（インポートAPI、セクション構造等）
dropped	検体が実行時に生成したファイル
behavior	検体実行時のAPIログ（PID、TID、API名、引数、返り値等）
processtree	検体実行時のプロセスツリー（親子関係）
summary	検体が実行時にアクセスしたファイル、レジストリ等の概要情報
target	解析対象検体のファイル情報（ハッシュ値等）
debug	検体解析時のCuckoo Sandboxのデバッグログ
strings	検体中に含まれる文字列情報
network	検体が実行時に行った通信の概要情報

PRACTICE Dataset 2013

データセット概要

- 総務省「国際連携によるサイバー攻撃予知・即応に関する実証実験」（略称：PRACTICE）の挙動観察システムで、マルウェアを長期観測（最大1週間）した際の通信トラフィック（マルウェア感染後の通信挙動）を示すデータ
- 提供するデータの内訳
  - 検体情報（ハッシュ値、AV4製品によるScan結果）
  - 通信トラフィックデータ（pcap形式）
- 対象検体
  - PRACTICEで観測中の、通信挙動で特徴的な挙動を示す検体を任意に抽出
- 対象検体数
  - 5検体
- 観測日時
  - 観測日時：2013/5/18～2013/5/25（最大1週間分のデータ）

データセット概要

- 解析環境
  - NTT SC研究所が開発した動的解析システム
  - 解析用OS：Microsoft Windows XP SP2
  - IPアドレス、デフォルトGW、DNSサーバはDHCPで割当
  - 解析環境の動作確認としてpcapに記録されているもの
    - www.google.co.jpにHTTPの（主に）HEADリクエストを送信
    - 時刻同期：ntp.jst.mfeed.ad.jp
    - グローバルIPアドレス確認：checkup.dyndns.org
    - DNSクエリ：gk-open10の名前解決
- AV検知結果
  - 各検体収集時点で最新の各社パターンファイルに基づく
    - practice_1: 2012/09/12
    - Practice_2: 2012/03/22
    - Practice_3: 2012/09/28
    - Practice_4: 2012/12/05
    - Practice_5: 2012/02/27

長期観測型通信情報

- 提供する長期観測の検体情報　※検体そのものは提供しない

データセット名	検体Hash（SHA-1）	挙動解析IP	AV検知結果		解析時間	ファイルサイズ
practice_1.pcap	5b9f78af4e5609c17fdff4d97e060d1a264b72d3	10.220.0.36	Kaspersky	未検出	start: 2013-05-18 02:35:06 end: 2013-05-25 11:59:53	10MB
			McAfee	PWS-Zbot.gen.alu		
			Symantec	未検出		
			TrendMicro	未検出		
practice_2.pcap	5944b5a106a75a7d0c4b7fe2f4099efb7ba79eae	10.220.0.37	Kaspersky	Backdoor.Win32.VanBot.cx	start: 2013-05-18 02:35:19 end: 2013-05-25 11:35:21	2.6MB
			McAfee	Generic BackDoor		
			Symantec	W32.Spybot.Worm		
			TrendMicro	WORM_MYTOB.IR		
practice_3.pcap	2fec8e24ac3c911955c37ddab6904b2e7db74309	10.220.0.38	Kaspersky	Trojan-Ransom.Win32.PornoAsset.abtn	start: 2013-05-18 02:35:36 end: 2013-05-20 02:00:01	494MB
			McAfee	ZeroAccess.hj		
			Symantec	Trojan.Zeroaccess!g19		
			TrendMicro	TROJ_GEN.RCCC7IT		
practice_4.pcap	12dba89f2c869ff6f12f8005dfb004628e2c983d	10.220.0.39	Kaspersky	Backdoor.Win32.ZAccess.ylb	start: 2013-05-18 02:35:55 end: 2013-05-20 02:00:00	231MB
			McAfee	ZeroAccess.hg		
			Symantec	Trojan.Gen		
			TrendMicro	TROJ_GEN.USBJ05ACN		
practice_5.pcap	093584d4f63d45fb46beb390ba9c10b73b394a88	10.220.0.40	Kaspersky	Trojan-Spy.Win32.SpyEyes.wb	start: 2013-05-18 02:36:16 end: 2013-05-25 11:59:58	4.3MB
			McAfee	Artemis!1E7C50EACE3D		
			Symantec	Trojan.Gen		
			TrendMicro	TSPY_SPYEYE.SME		

データセット例

```
$ capinfos practice_1.pcap
File name:           practice_1.pcap
File type:           Wireshark/tcpdump/... - libpcap
File encapsulation:  Ethernet
Packet size limit:   file hdr: 4096 bytes
Number of packets:   54403
File size:           10977185 bytes
Data size:           10106713 bytes
Capture duration:    638687 seconds
Start time:          Sat May 18 02:35:06 2013
End time:            Sat May 25 11:59:52 2013
Data byte rate:      15.82 bytes/sec
Data bit rate:       126.59 bits/sec
Average packet size: 185.77 bytes
Average packet rate: 0.09 packets/sec
SHA1:                dafceef264eb9c504ff26b2e81ae779dfb454ba4
RIPEMD160:           64ad22ab0f2c2da7fbdbfe2de04deb32735a36a1
MD5:                 9a5907924435873793 23b57094cd078b
Strict time order:   True
```

TCP Endpoints

Address	Port	Packets	Bytes	Tx Packets	Tx Bytes	Rx Packets	Rx Bytes
211.209.241.213	13109	418	209 432	196	194 048	222	15 384
94.137.177.75	26614	1 150	162 281	470	60 931	680	101 350
10.220.0.36	61532	181	137 653	83	5 676	98	131 977
211.75.189.231	20897	1 025	137 241	500	53 727	525	83 514
98.201.1							9 510
46.48.1							85 140
91.12							4 870
99.18							11 852
46.8.1							72 888
188.2							66 354
95.78.							6 620
24.24							6 326

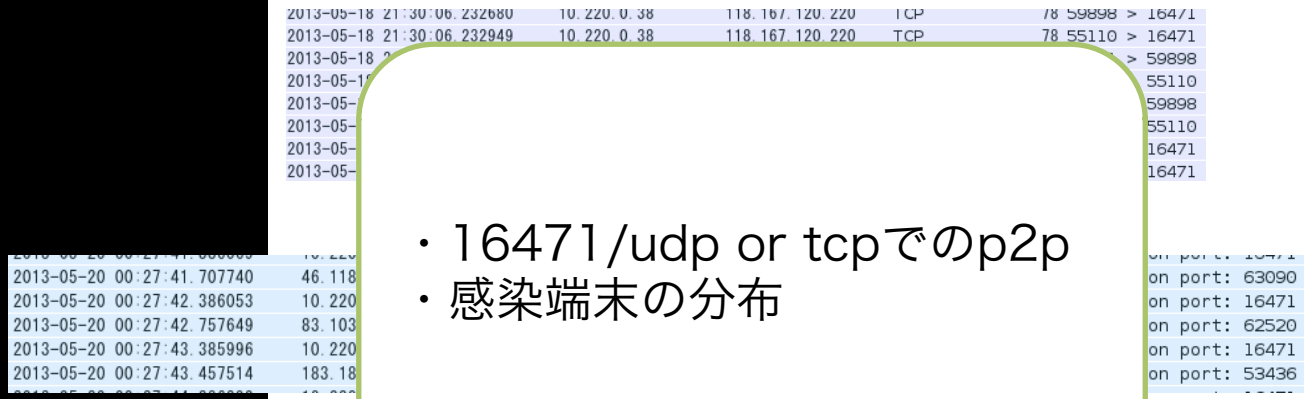
Addres							Bytes
10.220.							22 860
10.220.							12 860
10.220							10 808
194.94.							21 326
94.203.							71 819
75.34.30.							59 007
24.247.216.229	11992	600	133 630	317	80 383	283	53 287
123.110.175.202	15187	618	132 336	312	75 187	306	57 149
75.76.164.189	10788	598	132 125	314	77 960	284	54 165
86.54.238.222	22428	609	128 252	277	65 976	332	62 276

・TCP/UDPともにランダムのようなhigh port利用
・たまにhttpとか
・名前解決に失敗も多々
・UDPはホスト毎の送受信割合がある程度一定

```
$ tshark -r practice_1.pcap -q -z conv,ip | head -12
================================================================================
IPv4 Conversations
Filter:<No Filter>
                                |      <-      |      ->      |    Total     |  Rel. Start  |   Duration   |
                                | Frames  Bytes | Frames  Bytes | Frames  Bytes |              |              |
10.220.0.100      <-> 10.220.0.36          3340   333668   3344   426240   6684   759908    0.000539000   637201.6798
194.94.127.98     <-> 10.220.0.36           613   123042    602    63335   1215   186377  191.701388000   638470.3955
94.137.177.75     <-> 10.220.0.36           705   106025    495    67150   1200   173175  57003.557062000   307555.2854
46.48.235.250     <-> 10.220.0.36           697    85140    480    46530   1177   131670   69.778915000   637854.7095
46.8.115.89       <-> 10.220.0.36           668    73294    403    42589   1071   115883  273766.309451000   363971.2925
188.242.252.23    <-> 10.220.0.36           583    66354    478    46500   1061   112854  273833.027220000   364027.9457
211.75.189.231    <-> 10.220.0.36           527    83914    502    54559   1029   138473  113839.445968000   193919.9510
```

データセット例

```
$ capinfos practice_3.pcap
File name:           practice_3.pcap
File type:           Wireshark/tcpdump/... - libpcap
File encapsulation:  Ethernet
Packet size limit:   file hdr: 4096 bytes
Number of packets:   1160063
File size:           517963595 bytes
Data size:           499402563 bytes
Capture duration:    170664 seconds
Start time:          Sat May 18 02:35:36 2013
End time:            Mon May 20 02:00:00 2013
Data byte rate:      2926.23 bytes/sec
Data bit rate:       23409.82 bits/sec
Average packet size: 430.50 bytes
Average packet rate: 6.80 packets/sec
SHA1:                86efa5943323aec7c874ad413efe9d2ddbfca7a2
RIPEMD160:           d923294e579107d975eadd766ad82fcf1fd01618
MD5:                 8f41893919cdb2c8f94668d4919fba8c
Strict time order:   True
```



- 16471/udp or tcpでのp2p
- 感染端末の分布

```
$ tshark -r practice_3.pcap -q -z conv,ip | head -12
================================================================================
IPv4 Conversations
Filter:<No Filter>
                                |       <-      |  |       ->      |  |     Total     |  Rel. Start  |   Duration   |
                                | Frames  Bytes |  | Frames  Bytes |  | Frames  Bytes |              |              |
219.80.142.21       <-> 10.220.0.38          3452    218172      20     9958    3472    228130  3399.178702000    167228.7679
120.201.89.250      <-> 10.220.0.38          1135     70554    1154   417898    2289    488452   761.741767000    169177.2880
158.254.253.254     <-> 10.220.0.38          1869    113420       0        0    1869    113420   289.677549000    170294.6172
134.254.253.254     <-> 10.220.0.38          1824    110630       0        0    1824    110630   285.677644000    170295.6173
166.254.253.254     <-> 10.220.0.38          1822    109320       0        0    1822    109320   291.677416000    170291.6174
113.254.253.254     <-> 10.220.0.38          1820    109200       0        0    1820    109200   293.677414000    170291.6173
206.254.253.254     <-> 10.220.0.38          1806    108360       0        0    1806    108360   280.677486000    170295.6175
```

データセット利用状況

MWS Datasets		2008	2009	2010	2011	2012	2013
CCC Dataset	マルウェア検体	5	7	6	5	7	3
	攻撃通信データ	9	14	5	6	2	
	攻撃元データ	8	6	5	4		
MARS Dataset				1	1		
D3M				4	3	3	9
IIJ MITF Dataset						1	
FFRI Dataset							5
PRACTICE Dataset							3
NICTER Darknet Dataset							6
全部(データセット説明)			1[1]	1[2]	1[3]		1（本稿）
合計 ()内は学生発表の件数		22 (8)	28 (15)	22 (10)	20 (9)	13 (9)	25 (10)

※一部、複数のデータセット利用した論文あり

今後の課題

- 攻撃の進化に沿ったデータセットを継続的に提供
  - 標的型攻撃などのデータセット？？
  - 特定の（秘密）情報はどうマスキングする？？
  - Androidマルウェア？
- データセット間の相関が把握できる情報
- ホワイト（良性）データセットの提供
- データセットの安全な提供方法の確立
  - 今年度、一部のみNONSTOPにて提供
  - データセットをどのように安全かつ研究しやすく提供するか？

MWS 2013（MWS Cup 2013）は
これから始まります。
盛り上げて行きましょう！！

ご清聴ありがとうございました