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

NTTコミュニケーションズ株式会社 畑田充弘

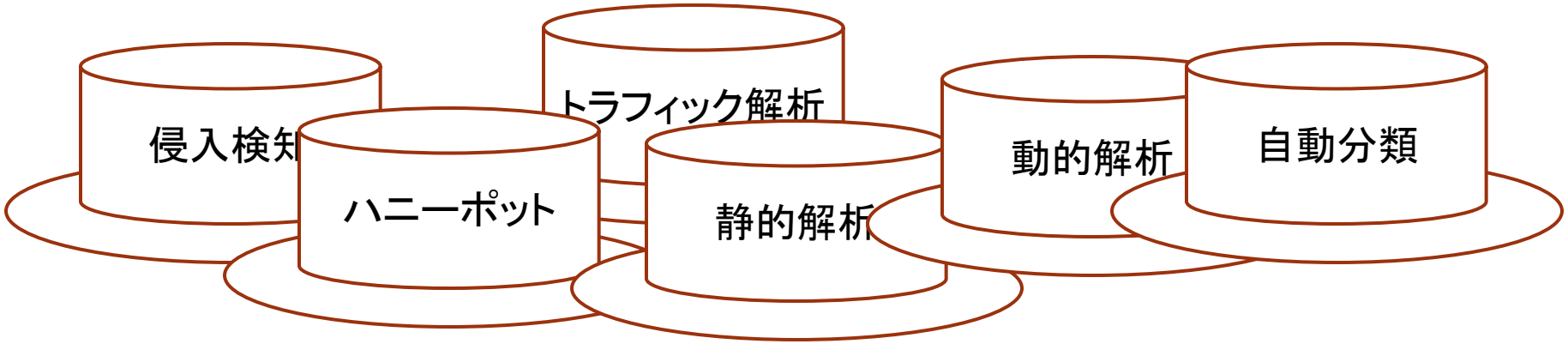
もくじ

- 背景と目的
- 研究用データセット
  - ～ CCC DATAset 2008 ～
- 研究成果の共有
  - ～ MWS 2008 ～
- 研究用データセット
  - ～ CCC DATAset 2009 ～
- 研究用データセットの要件と課題
- まとめ

はじめに

- 複雑化するマルウェアの脅威

- 多岐にわたる対策研究

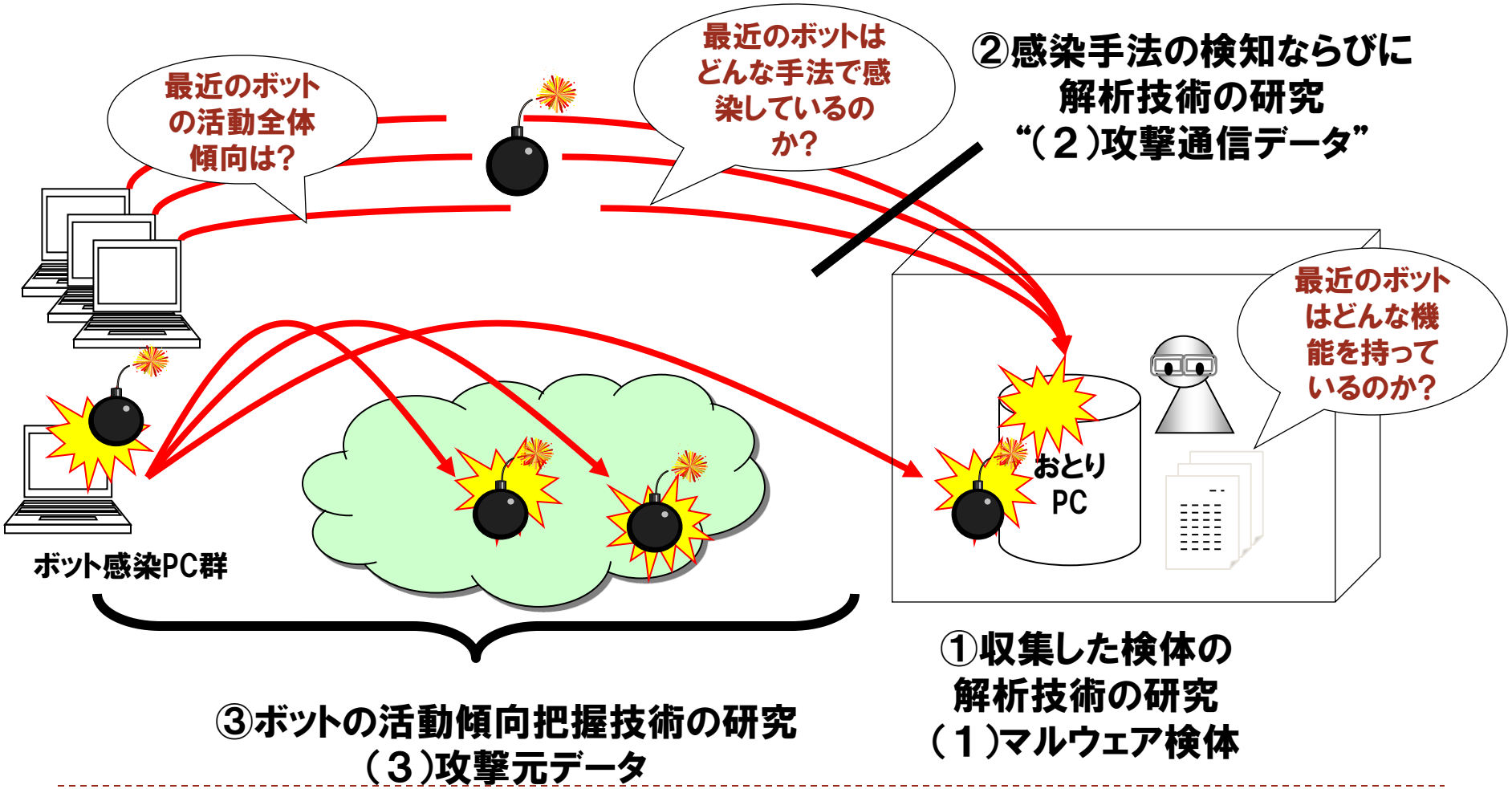


背景と目的

- マルウェア対策研究に用いられる評価用データ
  - 個別のハニーポットでデータ収集
  - 個別の実験環境で解析
- 各提案手法の客観的な評価が困難
  - 侵入検知の研究のための評価用トラフィックデータの公開
    - DARPA Intrusion Detection Evaluation Data Sets
      - 1998/1999/2000年
      - 学習用、学習後の検証用
  - この10年、、、
- サイバークリーンセンター（CCC）からのデータセット
- データセットを提供して成果を共有するワークショップ

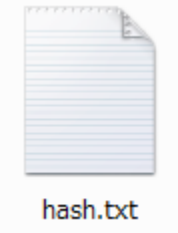
研究用データセット　～ CCC DATAset 2008 ～

- 想定研究と各データ



研究用データセット　～ CCC DATAset 2008 ～

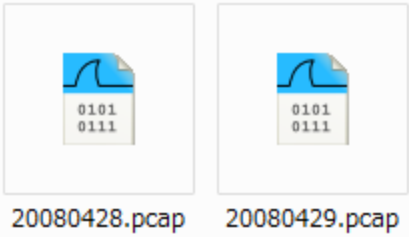
- （1）マルウェア検体
  - ハニーポットで収集したマルウェア検体のハッシュ値1件
  - 機能が豊富であり、耐解析性が高いという方針で選定




```
sha1sum: 805cb80b                                        7
md5sum:  aab9cb39
```

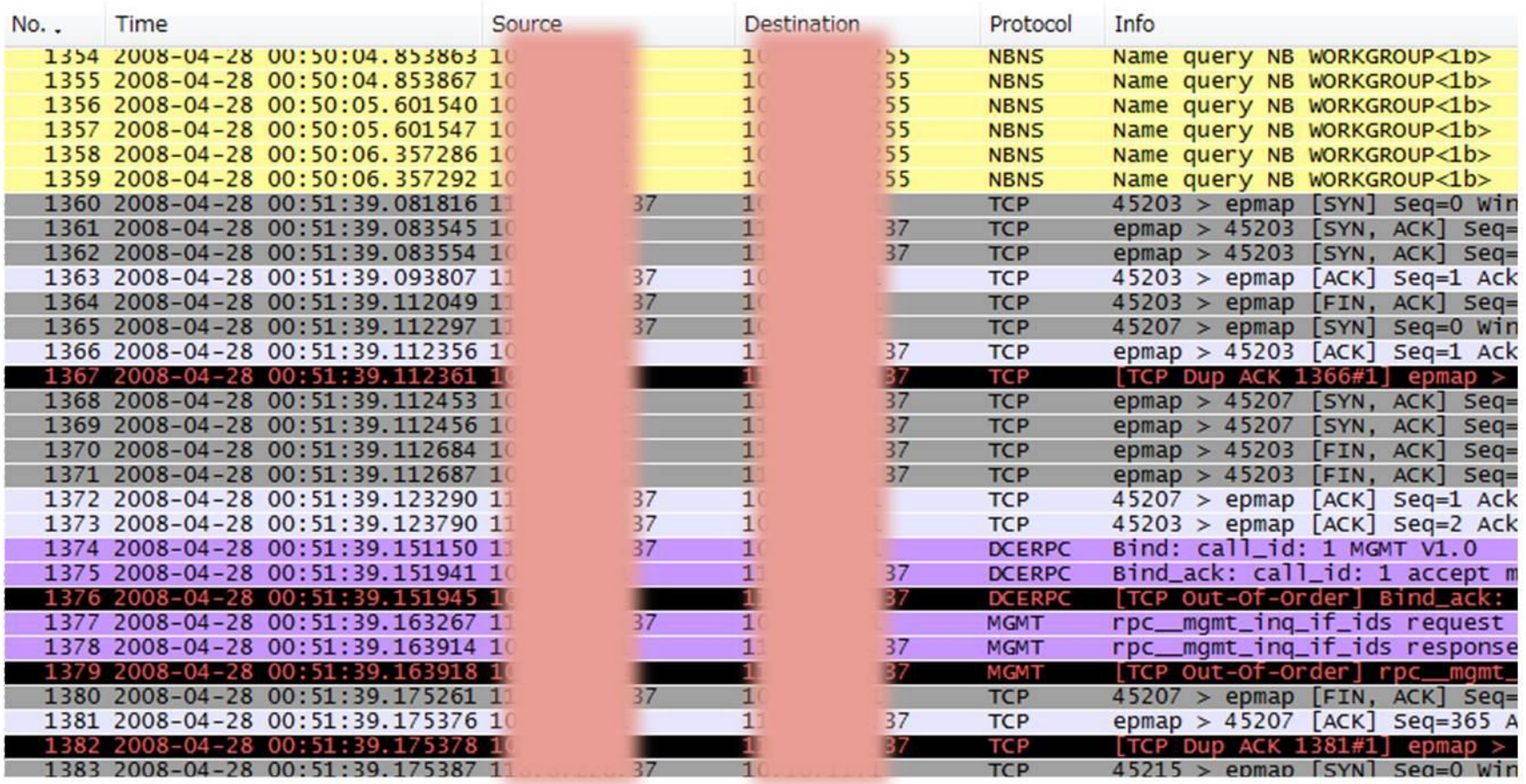
研究用データセット　～ CCC DATAset 2008 ～

- （2）攻撃通信データ
  - ハニーポット（2台）の通信をホストOS上でtcpdumpしたpcap
  - ハニーポットのOS（ゲストOS）
    - Windows 2000
    - Windows XP
  - ネットワーク接続環境
    - FTTH、動的IPアドレス
    - それぞれのハニーポットに物理IF割当
  - データ収集期間
    - 2008年4月28日／4月29日
  - 総パケット数：15,901,943
  - 約2.8GB



研究用データセット　～ CCC DATAset 2008 ～

- （2）攻撃通信データ

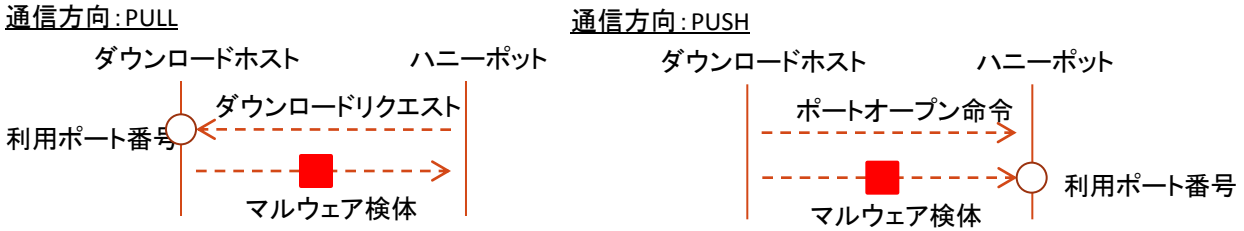


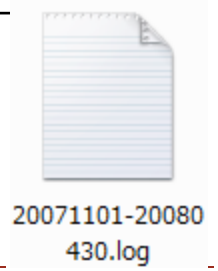
No.	Time	Source	Destination	Protocol	Info
1354	2008-04-28 00:50:04.853863	1(	1(55	NBNS	Name query NB WORKGROUP<1b>
1355	2008-04-28 00:50:04.853867	1(	1(55	NBNS	Name query NB WORKGROUP<1b>
1356	2008-04-28 00:50:05.601540	1(	1(55	NBNS	Name query NB WORKGROUP<1b>
1357	2008-04-28 00:50:05.601547	1(	1(55	NBNS	Name query NB WORKGROUP<1b>
1358	2008-04-28 00:50:06.357286	1(	1(55	NBNS	Name query NB WORKGROUP<1b>
1359	2008-04-28 00:50:06.357292	1(	1(55	NBNS	Name query NB WORKGROUP<1b>
1360	2008-04-28 00:51:39.081816	1	37 1(	TCP	45203 > epmap [SYN] Seq=0 Win
1361	2008-04-28 00:51:39.083545	1(	1	37 TCP	epmap > 45203 [SYN, ACK] Seq=
1362	2008-04-28 00:51:39.083554	1(	1	37 TCP	epmap > 45203 [SYN, ACK] Seq=
1363	2008-04-28 00:51:39.093807	1	37 1(	TCP	45203 > epmap [ACK] Seq=1 Ack
1364	2008-04-28 00:51:39.112049	1	37 1(	TCP	45203 > epmap [FIN, ACK] Seq=
1365	2008-04-28 00:51:39.112297	1	37 1(	TCP	45207 > epmap [SYN] Seq=0 Win
1366	2008-04-28 00:51:39.112356	1(	1	37 TCP	epmap > 45203 [ACK] Seq=1 Ack
1367	2008-04-28 00:51:39.112361	1(	1	37 TCP	[TCP Dup ACK 1366#1] epmap >
1368	2008-04-28 00:51:39.112453	1(	1	37 TCP	epmap > 45207 [SYN, ACK] Seq=
1369	2008-04-28 00:51:39.112456	1(	1	37 TCP	epmap > 45207 [SYN, ACK] Seq=
1370	2008-04-28 00:51:39.112684	1(	1	37 TCP	epmap > 45203 [FIN, ACK] Seq=
1371	2008-04-28 00:51:39.112687	1(	1	37 TCP	epmap > 45203 [FIN, ACK] Seq=
1372	2008-04-28 00:51:39.123290	1	37 1(	TCP	45207 > epmap [ACK] Seq=1 Ack
1373	2008-04-28 00:51:39.123790	1	37 1(	TCP	45203 > epmap [ACK] Seq=2 Ack
1374	2008-04-28 00:51:39.151150	1	37 1(	DCERPC	Bind: call_id: 1 MGMT V1.0
1375	2008-04-28 00:51:39.151941	1(	1	37 DCERPC	Bind_ack: call_id: 1 accept m
1376	2008-04-28 00:51:39.151945	1(	1	37 DCERPC	[TCP Out-Of-Order] Bind_ack:
1377	2008-04-28 00:51:39.163267	1	37 1(	MGMT	rpc__mgmt_inq_if_ids request
1378	2008-04-28 00:51:39.163914	1(	1	37 MGMT	rpc__mgmt_inq_if_ids response
1379	2008-04-28 00:51:39.163918	1(	1	37 MGMT	[TCP Out-Of-Order] rpc__mgmt_
1380	2008-04-28 00:51:39.175261	1	37 1(	TCP	45207 > epmap [FIN, ACK] Seq=
1381	2008-04-28 00:51:39.175376	1(	1	37 TCP	epmap > 45207 [ACK] Seq=365 A
1382	2008-04-28 00:51:39.175378	1(	1	37 TCP	[TCP Dup ACK 1381#1] epmap >
1383	2008-04-28 00:51:39.175387	1	37 1(	TCP	45215 > epmap [SYN] Seq=0 Win

研究用データセット　～ CCC DATAset 2008 ～

- （3）攻撃元データ
  - ハニーポット112台による6ヶ月間のマルウェア取得ログ
  - データ収集日
    - 2007年11月1日～2008年4月30日

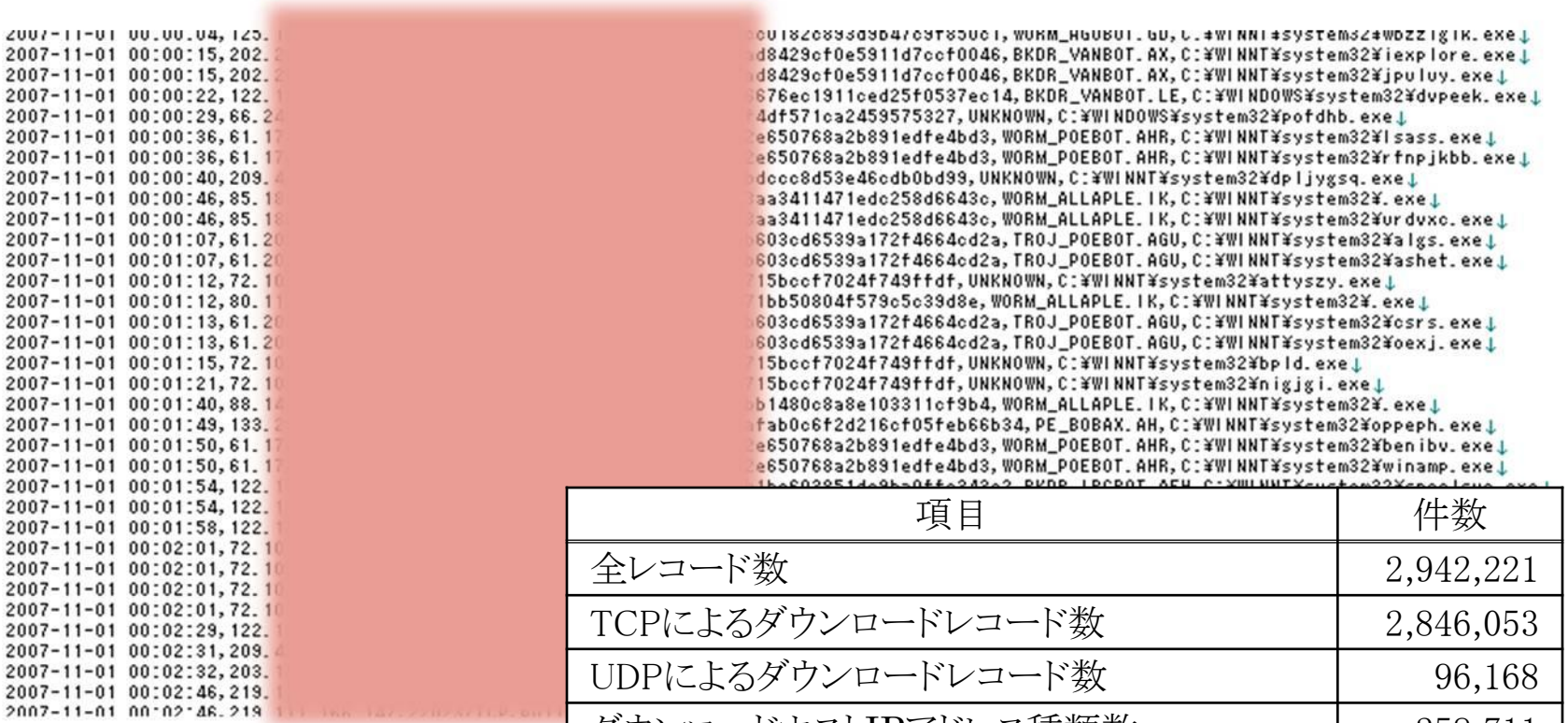
ログ項目	例（一部を*でマスク）
マルウェア検体の取得時刻	2007-11-01 00:02:01
ダウンロードホストIPアドレス	**.10.167.74
利用ポート番号／TCPまたはUDP	6251/TCP
通信方向	Pull
マルウェア検体のハッシュ値（SHA1）	*****a7e7edca3b787624c4edb6cc74d4dbd1b8f
ウイルス名称	PE_VIRUT.XV
ファイル名	C:¥WINNT¥system32¥cwgbiw.exe





研究用データセット　～ CCC DATAset 2008 ～

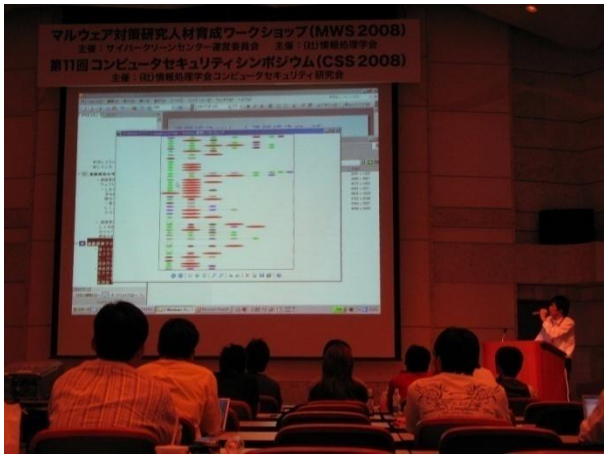
- （3）攻撃元データ



項目	件数
全レコード数	2,942,221
TCPによるダウンロードレコード数	2,846,053
UDPによるダウンロードレコード数	96,168
ダウンロードホストIPアドレス種類数	258,711
マルウェア検体のハッシュ値種類数	52,465
ウイルス名称種類数（UNKNOWN含まない）	1,081

研究成果の共有　～ MWS 2008 ～

- マルウェア対策研究人材育成ワークショップ2008
  - 日程：2008年10月8日（水）～10日（金）
  - 会場：沖縄コンベンションセンター
  - CSS2008と併催
  - 一般口頭発表22件（うち学生の部8件）

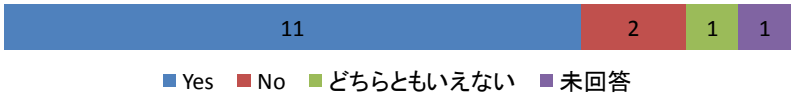


http://www.iwsec.org/mws/2008/

研究成果の共有　～ MWS 2008 ～

- 関係者アンケートから

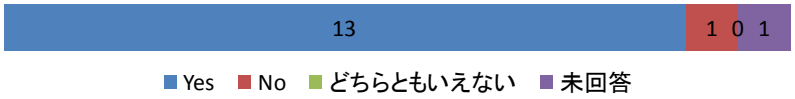
Q6.データセットにより従来実施できなかったことができたか



- ご意見（抜粋）
  - 独自に収集しているデータと比較することができ、その差異や共通する点など多くの知見を得ることができた。
  - 提案手法の有効性を評価することができ、理論を証明するための実践的なデータの必要性・重要性をあらためて感じた。
  - データセットによって、新たにマルウェアの研究を行うことができた。大学等、マルウェアのデータを取得することが難しい研究機関にとって、データセットの提供は非常に有意義なものであると思う。

7

Q7.データセットの使用により新たな研究課題の発見につながったか

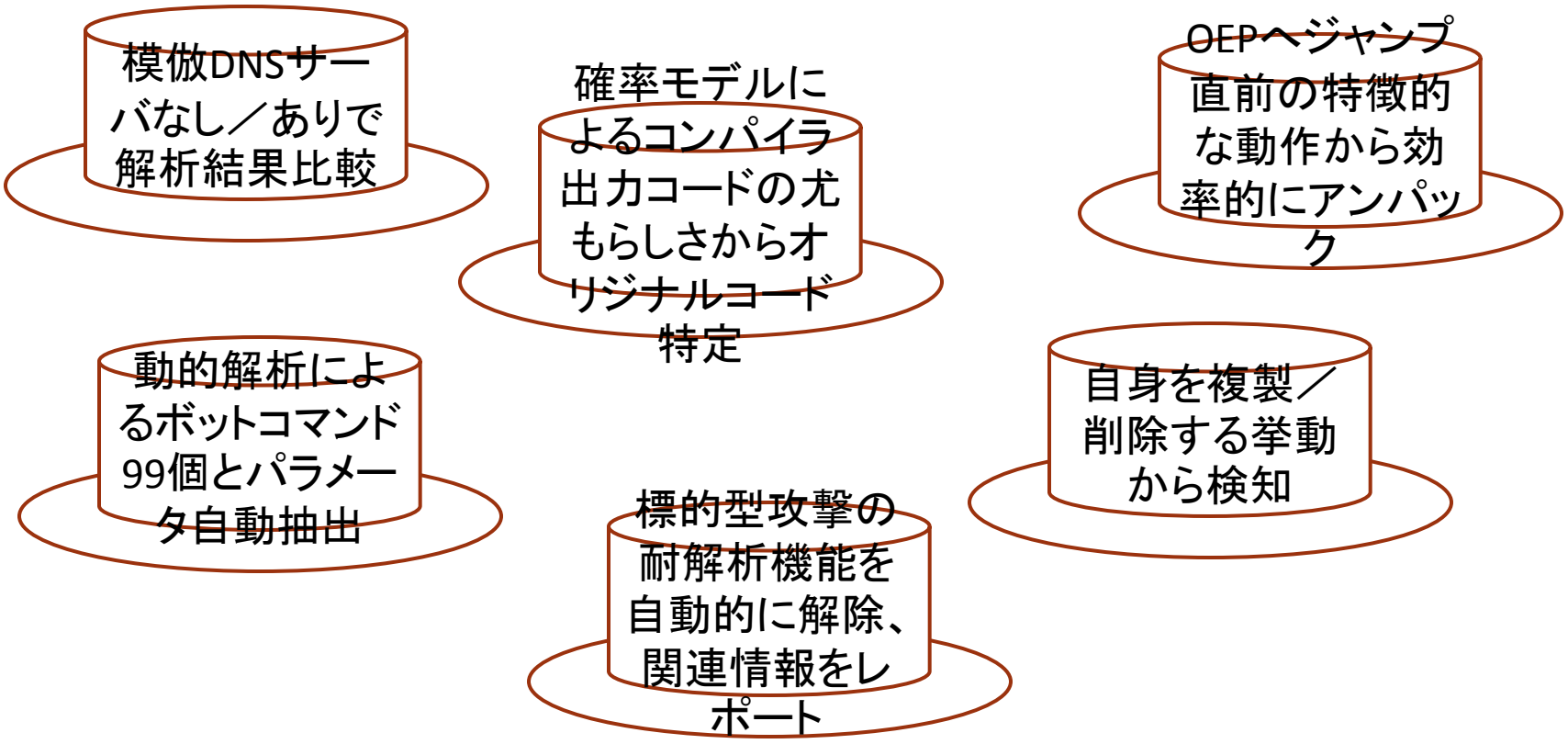


- ご意見（抜粋）
  - 異なるネットワークでは、ハッシュ値で比較する限り、収集した検体が一致する件数は思っていたよりも少ないという発表もあり、全体の傾向を知ることの困難さに課題を感じました。
  - 長期間の「攻撃元ログ」は各自で運用しているハニーポットでは収集するのが難しく新しい課題の発見などにつながる。逆に攻撃通信データのようなデータは一般に収集するのはそこまで難しくないので、そこから新しい研究課題の発見につなげるのは難しい。
  - データセットを利用することで、マルウェアの動作傾向を調査することができた。その結果から、対策手法の検討等を行うことができるため、新たな研究課題の発見につながったといえると考えている。

8

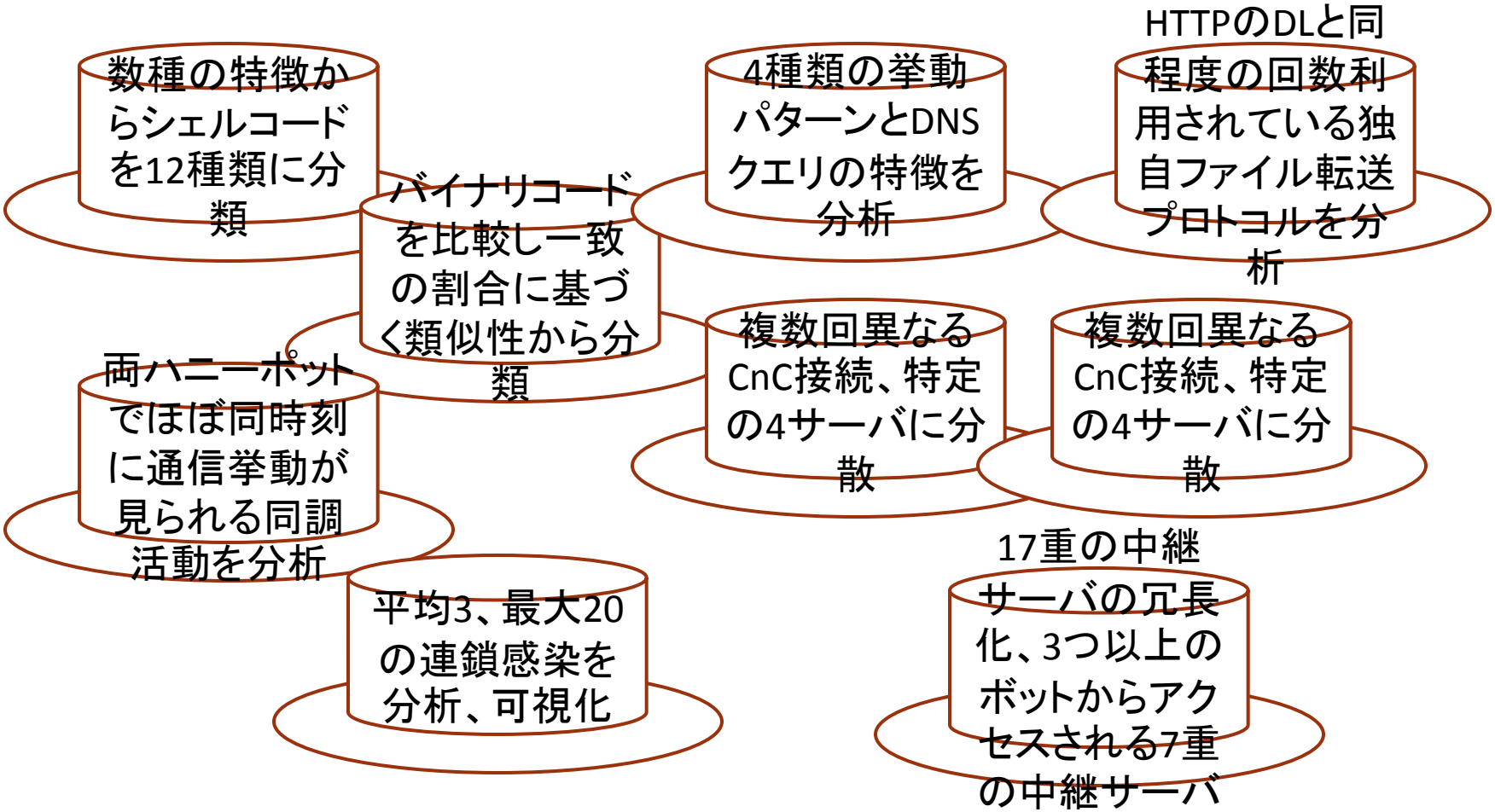
研究成果の共有　～ MWS 2008 ～

- （1）マルウェア検体　を用いた研究



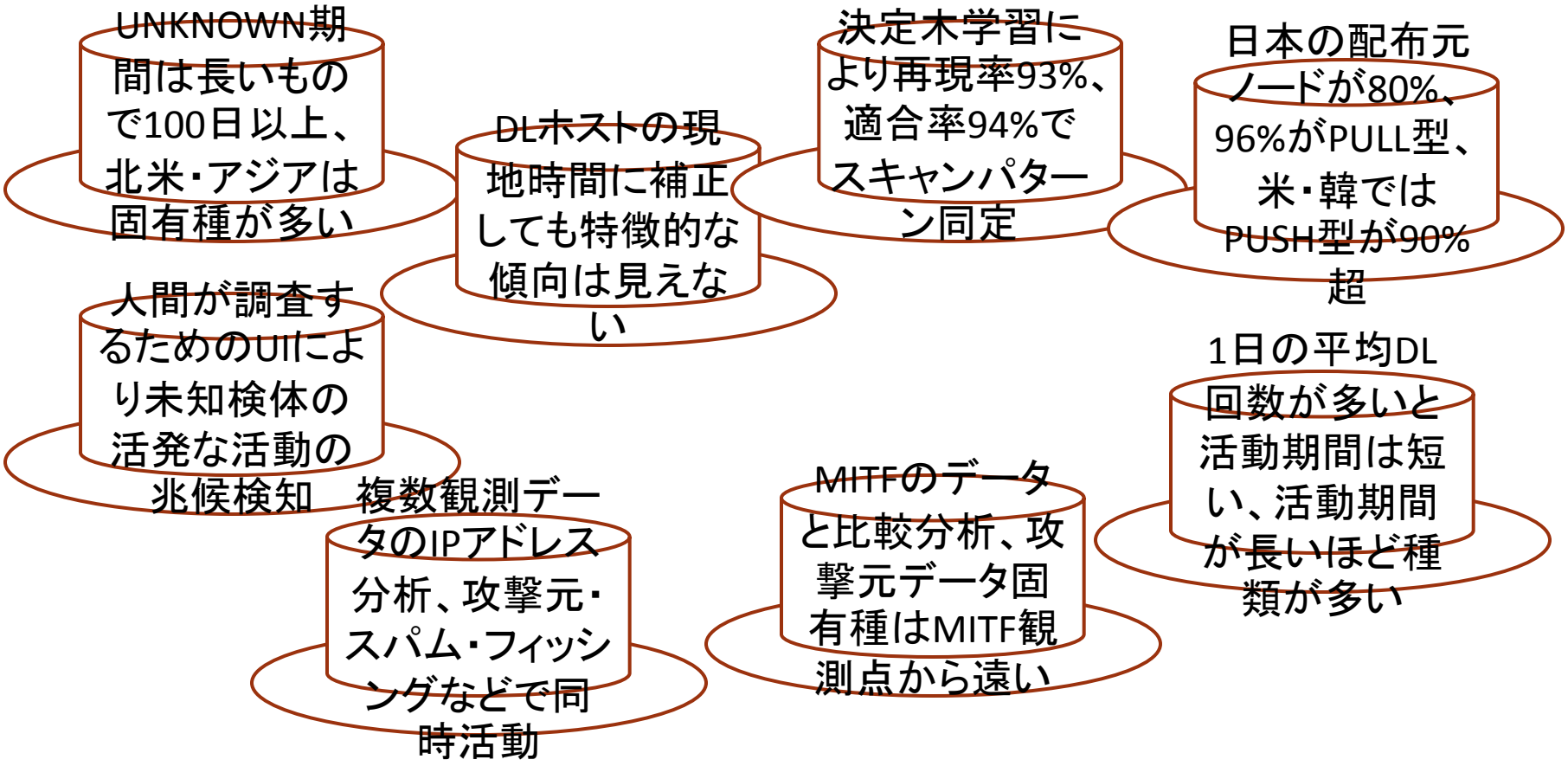
研究成果の共有　～ MWS 2008 ～

- （2）攻撃通信データ　を用いた研究



研究成果の共有　～ MWS 2008 ～

- （3）攻撃元データ　を用いた研究



研究用データセット　～ CCC DATAset 2009 ～

- MWS2008を終えての要望（関係者アンケートから）

Q9.研究を行って感じたデータセットへの要件（優先度順）

- 「マルウェア検体」について
  - 耐解析性が高い、ウイルス対策ソフトで検出できない、機能が豊富、一般に取得が困難、攻撃対象OS種類、その他（鮮度、量）
- 「攻撃通信データ」について
  - ハニーポットのグローバルIPアドレス情報、データ収集台数、データ収集期間、ハニーポットの動作特性、攻撃対象OS種類、その他（鮮度、攻撃元データとの照合）
- 「攻撃元データ」について
  - ハニーポットのグローバルIPアドレス情報（または識別子）、データ収集台数、データ収集期間、送信元・宛先のポート番号、障害による停止期間、期間中の構成変更情報、ハニーポットの動作特性、攻撃対象OS種類、その他（データの鮮度、攻撃通信データとの照合）

10

Q10.データセットとして提供されるのが望ましいデータ群

- Web感染型をはじめとして、様々な感染経路（入手経路）で得られたマルウェア検体や関連するデータを提供いただけると、その違いを調査したり、と色々興味深い研究につながると思います。
- 最近の検体はVMやデバッガを検知して活動をやめる機能を備えているため、素のPC上で検体を実際に起動させた際のPCの挙動（ファイル変化、プロセス起動、通信ポート開放、など）に関するデータが提供されていると助かります。
- 既存の解析結果にはない新たな挙動が見つかることもあるし、自分の解析力の目安にもなるので、人材育成を考慮し、「マルウェア検体」はすでに解析結果が公表されているもの、あるいは解析結果の模範解答があっても良い。

11

研究用データセット　～ CCC DATAset 2009 ～

- （1）マルウェア検体
  - ハニーポットで収集したマルウェア検体のハッシュ値10件
  - 分類（分類間で重複あり）
    - 解析結果を照合できる検体（9件）
      - 利用想定：検体の解析精度の評価
    - 関連性のある複数の検体（グループ1：3件、グループ2：2件）
      - 利用想定：検体間の関連性分析の評価
    - 特徴的な機能を有する検体（5件）
      - 利用想定：検体の特徴分析の評価

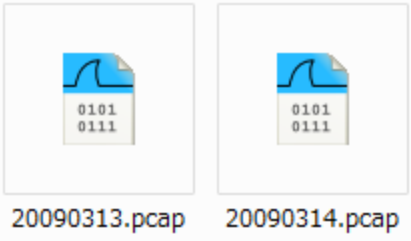
研究用データセット　～ CCC DATAset 2009 ～

- （1）マルウェア検体

```
## ハッシュ値は sha1sum md5sum の順で記載(計10種類)

■解析結果を照合できる(答え合わせができる)検体
1d23f28[redacted]1d794c
393f00c[redacted]6673e6
68ac29d[redacted]03667c
7190e43[redacted]0086d0
84e9c29[redacted]f2837f
cd9125d[redacted]4c3313
df75856[redacted]7196cc
d493913[redacted]ce23a1
fdf3bbc[redacted]461d2b

■関連性をもって解析ができる複数の検体
グループ1:
1d23f28[redacted]1d794c
393f00c[redacted]6673e6
84e9c29[redacted]f2837f

グループ2:
7190e43[redacted]0086d0
cd9125d[redacted]4c3313

■特徴的な機能を有する等、技術的に目を通しておきたい検体
68ac29d[redacted]03667c
df75856[redacted]7196cc
d493913[redacted]ce23a1
f8c19c1[redacted]0f9210
fdf3bbc[redacted]461d2b
```

研究用データセット　～ CCC DATAset 2009 ～

- （2）攻撃通信データ
  - ハニーポット（2台）の通信をホストOS上でtcpdumpしたpcap
  - ハニーポットのOS（ゲストOS）
    - Windows 2000
    - Windows XP
  - ネットワーク接続環境
    - FTTH、動的IPアドレス
    - それぞれのハニーポットに物理IF割当
  - データ収集期間
    - 2009年3月13日／3月14日
  - 総パケット数：3,511,850
  - 約580MB



研究用データセット　～ CCC DATAset 2009 ～

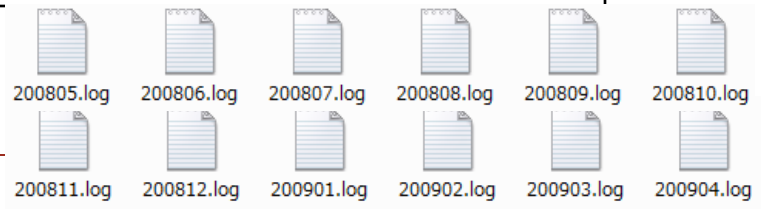
- （2）攻撃通信データ

No.	Time	Source	Destination	Protocol	Info
499	2009-03-14 00:04:58.977985		18.194	TCP	[TCP Out-of-Order] mtqp >
500	2009-03-14 00:04:58.997900	8.194	.1	TCP	4903 > mtqp [PSH, ACK] Seq
501	2009-03-14 00:04:58.997980		18.194	TCP	mtqp > 4903 [PSH, ACK] Seq
502	2009-03-14 00:04:58.997984		18.194	TCP	[TCP Out-of-Order] mtqp >
503	2009-03-14 00:04:59.018140	8.194	.1	TCP	4903 > mtqp [PSH, ACK] Seq
504	2009-03-14 00:04:59.018881		18.194	TCP	mtqp > 4903 [PSH, ACK] Seq
505	2009-03-14 00:04:59.018885		18.194	TCP	[TCP Out-of-Order] mtqp >
506	2009-03-14 00:04:59.034255	8.194	.1	TCP	4903 > mtqp [PSH, ACK] Seq
507	2009-03-14 00:04:59.034328		18.194	TCP	mtqp > 4903 [PSH, ACK] Seq
508	2009-03-14 00:04:59.034332		18.194	TCP	[TCP Out-of-Order] mtqp >
509	2009-03-14 00:04:59.049371	8.194	.1	TCP	4903 > mtqp [PSH, ACK] Seq
510	2009-03-14 00:04:59.252254		18.194	TCP	mtqp > 4903 [ACK] Seq=71 A
511	2009-03-14 00:04:59.252258		18.194	TCP	[TCP Dup ACK 510#1] mtqp >
512	2009-03-14 00:04:59.280864	.242	.1	TCP	http-alt > sbl [ACK] Seq=1
513	2009-03-14 00:04:59.280906		6.242	HTTP	Continuation or non-HTTP t
514	2009-03-14 00:04:59.280910		6.242	HTTP	[TCP Out-of-Order] Continu
515	2009-03-14 00:04:59.718366	.242	.1	TCP	http-alt > sbl [ACK] Seq=1
516	2009-03-14 00:05:00.193094	.242	.1	HTTP	Continuation or non-HTTP t
517	2009-03-14 00:05:00.193303		6.242	HTTP	Continuation or non-HTTP t
518	2009-03-14 00:05:00.193309		6.242	HTTP	[TCP Out-of-Order] Continu
519	2009-03-14 00:05:00.593117	.242	.1	TCP	http-alt > sbl [ACK] Seq=5
520	2009-03-14 00:05:00.593178		6.242	HTTP	Continuation or non-HTTP t
521	2009-03-14 00:05:00.593183		6.242	HTTP	[TCP Out-of-Order] Continu
522	2009-03-14 00:05:00.851593	.242	.1	HTTP	Continuation or non-HTTP t
523	2009-03-14 00:05:00.851690		6.242	HTTP	Continuation or non-HTTP t
524	2009-03-14 00:05:00.851694		6.242	HTTP	[TCP Out-of-Order] Continu
525	2009-03-14 00:05:00.860087		.1	DNS	Standard query A 127.0.0.1
526	2009-03-14 00:05:00.860094		.1	DNS	Standard query A 127.0.0.1
527	2009-03-14 00:05:00.862058		.1	DNS	Standard query A p3040-ipb

研究用データセット　～ CCC DATAset 2009 ～

- （3）攻撃元データ
  - ハニーポット94台による1年間のマルウェア取得ログ
  - データ収集日
    - 2008年5月1日～2009年4月30日

ログ項目	例（一部を*でマスク）
マルウェア検体の取得時刻	2009-04-01 00:01:58
送信元IPアドレス	honey035
送信元ポート番号	1034
宛先IPアドレス	**.215.1.206
宛先ポート番号	80
TCPまたはUDP	TCP
マルウェア検体のハッシュ値（SHA1）	*****86f2ec74727b14001cfe0b88af718797c91
ウイルス名称	WORM_AUTORUN.CZU
ファイル名	C:¥WINDOWS¥system32¥ptkj.exe



研究用データセット　～ CCC DATAset 2009 ～

- （3）攻撃元データ

項目	件数
全レコード数	2,470,766
TCPによるダウンロードレコード数	63,820
UDPによるダウンロードレコード数	61,275
ダウンロードホストIPアドレス種類数	269,730
マルウェア検体のハッシュ値種類数	67,055
ウイルス名称種類数（UNKNOWN含まない）	1,335

CCC DATAset 2008／2009の主な違い

項目	2008	2009
マルウェア検体		
検体数	1	10
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
攻撃通信データ		
ハニーポット	honey001, honey002	honey003, honey004
収集日	2008/4/28, 2008/4/29	2009/3/13, 2009/3/14
攻撃元データ		
ハニーポット数	112台	94台
ハニーポットID	なし	あり
収集期間	2007/11/1～2008/4/30	2008/5/1～2009/4/30

研究用データセットの要件と課題

- データの種類
  - 要件
    - ①プログラムされた動作を解析できる検体そのもの
    - ②ネットワークを介した感染・感染後の挙動データ
    - ③PC内部の挙動データ
    - ④必要となる前処理をした扱い易いデータ
    - ⑤データ収集時点でしか得られない補足データ
  - 考慮事項
    - ①はマルウェア検体としてハッシュ値提供
    - ②は攻撃通信データ
    - ④は攻撃元データとして多面的な分析ができるようログ項目を選定
  - 課題
    - ③のファイル・レジストリ操作、⑤のDNSレコード、ブラックリスト

研究用データセットの要件と課題

- データ収集環境の網羅性
  - 要件
    - 攻撃対象そのもの
      - OS種類、パッチ適用状況、AP導入状況、AP操作、各種設定
    - ネットワーク接続環境
      - ISP、IPアドレス帯、大域、アクセス制御
  - 考慮事項
    - 一般ユーザが多く利用するPC・インターネット環境
    - 国内主要ISPを混在
  - 課題
    - 物理的・論理的なリソースコスト
    - 自動解析のための相当数の検体数
    - 近年被害が拡大している受動的攻撃

研究用データセットの要件と課題

- データ収集の期間
  - 要件
    - 長期間にわたる連続性のあるデータ
    - すぐに提供できる最新のデータ
  - 考慮事項
    - 攻撃通信データは休前日・休日を選定
    - 攻撃元データは2008・2009で連続、2009は1年間に拡大
  - 課題
    - データ収集・提供の主体
    - 継続的な管理・提供

研究用データセットの要件と課題

- データ収集環境の運用情報
  - 要件
    - マルウェア活動の変化とデータ収集環境の変化
      - リセット周期、IPアドレス割当、障害対応、性能拡張
  - 考慮事項
    - 必要最低限の情報を意見交換会などで事前に共有
  - 課題
    - 機密性の高い技術ノウハウ
    - 攻撃者によるデータ収集環境の検知
    - 公開範囲・内容

まとめ

- マルウェア対策のための研究用データセット
  - ～CCC DATAset 2008/2009～
- 研究成果の共有
  - ～MWS2008～
- MWS2009
  - 一般口頭発表30件（うち学生の部15件）
  - （2）攻撃通信データの新たな活用　～ MWS Cup 2009 ～
  - MWSの新たな展開に向けたパネルディスカッション
  - 英語表記決定　anti-Malware engineering WorkShop

http://www.iwsec.org/mws/2009/