OLAP(多次元データ分析)を利用した攻撃元データの分析と検体の自動解析

永田大　堀合 啓一　田中 英彦
情報セキュリティ大学院大学

あらまし　OLAP とは，BI などで利用されるデータを多次元に解析する手法である．本論では，研究用データセット CCC DATAset 2008 の攻撃元データを OLAP による分析を行い，分析の方法とその結果を考察する．また，筆者らはマルウェアの自動解析システムを構築し運用しているが，このシステムの概要と研究用データセット CCC DATAset 2008 のマルウェア検体の自動解析を行った結果についても報告する．

Analysis of CCC DATAset 2008 using OLAP and Automated malware variant classification

Dai Nagata, Keiichi Horiai, Hidehiko Tanaka
INSTITUTE of INFORMATION SECURITY

Abstract The OLAP is a method to analyze multi-dimensional data in the area such as Business Intelligence. In this report, we describe the method of applying OLAP to the analysis of CCC DATAset 2008 and its results. Authors have been developing an automated malware analyzing system. We have adopted the system to analyze the specified malware in the CCC DATAset 2008 and we describe the outline of the system and the obtained results.

1　はじめに

OLAP（On-line Analytical Processing）とは，ビジネスインテリジェンス等で利用されるデータを多次元に解析する手法である．本論では，研究用データセット CCC DATAset 2008 の攻撃元データ(以降，攻撃元データ)を分析するに当たり OLAP を使用した．多次元でデータを扱うことにより，マルウェアの傾向分析を行った．また，筆者らは定点観測と連動したマルウェアの自動解析システムを構築し運用しており，このシステムを拡張し，研究用データセット CCC DATAset 2008 のマルウェア検体(以降，マルウェア検体)の挙動解析を行った．

2　OLAP の概要

OLAP は，ビジネスインテリジェンス（BI）と呼ばれる経営分析等で用いられる．企業の業務システム等で蓄積されたデータの分析・加工を行い，企業経営の意思決定に活用しようとする手法である．このように OLAP は，一般的には経営計画や企業戦略に利用される．BI ツールでは，論理的に 3 次元以上の視点を持つ分析用データベースから 2 次元のクロス集計表としてデータを取り出しデータの多次元的な分析を行うことができる．

本論では，攻撃元データについて多次元データとして解析することを試み，その手法として OLAP を利用することとした．OLAP にてデータを扱うことにより，基礎数値の把握と傾向分析を簡潔に行うことができるのではないかと考えた．

3　分析システムの概要

本研究では，OLAP のソフトウェアとして，

OpenStandia[1]を利用した．OpenStandia は，OpenOLAP をベースにしたフリーウェアである．OpenStandia では，データベースに mySQL が使われている．図 1 に構成を示す．

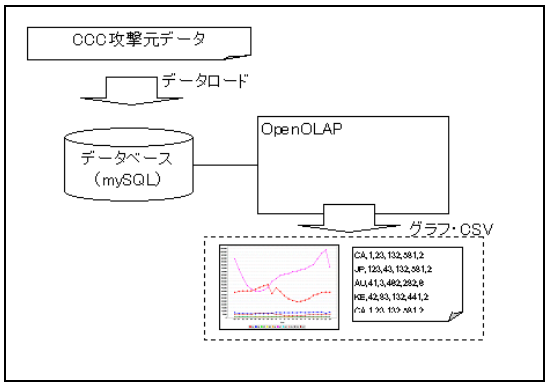


図 1 システムの構成

OLAP では，リレーショナルデータベースのデータを特定のタイミングで多次元データとして再構成する．再構成されたデータをキューブと呼ぶ．キューブに対して問い合わせを行うことによりグラフや CSV の元となる分析情報を取得することができる．またキューブは，複数のディメンジョンで構成される．ディメンジョンとは多次元的にデータを集計する際の軸である．

3.1 テーブル構造

OLAP でのデータ解析にあたり，mySQL にログのデータをロードするが，そのテーブル構成を表 1 に示す．ログの値と同じものに加え，マルウェアの Family 名や CountryCode を追加した．マルウェア名称についてカンマより前の部分を Family 名として扱う．例えば PE_BOBAX.AH であれば，PE_BOBAX を Family 名とする．追加した項目は，他のカラムからの導出が可能であるが，性能を考慮し，あらかじめデータを設定しておく方針とした．

3.2 ディメンジョンの構成

今回の分析では，表 2 に示すディメンジョンを作成した．これらの各ディメンジョンを組み合わせて多次元的にデータを分析した．

表 1 ログ管理テーブル

カラム名	説明
pk	プライマリキー（追加項目）
datetime	日時
ip	攻撃元 IP アドレス
port	ポート番号
direction	攻撃方向
hash	ハッシュ値
virusname	マルウェア名
filename	ファイル名
virusnamegroup	マルウェアの Family 名（追加項目）
countrycode	CN（追加項目）
24h	24 時間表記での受信時刻（追加項目）
24h_timezone	24 時間表記での受信時刻で日本との時差を考慮したもの（追加項目）

表 2 使用したディメンジョン

名称	説明
virusgroupname	マルウェアの Family 名を扱う
countrycode	攻撃元 IP アドレスから導出した国コードを扱う
24 時間	24 時間表記での時刻時間を扱う
ハッシュ値	検体のハッシュ値を扱う

4 分析結果

本システムを使用し次の分析を行った．

・全体基本情報の把握
・国別，時間帯別の分析
・ハッシュ値，マルウェア名称の分析

4.1 全体基本情報の把握

マクロ的な概要をつかむため，全体の基本と

なる統計情報を取得した．本システムでは，データをテーブルにロードしているため，全体の種類数や種類別出現数の情報は，SQL にて容易に取得可能である．表 3 に SQL で取得した件数情報を記述する．また表 4 にマルウェア名称についての種類別出現件数の情報を記述する．

表 3 全体基本情報 種類数

全ログレコード数	2,942,221 件
マルウェア名称による種類数	1,082 種類
ハッシュ値による種類数	52,465 種類
攻撃元 IP 種類数	258,711 種類
port/プロトコルによる種類数	63,820 種類
攻撃国別種類数	150 種類

表 4 マルウェア名称別出力件数（上位 5）

UNKNOWN	651,492
PE_BOBAX.AH	130,708
BKDR_VANBOT.AX	97,176
BKDR_VANBOT.AD	68,785
TROJ_POEBOT.AGU	66,649

4.2 国別，時間帯別グラフ

鬼頭[2]らは，不正ホストからの攻撃を時間軸での分析を行い，その中でそれぞれの国における日本時間を現地時間に修正し，時間帯別による攻撃数の変化のグラフ化を行っている．本システムで同様に，国別・時間帯別の集計を行った．図 2 は，出現数の多い国から 10 カ国（JP,CA,CN,US,TW,CZ,KR,DE,IN,PH）に対して，時差を考慮せずにグラフ化したものである．また，図 3 は，同じ国について時差を考慮してグラフ化したものである．

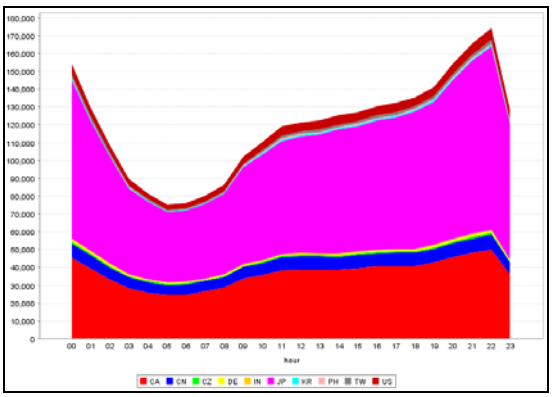


図 2 時間帯別攻撃数

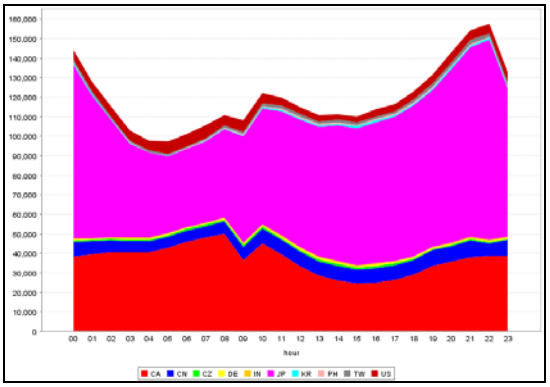


図 3 時間帯別攻撃数（時差補正）

今回の分析では，時差補正を行った場合に顕著な特徴見られなかった．

4.3 ハッシュ値，マルウェア名称の分析

ハッシュ値とマルウェア名称の関係を分析した．図 4 は，ウィルスパターンが更新されるまでの期間を視覚化したものである．横軸は月を示し，縦軸はマルウェアの検体を示す．UNKNOWN の期間が赤色で，名前が付くと黄色，さらに名前が変わると別の色で表現している．図 5 は，UNKOWN の検体へマルウェアの名称が付与されるまでの期間を示したものである．

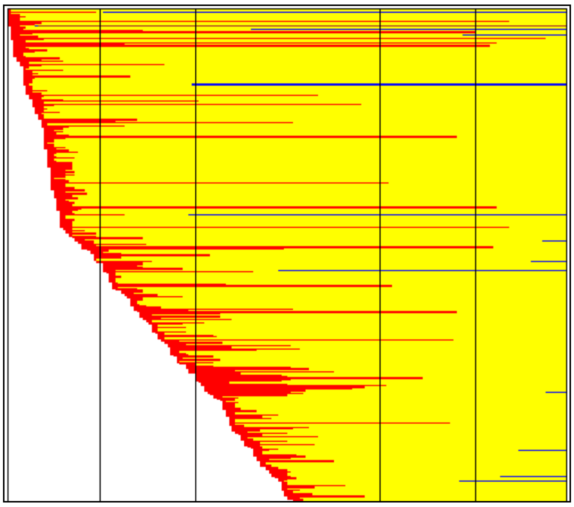
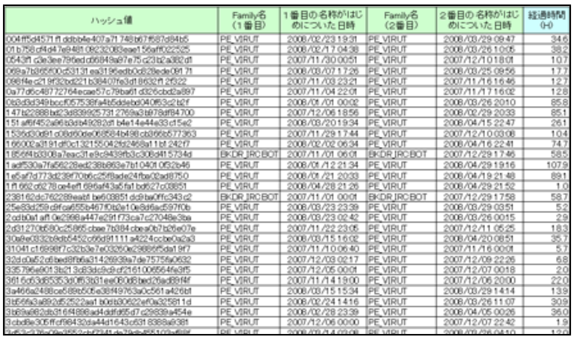


図 4 ウィルスパターン更新の可視化



ハッシュ値	Family名（1番目）	1番目の名称がはじめについた日時	Family名（2番目）	2番目の名称がはじめについた日時	経過時間(H)
004ff5d4571f1ddbb4e407a717f48b67f687d84b5	PE_VIRUT	2008/02/23 19:31	PE_VIRUT	2008/03/29 09:47	34.6
01b758cf4d47e94f81 08232083eae156aff022525	PE_VIRUT	2008/02/17 04:38	PE_VIRUT	2008/03/26 10:05	38.2
0543f1 c3e3ee79fedc66848a97e75c23b2a382d1	PE_VIRUT	2007/11/30 00:51	PE_VIRUT	2007/12/10 18:01	10.7
069a7b3655f00c53131ea3196edb0c828ede0f171	PE_VIRUT	2008/03/07 17:26	PE_VIRUT	2008/03/25 09:56	17.7
098f4ec219f32bd221b38407fe3d18632f12f322	PE_VIRUT	2007/11/03 23:21	PE_VIRUT	2007/11/16 16:46	12.7
0a77d6c48772764ecae57c79ba61d326cbd2a897	PE_VIRUT	2007/11/04 22:01	PE_VIRUT	2007/11/17 16:02	12.8
0b3d3d349bccf057538fa4b5ddebd040f63c2b2f	PE_VIRUT	2008/01/01 00:02	PE_VIRUT	2008/03/26 20:10	85.8
147b22888bd23d839925731276 9a3b978df84700	PE_VIRUT	2007/12/06 18:56	PE_VIRUT	2008/02/29 20:33	85.1
151af6f452a96b3db49282d1b4e14e44e33c15e2	PE_VIRUT	2008/03/20 19:34	PE_VIRUT	2008/04/15 22:47	26.1
1536d30d91c08d60de06584b498cb366b577363	PE_VIRUT	2007/11/29 17:44	PE_VIRUT	2007/12/10 03:08	10.4
166002a3f191df0c132155042fd2468a11bf242f7	PE_VIRUT	2008/02/02 06:34	PE_VIRUT	2008/04/16 22:41	74.7
1856f4b3308a7eac31e9c9439fb3c308d415734d	BKDR_IRCBOT	2007/11/01 06:01	BKDR_IRCBOT	2007/12/29 17:46	58.5
1adf530a7fa56228ed238b863e7b10401 0f32b46	PE_VIRUT	2008/01/12 21:34	PE_VIRUT	2008/04/29 19:16	107.9
1e5af7d773d239f70b6c25f8ade24fba02ad8750	PE_VIRUT	2008/01/21 20:33	PE_VIRUT	2008/04/19 21:48	89.1
1f1662c6278ce4ef1696af43a5fa1bd627c03851	PE_VIRUT	2008/04/28 21:26	PE_VIRUT	2008/04/29 21:52	1.0
238162dc76228 9eabf be603851dc9ba0ffc343c2	BKDR_IRCBOT	2007/11/01 00:01	BKDR_IRCBOT	2007/12/29 17:58	58.7
25e83d259c9fca655b467f0b2e10e8d6ad597f0b	PE_VIRUT	2008/03/23 23:39	PE_VIRUT	2008/03/29 03:51	5.2
2cdb0a1af10e2998a447e291f73ca7c27048e3ba	PE_VIRUT	2008/03/23 02:42	PE_VIRUT	2008/03/26 00:15	2.9
2d31270b680c25865cbae7b384cbea0b7b26e07e	PE_VIRUT	2007/11/22 23:05	PE_VIRUT	2007/12/11 05:25	18.3
30a9e0832b9db5452c66d91111a4224ccbe0a2a3	PE_VIRUT	2008/03/15 16:02	PE_VIRUT	2008/04/20 08:51	35.7
31041c16898f7c32b3e7e03260e2988 5f5da19f7	PE_VIRUT	2007/11/10 06:40	PE_VIRUT	2007/11/16 00:01	5.7
32dc0a52c6bed8fb6a314269 39a7de75756a0632	PE_VIRUT	2007/12/03 02:17	PE_VIRUT	2007/12/09 22:26	6.8
335796e9013b213d83dc9c9cf2161006564fe3f5	PE_VIRUT	2007/12/05 00:01	PE_VIRUT	2007/12/07 00:18	2.0
3616c63d65353d0f63b31ee080d8bed26ad89f4f	PE_VIRUT	2007/11/14 19:00	PE_VIRUT	2007/12/06 20:00	22.0
3a466a2488ce589b505e38f49763a0c561a426bf	PE_VIRUT	2008/03/15 15:34	PE_VIRUT	2008/03/29 14:14	13.9
3b56fa3a892d52522aa1b0db30622ef0a32581 1d	PE_VIRUT	2008/02/24 14:16	PE_VIRUT	2008/03/26 11:07	30.9
3b69a982db316f4898ad4ddfd65d7c29839a454e	PE_VIRUT	2008/02/28 23:39	PE_VIRUT	2008/04/05 00:26	36.0
3cbd8e305ffcf98432da44d1643c6318388a938f	PE_VIRUT	2007/12/06 00:00	PE_VIRUT	2007/12/07 22:42	1.9
[illegible]	PE_VIRUT	[illegible]	PE_VIRUT	[illegible]	[illegible]

図 5 マルウェア名称付与までの期間

これらの図によって，マルウェア名称の変化（定義体が更新される期間）を見ることができる．マルウェア名称が途中で変化しているもので，UNKNOWN を除外したハッシュ値は，247 件であった． 247 件の中で Family 名の変化に着目して確認したところ，Family 名が変わった検体が 109 件，Family 内だけで変わった検体が 138 件（内 126 件は PE_VIRUT Family）という結果であった．マルウェアの名称が途中で変わった 247 件のうち，ほぼ半数にあたる 126 件が PE_VIRUT Family の検体であった．

PE_VIRUT は，2008 年 1 月の CCC 分析レポート[3]によると，2007 年 11 月末より増えていて， 1 日 300 種の異なる検体が発生すると報告されている．マルウェア名称が変更されるということは，はじめの定義ファイル提供時の分析の時点から分析の結果が変わっているので，解析が難しいマルウェアや，新しいタイプのマルウェアと考えられる．

4.4 Excel によるグラフと簡易アニメーション化

OpenOLAP では，Viewer から分析結果を CSV 形式のファイルに出力することできる．その出力データを利用して，Excel による簡易アニメーションを行った．時系列に複数のシートにデータを配置した．グラフの元データを連続的に切り替えることにより，簡易アニメーションとして，傾向の変化を見る事ができる．図 6 は，国とマルウェア名称を軸に取り，アニメーション切り替えを時間帯によって実施した場合の表示例である．

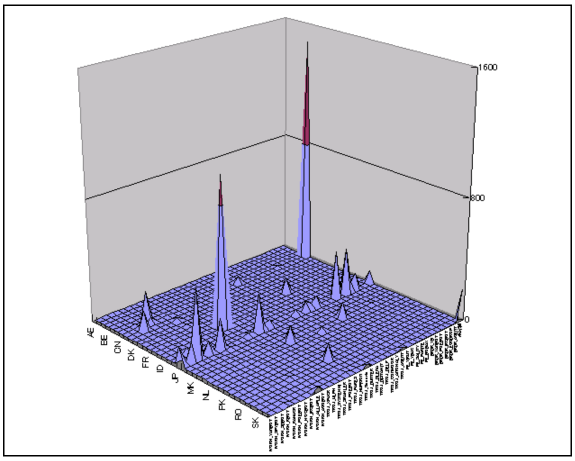


図 6 アニメーション表示例

5 マルウェア検体の自動解析

筆者らは定点観測と連動したマルウェアの自動解析システムを構築し運用している[4,5]．このシステムは，ハニーポットを定点観測のセンサーとして利用し，捕獲したマルウェアを一般のネットワークから隔離した安全な環境で実行して，その挙動を自動的に解析するものである．解析の対象は，マルウェアの実行前後の Windows のファイルやプロセス等の変化と，観測される通信の状況である．また，このシステ

ムは解析済みデータをＤＢへ蓄積し，新たに捕獲したマルウェアの挙動とＤＢ内のデータとの類似性を自動的に算出して，その名称を推定する機能を有している．マルウェアの実行環境は，感染後に必要となる Windows システムの復旧の容易性から仮想マシン上で構築している．

このシステムを利用して，指定されたマルウェア検体の解析を試みたが，arp パケットが観測されるのみで，その他の挙動は観測されなかった．近年のマルウェアの中には，自身が解析されることを妨害する機能を持ったものの存在が知られていることから，今回の検体がこれに該当すると想定し，仮想マシンを利用しない環境を構築して解析を試みた．

5.1 解析環境の概要

自動解析は解析全般の処理を制御する PC（制御 PC）とマルウェアを実行する PC（感染 PC）の２台で構成した．制御 PC は Linux を OS とし，解析全般の制御を行う他，次の処理を行う．

・解析析環境に必要なネットワークの模擬
・DNS,IRC など関連サーバの模擬
・マルウェア実行中の通信パケットの記録
・感染 PC から取得したマルウェア実行前後のログを比較して得られる挙動の変化の抽出と HTTP サーバと連動した解析結果の出力

一方の感染 PC は，マルウェアを実行する Windows OS 起動用のパーティションと，このパーティションのファイルを感染前の状態へ復旧させるための環境を備えた Linux 起動用のパーティションのデュアルブート構成とし，以下のステップで処理を行う．

(1) Linux のパーティションから起動し，感染前の正常な状態の Windows のディスク・イメージ・ファイルを，Windows 用のパーティションへコピー．

(2) ブートマネージャ Grub[6]の機能を利用し，当該 PC をリブートした際に，Windows のパーティションから１回だけ起動するように設定して感染 PC をリブート．

(3) 起動した Windows は，マルウェア実行前の状態を記録し，その後に制御 PC から解析対象のマルウェアファイルを受信して，このファイルを実行する．

(4) マルウェア実行後の状態を記録し，実行前のログとともに制御 PC へ転送する．

(5) 感染 PC 自身をリブートし,(1)のステップから繰り返す．

基本的には制御 PC と連携し，以上のステップを繰り返すことによって，全自動でマルウェアの解析が可能であるが，マルウェアの実行に伴う Windows システムのクラッシュ等の原因で(4),(5)が機能しない場合が発生する．この状態を検出し，制御 PC からの指令によって感染 PC をハード的にリブート（電源の OFF/ON）させる仕組みを持たせている．

5.2 解析結果

解析結果は，Web ブラウザで閲覧可能なシステムとして実装しているが，マルウェア検体の解析結果の表示例を図 7 に示す．図において②レジストリの改ざんが１件，③システム関連ファイルの生成と削除がそれぞれ１件，④ NVCOM プロセスの起動などが記録されている．また，⑤には起動したプロセスと，このプロセスが待ち受けているポート番号，⑥には HOSTS ファイルの改ざんの状態が表示される．解析対象の検体は，HOSTS ファイルの改ざんを行うことが知られているが[7]，本自動解析ではファイルの最後に１バイト(¥x0a)が追加されていることを確認している．これらの解析結果は，ウィルス対策製品ベンダーなどが Web 上で公開している情報[7]とほぼ一致している．⑦にマルウェアの実行に伴って発生した通信パケットの情報を示す．マルウェアは解析環へ実装した模擬 DNS サーバへホスト名 rx7.teensmutbox. com の FQDN の問い合わせを行い，模擬 DNS が指定した仮の IP アドレス 4.3.2.234 の IRC サーバへログインする様子が示されている．
(172.16.50.128 は感染 PC の IP アドレスである．)

このように，既知のプロトコルを利用するマルウェアについては，動的な挙動を観測することで，ボットネットの指令サーバを自動的に特定し，指令文字列を自動的に抽出することが可能である．

本自動解析システムはこれまでに述べた挙動をカテゴリデータへ変換してベクトル化し，ベクトル間のハミング距離を指標とした挙動の類似性を基に，マルウェアの名称を推定する機能を有している[5]．筆者らが仮想マシン上の環境で自動解析を行って蓄積した約 7,600 検体のデータとの挙動の類似性から推定したマルウェアの名称が①に示されている．この例では，推定した名称が BKDR_VANBOT.NM となり，実際の名称と一致しなかった．その原因としては，実行環境の違い（仮想マシン上と実マシン上）に加えて，蓄積している解析済の DB 内に，今回の検体と挙動が類似しているデータが含まれていないことが原因と推定している．

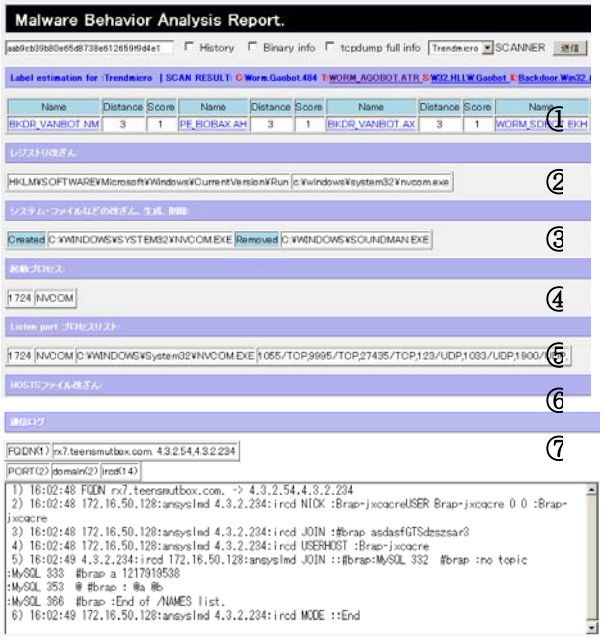


図 7 検体の自動解析結果

6 まとめ

OLAP を利用することにより，準備されたディメンジョンに対しては，容易にグラフ化・CSV 化をすることができるようになった．しかしながら，事前にディメンジョンやキューブを作成する必要があるためリアルタイムでの分析は難しい．

マルウェアの自動解析については，挙動の解析については所期の目的を達することができた．ただし，挙動の類似性を基にしてマルウェアの名称を推定する機能については，DB の充実やアルゴリズムの改善による精度の向上が今後の課題である．

参考文献

[1] OpenStandia http://www.nri-aitd.com/openstandia/)

[2] 鬼頭哲郎，仲小路博史，寺田真敏，菊池浩明「インターネット上の不正ホスト分布に関する社会的レイヤからの考察」情報処理学会研究報告．CSEC Vol.2007, No.71(20070719) , Jul.2008

[3] 2008 年 1 月 CCC 分析レポート https://www.ccc.go.jp/report/200801/0801monthly.html

[4]堀合啓一，今泉隆文，田中英彦「定点観測によるボットネットの観測と Malware の動的挙動解析システムの提案」情報処理学会論文誌 Vol.49 No.4 1-12, Apr.2008

[5]堀合啓一，今泉隆文，田中英彦「ハミング距離によるマルウェア亜種の自動分類」CSEC Vol.2008, No.45(20080523), May.2008

[6]GNU GRUB http://www.gnu.org/software/grub/

[7] WORM_AGOBOT.ATR http://www.trendmicro.co.jp/vinfo/virusencyclo/default5.asp?VName=WORM_AGOBOT.ATR&VSect=T