Vol.2009-CSEC-47 No.2
2009/12/18

MWS Cup 2009 活動報告 ～競技用通信データ作成について～

細井琢朗[†1]　畑田充弘[†2]

去る 2009 年 10 月 26 日に，MWS 2009 の企画として，MWS Cup 2009 が開催された．今回の競技では，感染ホストと通常ホストの通信データを先頭数分間のデータから区別することと，その感染しているマルウェアの名前とその後のネットワーク上での活動を推測することが出題された．我々はこの競技会の企画，運営を担当し，競技に用いる通信データを作成した．本発表では，その競技用通信データの作成について報告する．

MWS Cup 2009 Activity Report —about Construction of Contest Traffic Data—

HOSOI Takurou[†1] and Mitsuhiro Hatada[†2]

On October 26th, 2009, MWS Cup 2009 was held as one of projects in MWS 2009. Questions in this contest were discrimination between infected hosts' and normal hosts' traffic data of the first few minutes, and inference of malware names and subsequent network activities. We were in charge of planning and running this contest, and we made the traffic data for this contest. In this paper, we report on the construction of the contest traffic data.

1. はじめに

2009 年 10 月 26 日（月）～10 月 28 日（水）に，富山市において，共通のデータセット CCC DATAset 2009[1)] を研究用データセットとして活用し，研究成果を発表するワークショップ，マルウェア対策研究人材育成ワークショップ 2009（MWS 2009）が開催された．その中の企画として，共通のデータを使ったマルウェア対策の競技会 MWS Cup 2009 が，ワークショップの第一日に行われた[⋆1]．MWS Cup 2009 では，ワークショップの主眼の一つである，人材育成（特に学生実験）の主旨に沿うこと，また，産業界で使いたくなるようなツールや技術の発掘・評価を目的としており，以下の課題を正答することによる得点（競技時間：1 時間 20 分，50 点満点）と，その後に行われた解析手法の発表に対する採点委員からの採点（各発表 3 分間 + 質疑応答，50 点満点）での高得点を目指して，計 7 チーム[⋆2]が競い合った．

（各課題の選択肢は省略.）

課題 1　攻撃通信データを探し出せ。

（正答：+4 点 ／ 無解答：0 点 ／ 誤答：−2 点 ／ 20 点満点）

競技用 CD-ROM に収められている (W) が 5 個，(B) と (B′) で 5 個の合計 10 個の通信データファイル（pcap ファイル）のうち，(B) と (B′) を探し出し，その番号を解答欄に記入せよ．

番号には，各 pcap ファイルの拡張子を除いた数字を用いよ．

(W)　= マルウェアに感染していない PC の通信データ．攻撃を受けても感染しないパッチ適用済 PC の通信，P2P 通信，正規バイナリファイルの取得通信データ等を含むこともある．

(B)　= CCC DATAset 2009 の攻撃通信データとして未配布の 2009 年 3 月 15 日のうち，ハニーポットからマルウェアが検出されたタイムスロットの初めから数分の通信データ．

(B′)　= (W) と (B) を混合した通信データ．

課題 2　マルウェア名を言え．

（正答：+3 点 ／ 無解答：0 点 ／ 誤答：−1 点 ／ 15 点満点）

課題 1 で探し出した各通信データファイルにおいて，「ホストが感染しているウイルス名称」を以下の選択肢から選び，それぞれの解答欄に記入せよ．

尚，各解答欄には一つの選択肢しか入らない．

†1 東京大学
The University of Tokyo
†2 NTT コミュニケーションズ（株）
NTT Communications Corporation

⋆1 共通のデータを使ったネットワークセキュリティの競技会としては，KDD Cup 1999[2)] が有名である．
⋆2 チームを構成する人員の数に制限は設けなかった．

（注：）選択肢の名称は，CCC DATAset 2009 攻撃元データに従う．

……

課題 3 今後の通信パターンを予測せよ．

（正答：+3 点 ／ 無解答：0 点 ／ 誤答：−1 点 ／ 15 点満点）

課題 1 で探し出した通信データファイルの，「今後の通信パターン」について，以下の選択肢から該当するものを一つ選び，解答欄に記入せよ．

（注：）一つの選択肢が，複数の解答欄に該当することはない．

……

我々は MWS Cup 2009 企画担当の一員として，この課題に用いる，(W)，(B)，(B′) の通信データを作成した．本発表では，各通信データの作成方法の概略とその性質について報告する．また併せて，競技結果を基にした考察も行う．

2. 競技用通信データの作成指針

MWS Cup 2009 で出題した競技用通信データには，その目的のために，幾つかの条件を課した．

まず，競技に用いることから，以下の条件を満たす必要がある．

(1) 各通信データは，少なくとも競技者には開示可能である．

(2) 通信データの大きさは，競技時間内に CD-ROM から余裕を持って全て読み込める程度に小さい．

(3) 各通信データ，特に (B)，(B′) データについては，出題側で正解を正しく用意できる．

また，1 節で述べた MWS Cup 2009 の目的から，次の二つの条件も課せられる．

(4) 初心者（学生など）でもある程度解くことができる．

(5) 習熟者（実用ツールの開発者など）にとっても挑戦し甲斐がある．

これらの条件が満たされるかどうかは課題内容に大きく依存するが，競技用通信データもそれに沿うものでなければならない．

これらの条件から，各通信データは以下の作成指針の下に作成した．特に，各通信データから読み取れる基本情報（IP アドレス，パケットの取得時刻など）では (W) と (B)，(B′) が容易に区別できないように配慮した．

((W) データ)

- 一般の ISP を介してインターネットに接続した，マルウェアに感染していない PC ホスト上で取得した通信データを元に作成する．
  - このホストの OS は (B) データを取得したホストと同じもの[⋆1]にする．但し，通信データ取得時までに提供されたセキュリティパッチは全て適用させた[⋆2]．
  - このホストの IP アドレスは (B) データのものと合わせておく．
  - このホストが問い合わせる直近の DNS サーバの IP アドレスも，(B) データのものと同じになるようにネットワークの設定をしておく．
  - (B) データとの判別をある程度困難にするために，通信を取得している間，このホストの正規のユーザがこのホストを正常に使用する．

    正常な使用： このホストから意図的に攻撃（ネットワーク攻撃など）を行わず，またこのホストを意図的に危険に晒すこと（安全かどうかわからない Web サイトの閲覧など）も行わない，その他のホスト使用．
  - 通信データ取得後に，市販のウィルス対策ソフトウェア[⋆3]を用いて，このホスト内に感染が無いことを確認する．
- 通信データに保存される各パケットの取得開始時刻は，後の操作で合わせる（データ取得時には特に考慮しない）．
- 通信データに保存されている各パケットの MAC アドレスは，後の操作で一律に変更する．

((B) データ)

- CCC DATAset 2009 の攻撃通信データとして配布された二日間のデータに続く，未配布の 2009 年 3 月 15 日の通信データの中の，一台のハニーポットの通信データから作成する．
- 通信データに付属する攻撃元データにおいて何らかの感染が確認された時刻の周辺のパケットを取り出して，出題用の通信データを作成する．
- 出題用の通信データには，以下の性質を満たすものを用いる．
  - 感染開始から，感染後の挙動までの通信が含まれている．
  - 通信データ上では感染が認められても，何らかの理由により攻撃元データに記録されていない通信があり，陽性と判断された本来の活動と動作の切り分けが困難な通信データではない．
  - 選ばれた他の (B) データとは異なるマルウェア名の感染が確認できる．

⋆1 Microsoft Windows XP
⋆2 (B) データを取得したホストはハニーポットであり，一部のセキュリティパッチは適用されていない．
⋆3 TrendMicro ウィルスバスター 2008

（ここで判明したマルウェア名と活動を正解として用いる.）

- 通信データに保存されている各パケットの取得開始時刻は，後の操作で合わせる．
- 通信データは後の操作で Layer-2 のフォーマットを Ethernet のフォーマットに変換する．その際，各パケットの MAC アドレスは一律の値にする．

((B′) データ)

- (W) と (B) それぞれから，うまく混ざり合う適当な通信データを一つずつ選び，一つの通信データに併せる．

これとは別に，(1) の条件を満たすのはデータの作成方法の工夫だけでは不可能であったため，MWS Cup 2009 の競技の開始前に，競技参加者全員の「MWS Cup 2009 への参加に関する同意書」への署名を，出場した各チーム毎に集めた．この同意書は CCC DATAset 2009 の使用の際に結ぶ「研究用データセットの使用に関する契約書」に準じた内容になっており，これを以って，競技用通信データの漏洩などへの対策の一つとした．また，競技用通信データの入った CD-ROM は MWS Cup 2009 企画担当者が厳重に管理し，競技開始直前に必要数を配布し（データのハードディスクなどへのコピーは禁止した），競技終了後に全て回収した．

3. 競技用通信データの作成方法

MWS Cup 2009 で出題した競技用通信データは，2 節の指針に従い，それぞれ次のように作成した．

3.1 (W) データの作成方法

今回は MWS Cup 2009 のために，5 個の (W) データと，2 個の (B′) データ作成用材料データを作成した．

(1) 2 節の通りに整え，インターネットに接続した PC ホスト上で，ツールを用いて通信を取得し，データファイルに保存する．その間，このホストの正規のユーザがこのホストを正常に使用する．
この使用方法を変えた，幾つかの通信データファイルを作る．

(2) 作成したデータファイルに含まれる各パケットの MAC アドレスを，ツールを用いて全て一律に上書きする．

(3) 上記の通信データファイルから，ファイルサイズ，パケット数，経過時間，含まれるパケットを勘案した上で，ツールを用いて適当な部分を切り出す．その際，連続した NetBIOS のパケットなどの余計なパケットがなるべく含まれないようにする．

(4) 含まれている NTP パケットを全て削除する．

(5) ツールを用いて，通信データの開始時刻を全て「2009-03-15 00:00:00」（1 秒未満はそのまま）に合わせる．

(6) 通信データファイルのタイムスタンプを全て「2009-10-19 12:00:00」に変更する．

3.2 (B) データの作成方法

今回は MWS Cup 2009 のために，3 個の (B) データと，2 個の (B′) データ作成用材料データを作成した．

(1) CCC DATAset 2009 の攻撃通信データとして配布された二日間のデータに続く，未配布の 2009 年 3 月 15 日の通信データの中から，ツールを用いて適切な一台のハニーポットのデータを抽出する．

(2) この通信データは，その取得環境のために，一部のパケットについてはそのコピーも含む（非常に近い時刻に同じ内容のパケットが二つある）．このコピーをツールを用いて削除する．

(3) この通信データは，その取得環境のために，その Layer-2 のフォーマットが “pseudo-protocol (Linux-SLL)” フォーマットになっている．これをツールを用いて Ethernet のフォーマットに変換する．その際，MAC アドレスは全て一律のものにする．

(4) 上記の通信データファイルから，ツールを用いて，2 節の指針に沿った，マルウェアの感染とその後の活動を含む適当な部分を幾つか切り出す．その際，連続した NetBIOS のパケットなどの余計なパケットがなるべく含まれないようにする．

(5) 含まれている NTP パケットを全て削除する．

(6) 上記の通信データファイルから，課題 3 で出題される，今後の通信パターンの予測の範囲に当たる部分を削除する．

(7) ツールを用いて，通信データの開始時刻を全て「2009-03-15 00:00:00」（1 秒未満はそのまま）に合わせる．

(8) 通信データファイルのタイムスタンプを全て「2009-10-19 12:00:00」に変更する．

3.3 (B′) データの作成方法

今回は MWS Cup 2009 のために，(W)，(B) データそれぞれ 2 個の材料データから，2 個の (B′) データを作成した．

(1) 3.1 節で作成した，(B′) データ作成用の (W) データと，3.2 節で作成した，(B′) データ作成用の (B) データから，それぞれ一つずつ選び，ツールを用いて一つの通信データに併せる．

(2) 通信データファイルのタイムスタンプを全て「2009-10-19 12:00:00」に変更する.

4. 競技用通信データの性質

MWS Cup 2009 で出題した競技用通信データは，2 節の指針に従い，3 節の方法で作成した，そのため，以下に示す性質を持つ.

- Layer-2 のフォーマットは Ethernet である．また，その MAC アドレスは一律に変更されている.
- チェックサム値は一般には正しくない.
- パケット取得の開始時刻は変更されており，どのデータでも「2009-03-15 00:00:00」である（データファイル内でのパケット取得の相対時刻は変更されていない）．またデータファイルのタイムスタンプは「2009-10-19 12:00:00」である.

  このようにパケット取得の開始時刻を全て同じ時刻にすると，時刻情報を利用する一部の検知技術（参考 : 文献 3)）の利用が妨げられる．しかし，日付のみの変更で留めるのは，(B) データの時刻に合わせて (W) データを取得する困難さを考えると現実的ではない.
- NTP のパケットは含まれない.
- IP アドレスは変更されていない.

  これは通信データを取得したホストにローカル IP アドレスを割り当てている（DNS もローカルのものを使っている）ために可能となった．このことにより，課題を解く際に，IP アドレスの値そのものを利用することができる．また，IP データグラム内の IP アドレスのみを変換した際に起こる，TCP パケット内の IP アドレスとの不整合とそれに起因する通信データとしての劣化を回避している.

競技用通信データの簡単な統計情報を表 1 に掲載する． この表から，(W) データと (B) , (B′) データは，一点（最大パケットサイズ）を除き，簡単な統計量だけでは区別し難いものになっていることがわかる.

5. 競技結果を基にした考察

MWS Cup 2009 の競技結果の集計値は，表 2 の通りとなった． この集計から，今回出題した競技用通信データは，少なくとも競技参加者の全員に扱えるものであったこと，また容易には全問正解ができないように設定された課題の性質を損なうものではなかったことがわかる．即ち，2 節で設けた (4) , (5) の条件（初心者でも扱え，習熟者にも挑戦し甲斐がある）に沿った通信データであったと言える．さらに，得点が（特に課題 2 , 課題 3 で）ある程度ばらついたことから，得点により優劣を競う競技に使えるデータであったことがわかる．ここで，課題 1 ではばらつきが殆ど出なかったことは，このようなデータを作成する際の今後の課題として残る（(W) データの作成にさらに工夫を加える，など）．また，課題 2 , 課題 3 での得点が少ないのは，課題の難しさもあるが，(B) , (B′) データを作成する際に，感染が判明してからの通信を大きく削除したことが最大の原因と考えられる．今回 MWS Cup 2009 に出題した競技用通信データは，人的，時間的制約から，通信データに付属する攻撃元データから判っているマルウェアの取得をするかしないかの内に，通信データを打ち切ったものになっている．この打ち切り開始時点を後ろにずらすことは，課題設定との兼ね合い（課題 3 にある，今後の活動の予測の設問）もあり，課題 1 に対しての通信データの工夫より難しい.

表 1 競技用通信データの簡単な統計情報.

	経過時間	パケット数				パケットサイズ		
	(秒)	ICMP	TCP	UDP	総数	最小	平均	最大
(B) TROJ_AGENT.ARWZ	188 s	0	160	9	169	40	490	1454
(B) TROJ_DLOADR.CBK	407 s	1	235	11	247	40	395	1454
(B) WORM_ALLAPLE.IK	18 s	4	155	0	159	40	444	1454
(B′) BKDR_RBOT.ASA + PING（“www.yahoo.co.jp” へ）	946 s	0	90	13	103	40	153	1454
(B′) PE_VIRUT.AV + “cmd.exe” のダウンロード	362 s	1	324	29	354	40	214	1454
(W) PING（“www.yahoo.co.jp” へ）	307 s	16	56	4	76	40	86	856
(W) MWS 2009 ウェブサイトの閲覧	14 s	4	169	0	173	40	140	1452
(W) Svsql ウェブサイトの閲覧	347 s	0	233	8	241	40	694	1452
(W) “ffftp.exe” のダウンロード	23 s	8	161	0	169	40	142	1452
(W) Cygwin のダウンロードとインストール	29 s	0	182	0	182	40	779	1452

表 2 各課題の最小得点，平均点，最大得点.

	課題 1	課題 2	課題 3	合計
満点	20	15	15	50
最大得点	20	7	7	28
平均点	17.4	2.6	-1.6	18.4
最少得点	14	-3	-5	7
下限点	-10	-5	-5	-20

表 3 競技用通信データ別の正答，無解答，誤答の集計．

	課題 1		課題 2			課題 3		
	偽陽性	偽陰性	正答	無解答	誤答	正答	無解答	誤答
(B) TROJ_AGENT.ARWZ	–	0/7	5/7	1/7	1/7	1/7	1/7	5/7
(B) TROJ_DLOADR.CBK	–	0/7	0/7	1/7	6/7	0/7	1/7	6/7
(B) WORM_ALLAPLE.IK	–	0/7	2/7	3/7	2/7	2/7	3/7	2/7
(B′) BKDR_RBOT.ASA + PING（“www.yahoo.co.jp” へ）	–	0/7	3/7	1/7	3/7	1/7	1/7	5/7
(B′) PE_VIRUT.AV + “cmd.exe” のダウンロード	–	3/7	1/7	2/7	1/7	0/7	2/7	2/7
(W) PING（“www.yahoo.co.jp” へ）	0/7	–						
(W) MWS 2009 ウェブサイトの閲覧	2/7	–						
(W) Svsql ウェブサイトの閲覧	0/7	–						
(W) “ffftp.exe” のダウンロード	1/7	–						
(W) Cygwin のダウンロードとインストール	0/7	–						

競技結果の集計を細かくし，正答，無解答，誤答の結果を競技用データ別に集計したものが，表 3 である． この集計から，課題を解くに当たって難易度が高い通信データと低いデータがあることがわかる．例えば課題 2 を見ると，(B) データの一番目では多くのチームが正答しているのに対し，(B) データの二番目ではほぼ全てのチームが誤答している．但し，これは競技用通信データの性質ではなく，対応するマルウェアの性質（亜種が多く判別が難しい，など）が結果に反映しているのが要因かもしれない．一方，(W) データを見ても，誤答（偽陽性）が二つのデータに分かれて現れている．これは明らかに競技用通信データの性質が要因であり，使用したデータがある程度競技に向いていたことを示している．

6. ま と め

以上，MWS Cup 2009 に出題した競技用通信データの作成方法とその性質について報告し，競技結果の集計からデータに対して幾つかの考察を加えた．それらをまとめると，少人数で短期間に作成した競技用通信データではあったが，MWS Cup 2009 の競技課題に堪えるものであったと言える．一方，改善すべき点も幾つか見付かった．それらの内の幾つかは改善方法とその効果が明らかであり，次の機会があれば是非改善したいと考えている．残念ながら，一部の点については現時点では現実的な方法が見付からず，今後の課題として残っている．

謝辞 MWS Cup 2009 の企画担当として共に尽力して下さった，竹森敬祐氏，菊池浩明先生に深く感謝致します．また，MWS Cup 2009 に出場していただいた競技者の皆様，競技の場を提供して下さった MWS 2009 ， CSS 2009 の両実行委員会にもこの場を借りて御礼申し上げます．

参 考 文 献

1) 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，コンピュータセキュリティシンポジウム 2009（CSS 2009），富山県富山市，情報処理学会，pp.1–6 (2009).
2) : KDD Cup 1999 (Computer network intrusion detection). http://www.kdd.org/kddcup/index.php?section=1999&method=info.
3) Samano, V. J.Z., Hosoi, T. and Matsuura, K.: Time Categorization in a Social-Network-Analysis Spam Filter, *Workshop on Information Security Applications 2008 (WISA 2008)*, Jeju Island, Korea (2008). (CD-ROM).