2010-10-19(Tue)
マルウェア対策研究人材育成ワークショップ(MWS2010)

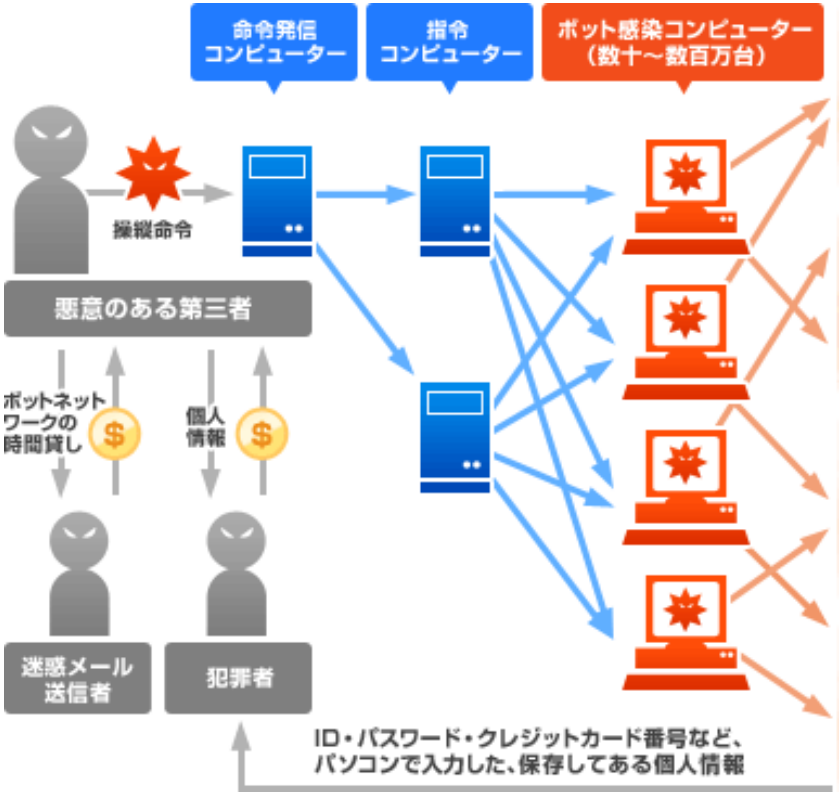
マルウェア対策のための研究用データセット ～MWS 2010 Datasets～

畑田充弘（NTTコミュニケーションズ）、中津留勇（JPCERT/CC）、
秋山満昭（NTT PF研）、三輪信介（NICT）

目次

- 課題と目的
- MWS 2010 Datasets
  - CCC DATAset 2010
  - MARS for MWS 2010
  - D3M 2010
- 実績（速報）
- おわりに

課題



(https://www.ccc.go.jp/bot/index.html)

- 共通の教材がないこと
- 研究用データを入手すること自体が難しくなっていること
- DARPAの最新は2001年、CDX Datasetsは2009年だがマルウェアによる攻撃ではない

目的

- 研究用データセットの提供
  - CCC DATAset 2008
  - CCC DATAset 2009

- 研究成果を共有する場、切磋琢磨する環境作り
  - MWS 2008
  - MWS 2009

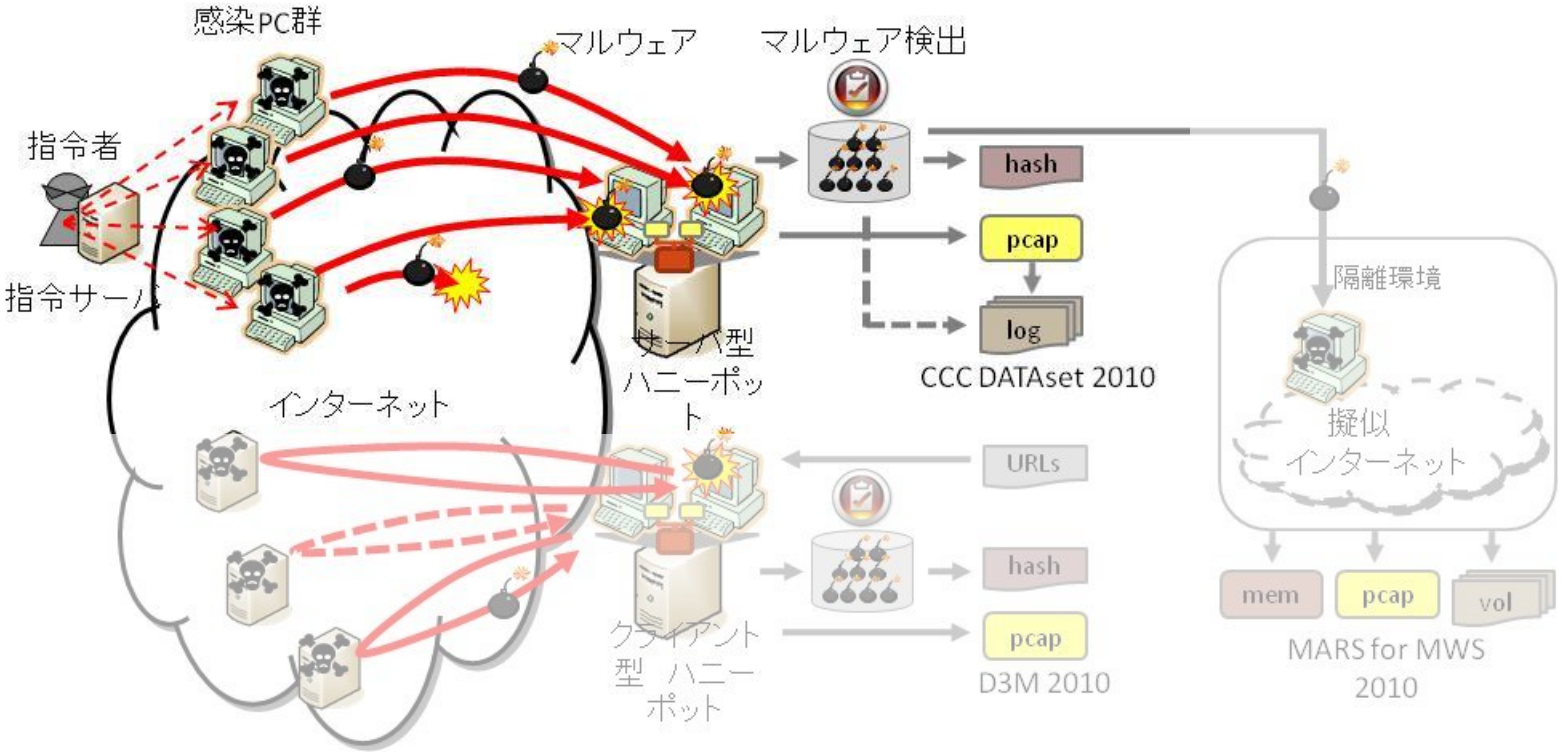
MWS 2010 Datasets

- CCC DATAset 2010 （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
- MARS for MWS 2010 （CSEC研究会/NICT）
  - CCC DATAset 2010のマルウェア検体を擬似環境で動作させて得た一次解析結果のデータ群
- D3M 2010 （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 DATAset 2010

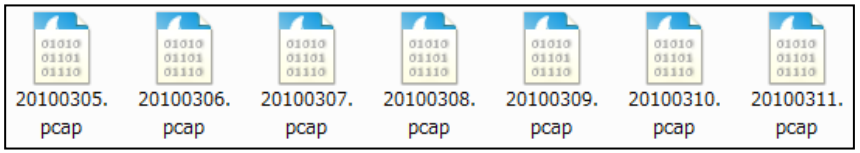
- マルウェア検体
- 攻撃通信データ
- 攻撃元データ

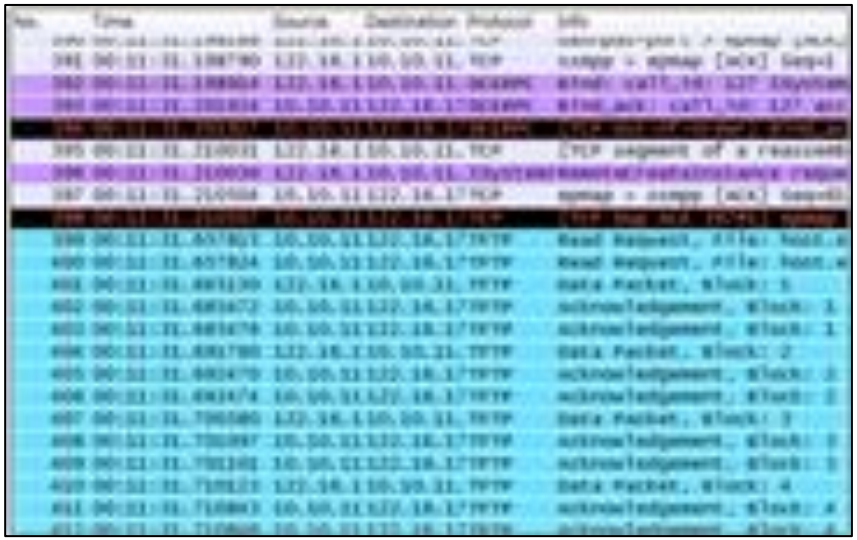


CCC DATAset 2010 – マルウェア検体

- 50検体のハッシュ値
- 解析結果を照合できる検体：10検体
- 未知検体：40検体

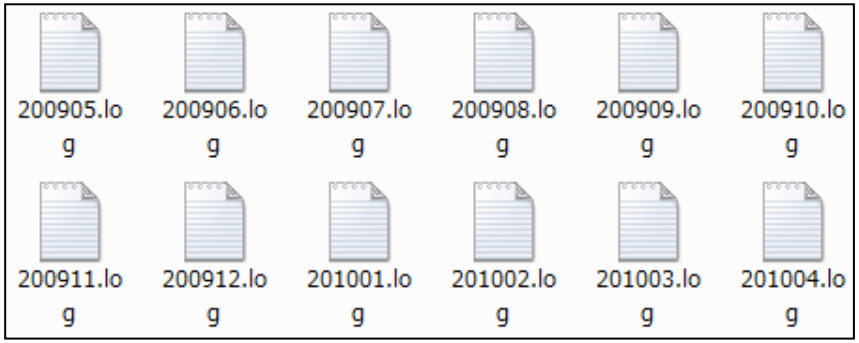
CCC DATAset 2010 – 攻撃通信データ





- サーバ型ハニーポット2台のパケットキャプチャデータ
- Windows XP SP1+α
- 2010年3月5日～11日
- 2200万パケット、3.5GB

CCC DATAset 2010 – 攻撃元データ



- ハニーポット92台のマルウェア取得時のログ
- 116万件（TCP：105万件）
- 3万種（ハッシュ）、1000種（ウイルス名称）の検体

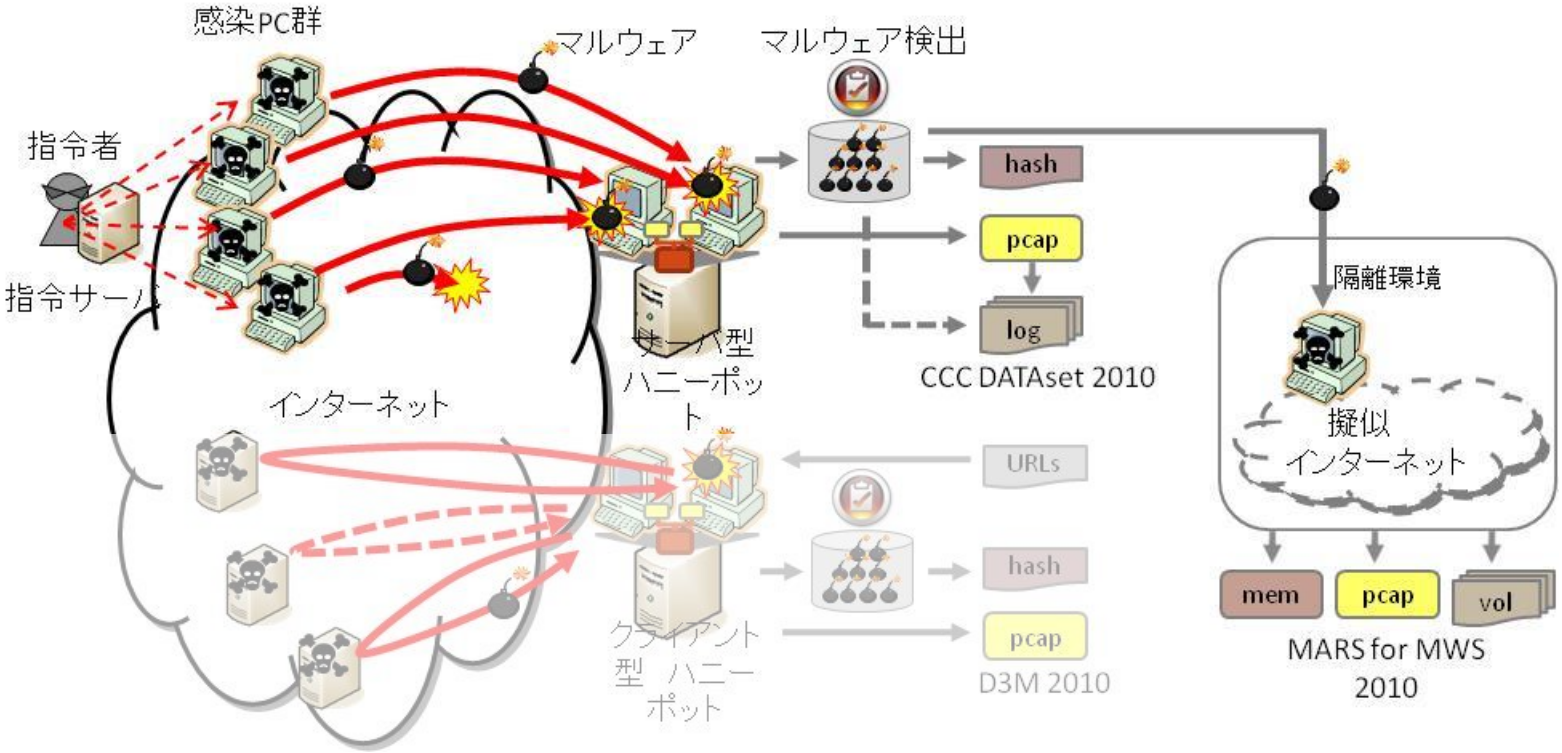
ログ項目	例（一部を*でマスク）
マルウェア検体の取得時刻	2010-03-05 03:02:41
送信元IPアドレス	honey001
送信元ポート番号	1028
宛先IPアドレス	**. 243.167
宛先ポート番号	5824
TCPまたはUDP	TCP
マルウェア検体のハッシュ値（SHA1）	*****bc3c850cf68a39c9e8013f2169d408a9d90
ウイルス名称	WORM_DOWNAD.AD
ファイル名	C:¥WINDOWS¥system32¥dhnlr.dll

CCC DATAset 2010 – 2008/2009比較

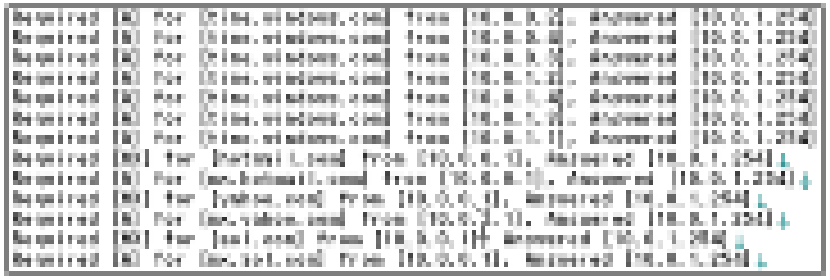
項目	2008	2009	2010
マルウェア検体			
検体数	1	10	50
選定条件	多機能，解読困難	解析結果あり，関連性のある複数検体，特徴的な機能	解析結果あり，特徴的な機能，2010年1月～3月に収集した未知検体
攻撃通信データ			
ハニーポット	honey001，honey002	honey003, honey004	honey001，honey002
収集日	2008/4/28，2008/4/29	2009/3/13，2009/3/14	2010/3/5～2010/3/11
総パケット数	15,901,943	3,511,850	22,486,674
攻撃元データ			
ハニーポット数	112台	94台	92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
収集期間	2007/11/1～2008/4/30	2008/5/1～2009/4/30	2009/5/1～2010/4/30
全レコード数	2,942,221	2,470,766	1,162,093

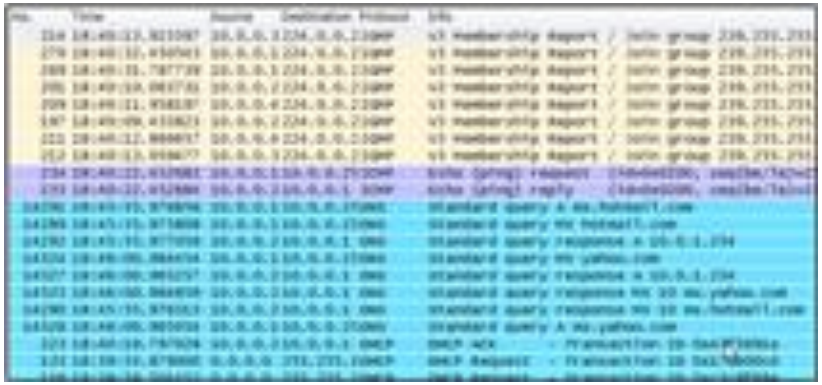
MARS for MWS 2010

- CCC DATAset 2010のマルウェア検体を擬似環境で動作させて得た一次解析結果のデータ群
- 解析結果の応用によるマルウェア解析技術の研究



MARS for MWS 2010

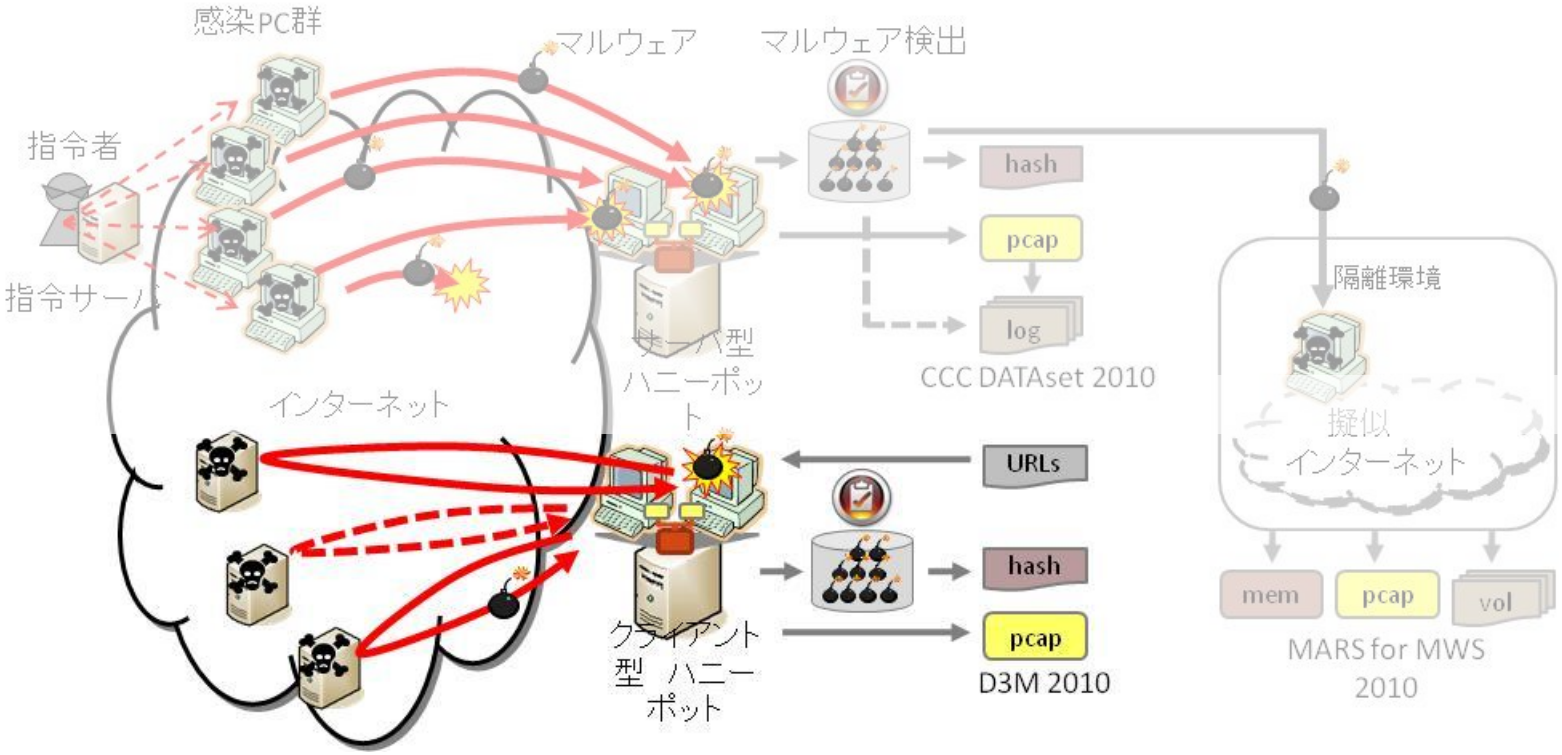




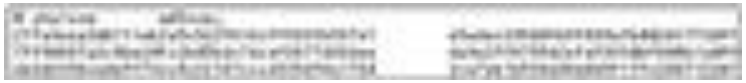
- 分類情報
  - 検体ファイル名、ハッシュ値
- 動的解析結果
  - 検体動作時のメモリダンプ、パケットダンプ、模倣DNSへのアクセス記録
- 静的解析結果
  - strings結果、メモリダンプに対するVolatility Frameworkによる解析結果

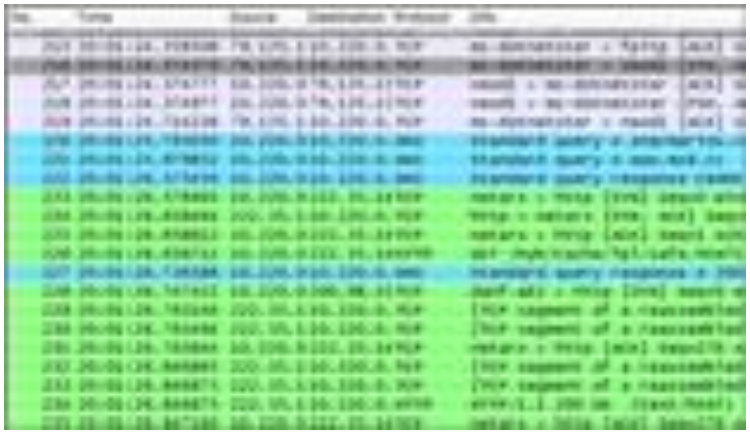
D3M 2010

- 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
- マルウェア解析技術、感染手法の検知技術の研究



D3M 2010





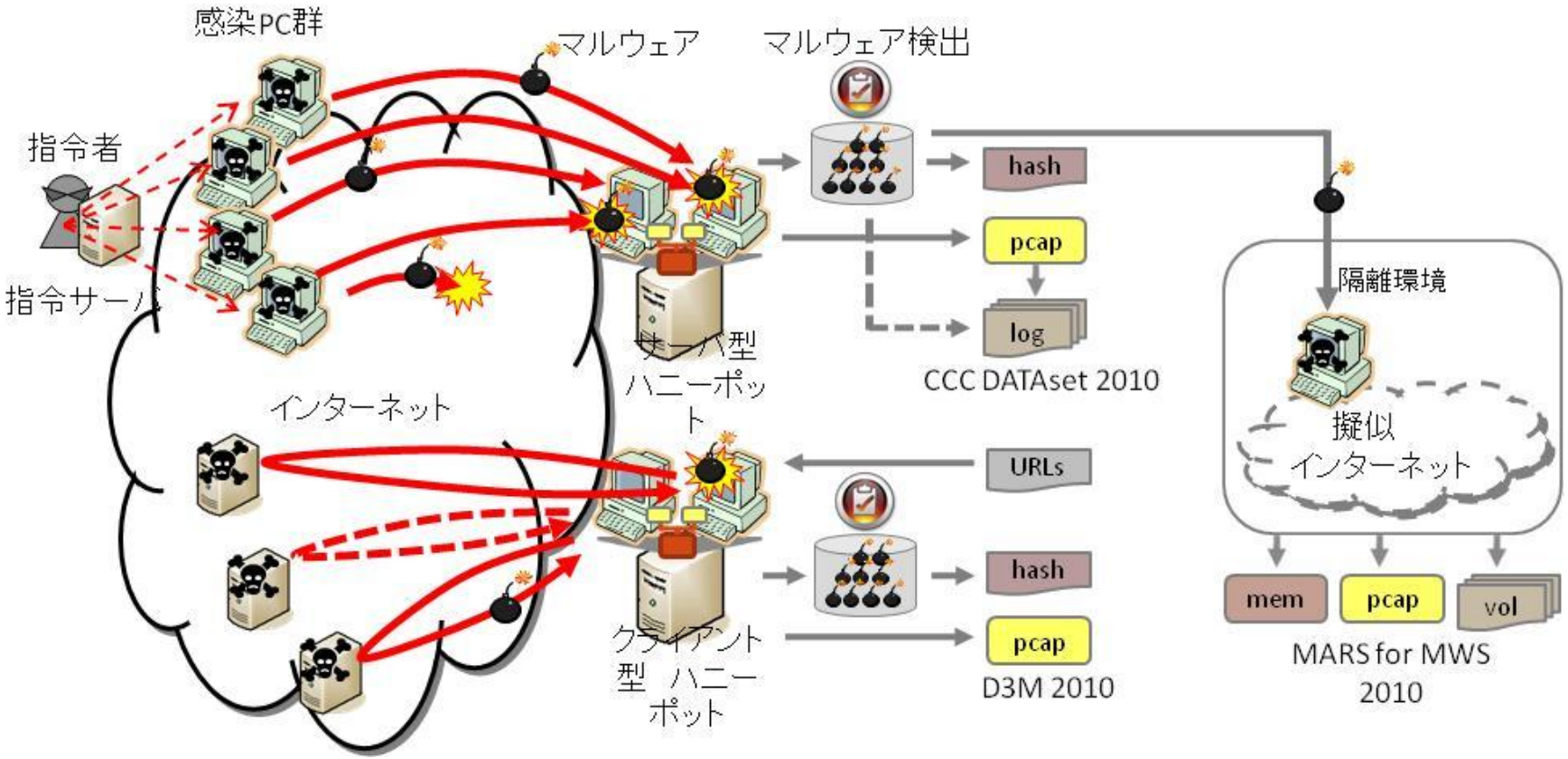
- マルウェア検体
  - 3検体のハッシュ値
  - Gumblar：8080系
- 攻撃通信データ
  - Windows XP SP2
  - 2010年3月8、9、11日
  - 入り口URL
    malwaredomainlist.com
    - 各日で205、180、172個
  - 133MB

研究用データセットの利用実績

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

今後の課題

- データ種類、収集環境の網羅性、期間、運用/構成情報
- 正常データも



MWS201x　これからが本番

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)
- 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とワークショップを通じた研究成果の共有, 情報処理学会シンポジウムシリーズ, Vol.2009, No.11, CSS2009(MWS2009), pp.1-8 (2009.10)
- 畑田充弘: 研究用データセットを用いたマルウェア対策研究人材育成ワークショップ, 情報処理, Vol.51, No.3, pp.284-287 (2010.03)
- 竹森敬祐, 他: MWS Cup 2009, 情報処理, Vol.51, No.3, pp.296-299 (2010.03)
- マルウェア対策研究人材育成ワークショップ, http://www.iwsec.org/mws/2010/
- 三輪信介, 他: 小規模攻撃再現テストベッドによる動作記録データセットの生成, 情報処理学会シンポジウムシリーズ, Vol.2009, No.11, CSS2009(MWS2009), pp.931-936 (2009.10)
- N. Petroni, et al: FATKit: A Framework for the Extractionand Analysis of Digital Forensic Data from Volatile System Memory, Digital Investigation Journal 3(4) (2006.12)
- 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)