

# マルウェア対策のための研究用データセット ～ MWS 2011 Datasets ～

MWS2011

Oct. 19, 2011

畑田充弘（NTTコミュニケーションズ）

中津留勇（JPCERT/CC →ラック）

秋山満昭（NTT情報流通プラットフォーム研究所）

- 課題、関連動向、目的
- MWS 2011 Datasets
  - CCC DATASet 2011
  - D3M 2011
- 研究用データセットの利用実績
- 今後の課題と対策

- 共通の教材がないこと
  - 標準画像データベース (SIDBA)
  - ATRの音声データベース
- 研究用データを手に入れること自体が難しくなっていること
  - ハニーポット運用
  - マルウェア収集・保管・配布

- DARPA
  - 2000年が最新、10年前ながら未だに利用されている
- 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
- BADGERS2011
  - 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プロジェクト

# 目的



- 研究用データセットの提供
  - CCC DATASET 2008,2009
  - MWS 2010 Datasets
  - MWS 2011 Datasets
- 研究成果を共有する場、切磋琢磨する環境作り
  - MWS 2008-2010
    - MWS Cup 2009,2010
  - MWS 2011
    - MWS Cup 2011

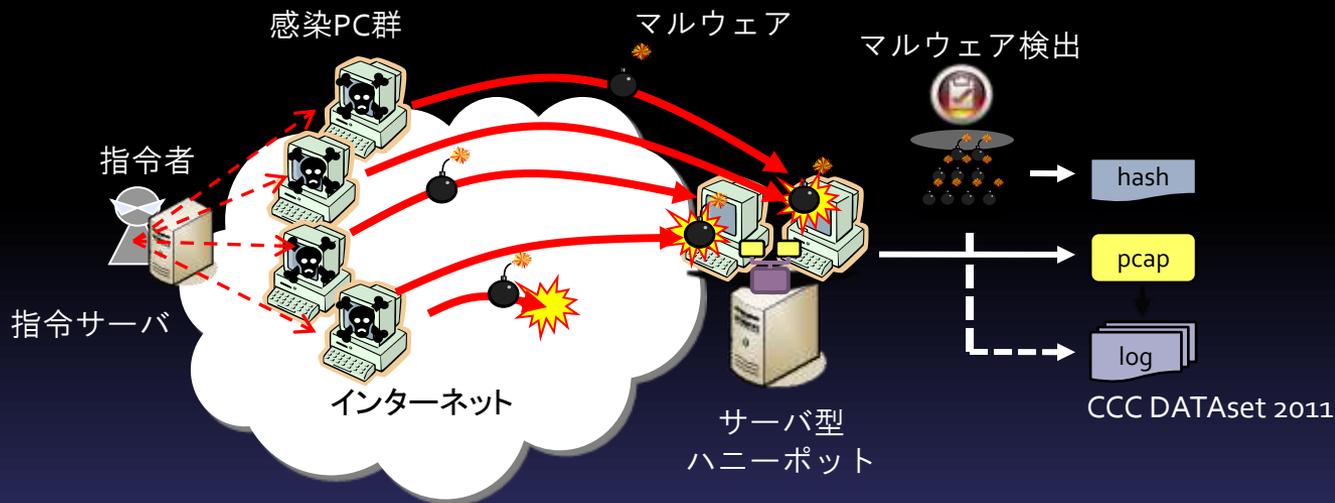
# MWS 2011 Datasets

- CCC DATASET 2011 (NTTコミュニケーションズ/CCC)
  - CCCが提供するサーバ型ハニーポットで収集したデータ群
- D3M 2011 (NTTコミュニケーションズ/NTT PF研)
  - クライアント型ハニーポットで収集したWeb感染型マルウェアに関するデータ群

(提供元/作成元)

# CCC DATAsset 2011

- マルウェア検体
- 攻撃通信データ
- 攻撃元データ



# CCC DATAsset 2011 - マルウェア検体



- 50検体のハッシュ値
- 解析結果を照合できる検体：10検体
- 2011年1月に収集した未知検体：40検体
- 用途；マルウェア解析技術

1A2-4:

CCC DATAsset 2011

マルウェア検体解説

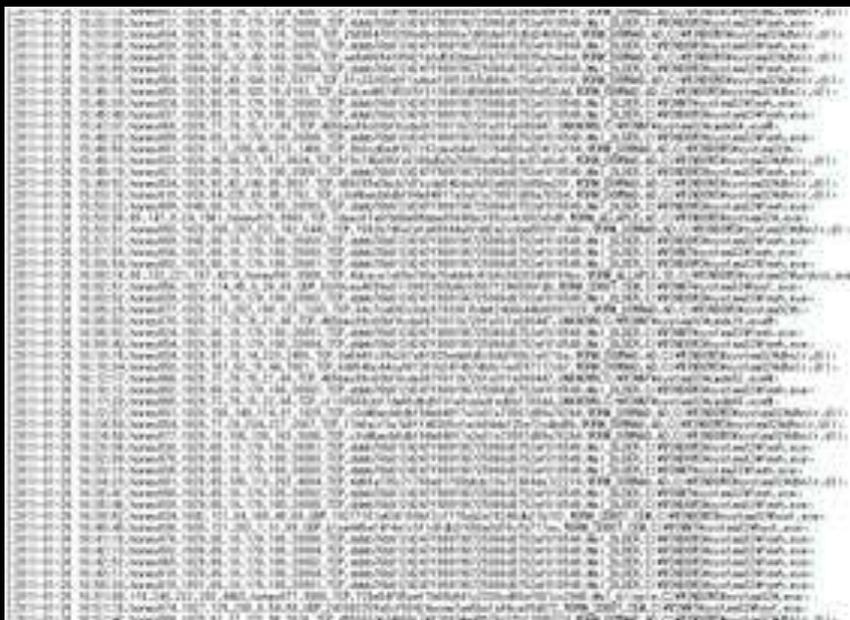
宮崎清隆さん(JPCERT)

# CCC DATASET 2011 - 攻撃通信データ

No.	Time	Source	Destination	Protocol	Length	Info
7366	7863.13839	10.10.110.11.11.11	10.10.110.11.11.11	DNS	82	standard query A blah.swapetfame.com
7367	7863.14064	10.10.110.11.11.11	10.10.110.11.11.11	DNS	98	standard query response A 183.104.31.22x
7368	7863.16075	10.10.110.11.11.11	104.78CF	TCP	84	ftp->frs > 7878 [CON] seq=0 win=64240 len=0 wsc
7369	7863.16075	10.10.110.11.11.11	104.78CF	TCP	84	ftp->frs > 7878 [CON] seq=0 win=64240 len=0 wsc
7370	7863.18226	10.10.110.11.11.11	104.78CF	TCP	36	initial-req > 8024 [FIN, ACK] seq=33 ack=210 win
7371	7863.18226	10.10.110.11.11.11	104.78CF	TCP	36	initial-req > 8024 [FIN, ACK] seq=33 ack=210 win
7372	7863.43187	10.10.110.11.11.11	104.78CF	TCP	84	7878 > ftp->frs [CON, ACK] seq=0 ack=2 win=188
7373	7863.43187	10.10.110.11.11.11	104.78CF	TCP	84	7878 > ftp->frs [CON, ACK] seq=0 ack=2 win=188
7374	7863.43187	10.10.110.11.11.11	104.78CF	TCP	36	ftp->frs > 7878 [ACK] seq=3 ack=1 win=64240 len
7375	7863.43187	10.10.110.11.11.11	104.78CF	TCP	36	ftp->frs > 7878 [ACK] seq=3 ack=1 win=64240 len
7376	7863.43187	10.10.110.11.11.11	104.78CF	TCP	108	ftp->frs > 7878 [FIN, ACK] seq=1 ack=3 win=64240
7377	7863.43187	10.10.110.11.11.11	104.78CF	TCP	108	ftp->frs > 7878 [FIN, ACK] seq=1 ack=3 win=64240
7378	7863.54679	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Request, FID: 0x4000, FATH: 0x
7379	7863.54679	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Response, FID: 0x4000
7380	7863.54679	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Request, FID: 0x4000, FATH: 0x
7381	7863.54679	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Response, FID: 0x4000
7382	7863.72282	10.10.110.11.11.11	104.78CF	TCP	82	7878 > ftp->frs [ACK] seq=0 ack=31 win=5848 len
7383	7863.72282	10.10.110.11.11.11	104.78CF	TCP	82	7878 > ftp->frs [ACK] seq=0 ack=31 win=5848 len
7384	7863.72282	10.10.110.11.11.11	104.78CF	TCP	120	7878 > ftp->frs [FIN, ACK] seq=1 ack=31 win=0 len
7385	7863.72282	10.10.110.11.11.11	104.78CF	TCP	120	7878 > ftp->frs [FIN, ACK] seq=1 ack=31 win=0 len
7386	7863.90884	10.10.110.11.11.11	104.78CF	TCP	36	ftp->frs > 7878 [ACK] seq=31 ack=61 win=64176 len
7387	7863.90884	10.10.110.11.11.11	104.78CF	TCP	36	ftp->frs > 7878 [ACK] seq=31 ack=61 win=64176 len
7388	7863.90884	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Request, FID: 0x4000, FATH: 0x
7389	7863.90884	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Response, FID: 0x4000
7390	7863.90884	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Request, FID: 0x4000, FATH: 0x
7391	7863.90884	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Response, FID: 0x4000
7392	7863.90884	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Request, FID: 0x4000, FATH: 0x
7393	7863.90884	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Response, FID: 0x4000
7394	7863.90884	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Request, FID: 0x4000, FATH: 0x
7395	7863.90884	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Response, FID: 0x4000
7396	7863.90884	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Request, FID: 0x4000, FATH: 0x
7397	7863.90884	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Response, FID: 0x4000
7398	7863.90884	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Request, FID: 0x4000, FATH: 0x
7399	7863.90884	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Response, FID: 0x4000
7400	7863.90884	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Request, FID: 0x4000, FATH: 0x
7401	7863.90884	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Response, FID: 0x4000
7402	7863.90884	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Request, FID: 0x4000, FATH: 0x
7403	7863.90884	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Response, FID: 0x4000
7404	7863.90884	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Request, FID: 0x4000, FATH: 0x
7405	7863.90884	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Response, FID: 0x4000
7406	7863.90884	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Request, FID: 0x4000, FATH: 0x
7407	7863.90884	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Response, FID: 0x4000
7408	7863.90884	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Request, FID: 0x4000, FATH: 0x
7409	7863.90884	10.10.110.11.11.11	104.78CF	TCP	180	4r create Andx Response, FID: 0x4000

- サーバ型ハニーポット2台のネットワークキャプチャデータ
- Windows XP SP1+α
- 2010年8月18日～31日, 2011年1月18日～31日
- 2300万パケット、3.8GB
- 用途：感染手法の検知ならびに解析技術

# CCC DATAsset 2011 - 攻撃元データ



- ハニーポット72台のマルウェア取得時のログ
- 16万件（TCP：14万件）
- 1.2万種（ハッシュ）、約300種（ウイルス名称）の検体

用途：ボットの活動傾向把握技術

ログ項目	例（一部を*でマスク）
マルウェア検体の取得時刻	2011-01-14 18:20:01
送信元IPアドレス	honey016
送信元ポート番号	1029
宛先IPアドレス	**.*.179.100
宛先ポート番号	20000
TCPまたはUDP	TCP
マルウェア検体のハッシュ値（SHA1）	*****6b8124247f988f96725066d3752ef018549
ウイルス名称	Mal_DLDER
ファイル名	C:\¥WINNT¥system32¥fewh.exe

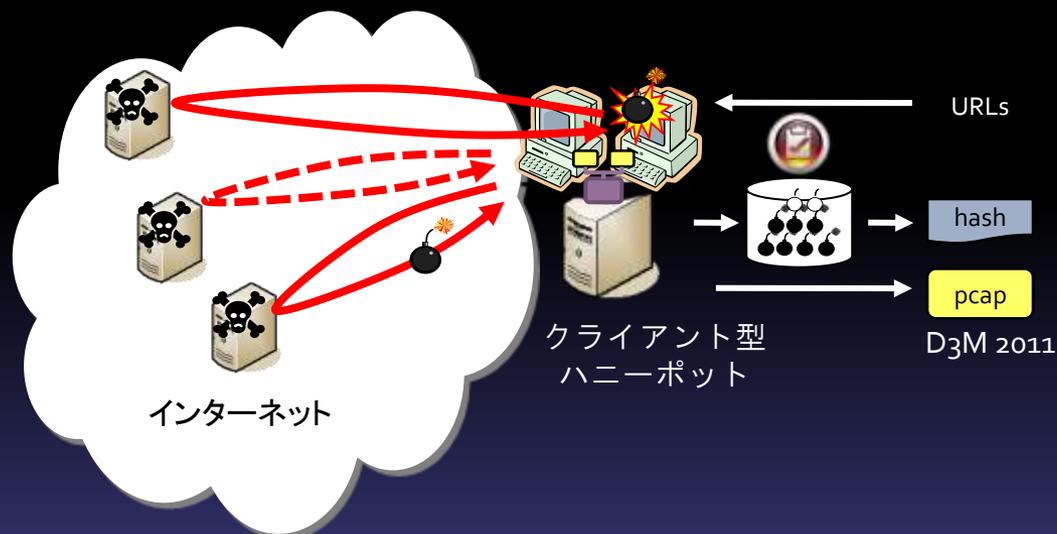
# 攻撃元データの基本情報

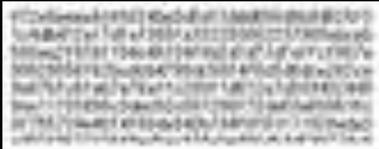
項目	件数
全レコード数	158,734
TCPによるダウンロードレコード数	136,251
UDPによるダウンロードレコード数	22,483
ダウンロードホストIPアドレス種類数	89,122
マルウェア検体のハッシュ値種類数	12,591
ウイルス名称種類数（UNKNOWN含まない）	316

# CCC DATAsEt 2008 - 2011比較

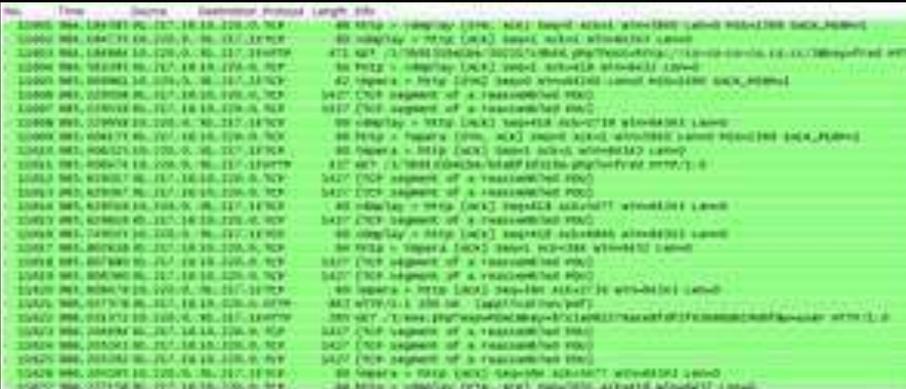
項目	2008	2009	2010	2011
マルウェア検体 (ハッシュ値)				
検体数	1	10	50	50
選定条件	多機能, 解読困難	解析結果あり, 関連性のある複数検体, 特徴的な機能	解析結果あり, 特徴的な機能, 2010年1月～3月に収集した未知検体	解析結果あり, 特徴的な機能, 2011年1月に収集した未知検体等
攻撃通信データ (パケットキャプチャ)				
ハニーポット	honey001, honey002	honey003, honey004	honey001, honey002	honey001, honey002
収集日	2008/4/28, 2008/4/29	2009/3/13, 2009/3/14	2010/3/5～2010/3/11	2010/8/18～2010/8/31, 2011/1/18～2011/1/31
パケット数	15,901,943	3,511,850	22,486,674	23,009,309
攻撃元データ (ログ)				
ハニーポット数	112台	94台	92台	72台
ハニーポットID	なし (ダウンロードホストと通信方向のみ)	あり	あり	あり
収集期間	6ヶ月間 2007/11/1～2008/4/30	1年間 2008/5/1～2009/4/30	1年間 2009/5/1～2010/4/30	9ヶ月間 2010/5/1～2011/1/31
レコード数	2,942,221	2,470,766	1,162,093	158,734

- クライアント型ハニーポット(Marionette)で収集したWeb感染型マルウェアに関するデータ群
- 用途：マルウェア解析技術、感染手法の検知技術の研究





- マルウェア検体
  - 34検体のハッシュ値



- 攻撃通信データ
  - ハニーポット10台
  - Windows XP SP2
  - 2011年2月8、14、16日
  - 入り口URL  
malwaredomainlist.com
  - 各日で65、100、118個
  - 33MB



# MWSにおける研究用データセットの利用実績

		2008	2009	2010	2011
CCC DATAsset	マルウェア検体	5	7	6	5
	攻撃通信データ	9	14	5	6
	攻撃元データ	8	6	5	4
MARS				1	1
D3M				4	3
総括			1	1	1
合計		22 (8)	28 (15)	22 (10)	20 (9)

# 今後の課題と対策

## ■ 課題

- 最新の脅威を捉えた研究用データセットの収集・作成・蓄積や利用環境の構築・提供など包括的なフレームワーク

## ■ 対策

- 情報処理学会コンピュータセキュリティ研究会の配下にMWS組織委員会を設立（2011.7）  
<http://www.iwsec.org/mws/2011/committee.html>
- 年次のMWS開催とは別に、中長期的な視点でマルウェア対策研究、人材育成に取り組む

# 参考文献

- MIT Lincoln Laboratory, DARPA Intrusion Detection Evaluation Data Sets, <http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/index.html>
- B. Sangster, et al.: Toward Instrumenting Network Warfare Competitions to Generate Labeled Datasets, 18th USENIX Security Symposium CSET'09 (2009.08)
- BADGERS2011: Building Analysis Datasets and Gathering Experience Returns for Security, <http://iseclab.org/badgers2011/> (2011.04)
- PREDICT: the Protected Repository for the Defense of Infrastructure Against Cyber Threats, <https://www.predict.org/>
- サイバークリーンセンター, <https://www.ccc.go.jp/>
- 畑田充弘, 他: マルウェア対策のための研究用データセット ～MWS 2010 Datasets～, CSS2010(MWS2010) (2010.10)
- マルウェア対策研究人材育成ワークショップ2011, <http://www.iwsec.org/mws/2011/>
- Mitsuaki Akiyama, 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 (2010.05)