ボットネットはいくつあるか？ ダウンロードログからの線形独立な基底数

松尾 峻治 †　　菊池 浩明 †　　寺田 真敏 ††

† 東海大学院工学研究科情報理工学専攻
259-1292 平塚市北金目 1117ozuma,kikn@cs.dm.u-tokai.ac.jp
†† 日立製作所 Hitachi Incident Response Team(HIRT)
212-8567 神奈川県川崎市幸区鹿島田 890　日立システムプラザ新川崎

あらまし 本研究では，CCC DATAset 2009 の攻撃元データに主成分分析を適用することで，効率よく独立なボットネットを抽出する手法を提案する．主要な結論は，独立した 4 つのボットネットが判別できたことと，1 年間が 5 つのフェーズに分割されたことである．

How Many Botnets are Running? Number of Linear Independent Basses in Log of DownLoading

Shunji Matsuo †　　Hiroaki Kikuchi †　　Masato Terada††

†Graduate School of Engineering, Tokai University, 1117 Kitakaname , Hiratsuka , Kanagawa 259-1292
††Hitachi,Ltd.Hitachi Incident Response Team(HIRT), 890 Kashimada , Kawasaki , Kanagawa 212-8567

Abstract In this study, we propose a new method to extract independent botnets from the CCC DATAset 2009, the log of downloading servers observed by distributed honeypots, by applying a technique of principal component analysis. The main results include that four independent botnets are distinguished and a year is divided into five phases.

1　はじめに

ボットは，コンピュータウイルスの一種で，感染したコンピュータをネットワークを通じて外部から操ることを目的として作成されたプログラムである．ボットネットはボットから構成されるネットワークであり，ハーダと呼ばれる指令者から攻撃コマンドを受け取ることにより，スパムメールの送信，DDoS 攻撃，ホスト内にある個人情報取得など，多様な不正行為の基盤として利用されている．プログラムの構成変更などを通して，アンチウィルスソフトのシグネチャによる検知や早期発見を困難にしている．

[1] によると，大規模ボットネット “Donbot ’’は一日に 100 億のスパムを送信しているという．[2] は，3,600 台を超える C&C サーバ数の推移と位置を定期的に報告している．これだけ多くのボットネットが活動しているその一方で，C&C サーバを特定されないように長期間にわたり同一のボットネットを使い続けることはなく [3]，それゆえその規模や実態を探るのは困難であった．

そこで，本研究では，ボットのダウンロードサーバ（以下 DL サーバ）とダウンロード時刻に相関があると考え，CCC DATAset 2009[8] の攻撃元データにて観測されたソースアドレスとマルウェアのダウンロード時刻の相関に着目し，それらの間に成立する特定のパターンを機械的に抽出することを試みる．DL サーバはどこかひとつ以上のボットネットに属しており，ボットネット単位で活動の増減があることを仮定すると，ハニーポットで観測しているのはそれらの独立したボットネットの線形和である．線形和の組は長期間に渡るダウンロードの履歴により与えることが出来る．従って，その中から線形独立なベクトルを抽出すれば，それが独立したボットネットの数と規模を与えるはずである．

本論文では，固有値解析の手法を導入することで，効率よく独立なボットネットを抽出する手法を提案する．本提案手法で用いることで，

- 独立したボットネットの個数
- ボットネット間の勢力の時系列変化

を明らかにする．CCC DATAset のデータを用いた実験結果を報告し，提案方式の有効性を示す．

2 提案方式

ボットネットは，(1) 脆弱性を利用して最初の攻撃を行い投入されるダウンローダー，(2) ポートスキャンなどのコマンドコントロールするための各種のツールを提供するダウンロードサーバ，(3) 全制御を行う C&C サーバから成っている．ダウンロードはさらに，長期間オンラインで多くのホストと通信する PE 型，短期間で独自のツールを提供する W(Worm) 型，T(Troy) 型などに細分化され，[4] によると，

$$PE \to W \to T$$

といった一連の動作で攻撃を繰り返している例が報告されている．そこで，一般的に各ボットネットに，この PE，W，T の組が複数あり，連携して攻撃しているとすると仮定する．

今 A と B の二つの独立したボットネットが，各々，(PE, W, T) に対応するサーバ (S_1, S_2, S_3) と (S_1, S_2, S_3) を持っていると考える．この時，各サーバ S_i が単位時間当たりの DL 回数を

	S_1	S_2	S_3	S_4	S_5	S_6
A	3	1	1	0	0	0
B	0	0	0	4	2	1

と仮定する．4 月には A のみが活動し，5 月には互いに争い，6 月には B が優勢になったならば，その間に試みられる DL 回数は

	S_1	S_2	S_3	S_4	S_5	S_6	
4 月	30	10	10	0	0	0	$10A$
5 月	9	3	3	4	12	6	$3A+3B$
6 月	3	1	1	40	20	10	$A+10B$

であったとしよう．

では，ここで，このダウンロードの観測データが与えられたとき，ボットネットの数がいくつあるといえるだろうか?次のように推定を試みよう．まず観測データを，簡略化のため，サーバ集合と観測点の集合についての DL 回数を持つ配列

$$X = \begin{pmatrix} 30 & 10 & 10 & 0 & 0 & 0 \\ 9 & 3 & 3 & 12 & 6 & 3 \\ 3 & 1 & 1 & 40 & 20 & 10 \end{pmatrix}$$

とおく．この場合，独立したボットネットは，X における線形独立な行ベクトル (A と B の二つ) である．仮定より，X のどの行も A と B の線形和で表されているからである．3ヶ月の例で説明したが，十分な観測期間を与えれば，独立したベクトルが識別できると考える．

そこで，主成分分析としてよく知られている次の手法を適用する．列の平均値で正規化した $\grave{C}$ の共分散行列 $V = \grave{C}\grave{C}^\top$ の固有ベクトルは，互いに直交した列ベクトルを効率よく与える．例えば，X について固有値を求めると，固有値の大きい順から

$$\begin{aligned} \boldsymbol{u}_1 &= (-0.5, -0.2, -0.2, 0.7, 0.4, 0.2) \\ \boldsymbol{u}_2 &= (0.8, 0.3, 0.3, 0.5, 0.2, 0.1) \end{aligned}$$

の 2 つの固有ベクトルを得る．これが，元のボットネット A と B の個数に対応している[1]．これを直交基底と呼び，内積 $u_1 \cdot u_2 = 0$ と直交し，ノルム $||\boldsymbol{u}_i|| = 1$ の単位ベクトルである性質を持っている．

この直交基底 $\boldsymbol{u}_1$，$\boldsymbol{u}_2$ を用いれば，X の任意の行はこれらの基底（ボット）の線形和（合成）で現されることを示す．例えば X の 1 行目 $\boldsymbol{x}_1$ は

$$\boldsymbol{x}_1 = y_1\boldsymbol{u}_1 + y_2\boldsymbol{u}_2 + \boldsymbol{x}_0$$

で表すことができる．ただし，$\boldsymbol{x}_0$ は行についての平均ベクトル $\boldsymbol{x}_0 = (14, 4.6, 4.6, 17.3, 8.6, 4.3)$ である．

直交性のため，係数 y_1 は

$$\begin{aligned} y_1 &= \boldsymbol{x}_1 \cdot \boldsymbol{u}_1 = -26 \\ y_2 &= \boldsymbol{x}_2 \cdot \boldsymbol{u}_2 = 4 \end{aligned}$$

で求められる．この y_1, y_2 は主軸 $\boldsymbol{u}_1$，$\boldsymbol{u}_2$ に対する係数である．4 月のダウンロード観測数に該当する $\hat{\boldsymbol{x}_1} = \boldsymbol{x}_1 - \boldsymbol{x}_0$ の一行目 $\grave{C} = (16, 4.4, 4.4, -17.3, -8.6, -4.3)$ は，二つのボットネットとその勢力を表す二つの係数で得られる式

$$-26\boldsymbol{u}_1 + 4\boldsymbol{u}_2 = (16.2, 6.4, 6.4, -16.2, -9.6, -4.8)$$

[1] 実際には，$\boldsymbol{u}_1$,$\boldsymbol{u}_2$ が直接 A，B になっているわけではなく，$\boldsymbol{u}_1 = B - A$，$\boldsymbol{u}_2 = A + B$ に該当している

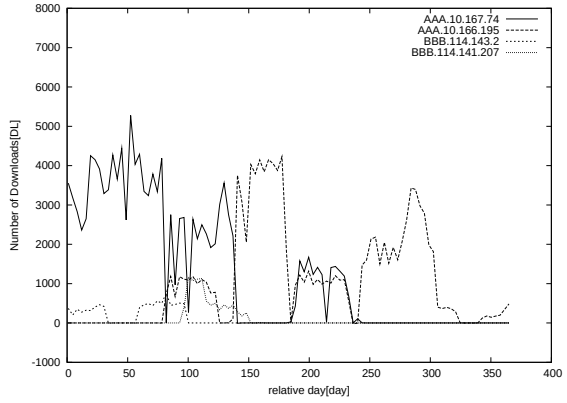

図 1: 上位 4 ソースアドレスでの日毎の DL 数推移

で得られる．これより，十分な精度でボットネットの独立線形の行ベクトルの近似出来ており，独立した 2 つのボットネットで十分であることがわかる．

3 実験

3.1 目的

2 節で述べた手法を CCC DATAset 2009 攻撃元データ [8] に適用して，次の項目を明らかにすることを目的とする．

- 独立したボットネットの数,
- ボットネットの活動期間と年間を通じた変化.

ことを目的とする．ことを示す.

3.2 実験データ

本実験のオリジナルデータとして，2008/5/1 から 2009/4/30 までの攻撃元を観測するハニーポットで収集された “CCC DATAset 2009”[8] の「攻撃元データ」を用いた．また本実験では，この攻撃元データ内で取得されたソースアドレスの DL 数で上位 100 個に対象を絞り，検証をおこなった．

対象となる実験データの基本統計量（表 1），活動状況（表 2，図 1）を示す．表 1 に記載した活動日数は，2008/05/1 から 2009/4/30 までの各ソースアドレスが観測された総日数である．表 2 は，表 1 のソースアドレスについての毎日のマルウェアダウンロード数の一部であり，図 1 は表 2 の上位 4 つのソースアドレス通年のダウンロード数の推移である．

表 2: 日ごとのＤＬ数

	Top1	Top2	⋯	Top99	Top100
2008/5/1	3555	0		0	0
2008/5/2	3510	0		0	0
⋮					
2009/4/29	0	788		0	0
2008/4/30	0	486		0	0

3.3 実験方法

1. 表データを $m = 365$ 行，$n = 100$ 列の配列 X で表わす．$\boldsymbol{x}_0$ は $x_{0j} = 1/m \sum_i^m x_{ij}$ で定義される平均ベクトルとする．共分散行列 V は，C を $C = X - \boldsymbol{x}_0$ と X を正規化し，$V = C \cdot C^\top$ で定める．この時，固有値の大きな順に固有値 $\lambda_1, \ldots, \lambda_n$，固有ベクトル $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n$ を基底と呼ぶ.

2. $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n$ を用いて，X を直交展開する．すなわち $i = 1, \ldots, 100$ について

$$y_i = \boldsymbol{x}_i \cdot \boldsymbol{u}_i$$

で与えられる主成分 (周波数成分)$y_1, y_2, \ldots, y_n$ を求める.

3. (y_1, y_2) 空間上で，$m = 365$ 個の観測データを可視化する．明らかな特徴的な変化を見つけ，いくつかのフェーズに分ける.

3.4 実験結果

表 2 を X として求めた主軸 (基底) を表 3 に示す．これらの基底について X を展開した時の主成分 y_1，y_2 の推移を図 3 に示す．X 軸を y_1，Y 軸 y_2 という散布図で主成分の分布を図 4 に示す．同様に y_3，y_4 の推移を図 5 に，y_2，y_4 の推移を図 6 に，X 軸を y_2，Y 軸 y_4 の散布図を図 7 にそれぞれ示す. 固有値計算には Octave[5] を使用した.

3.5 考察

3.5.1 独立したボットネット数について

図 4 より，二つの phase 1，phase 3，5 の点がきれいな直線を描いていることより，少なくとも強い影響力を持つ二つのボットネット勢力があることがわかる．これより，本実験データの攻撃元データか

表 1: DL 数上位 20 位のソースアドレス

順位 Pod	IP アドレス	総数 [DL]	活動日数 [day]	平均 [DL/day]	ユニーク Hash 数	ユニーク Honey 数
1	AAA.10.167.74	462246	184	2512.21	119	91
2	AAA.10.166.195	399562	249	1604.67	48	92
3	BBB.114.143.2	33283	73	455.93	29	82
4	BBB.114.141.207	32202	53	607.58	37	78
5	CCC.215.1.206	26780	62	431.94	7	59
6	DDD.95.79.6	19641	117	167.87	99	85
7	AAA.10.169.26	14951	223	67.04	52	82
8	EEE.48.75.63	11699	148	79.05	100	69
9	CCC.18.161.250	10060	121	83.14	131	68
10	AAA.8.143.164	5099	70	72.84	40	81
11	FFF.202.252.41	4901	43	113.98	13	87
12	GGG.131.76.60	4659	71	65.62	52	74
13	CCC.215.1.226	4492	76	59.11	36	68
14	HHH.247.2.38	4262	61	69.87	166	77
15	III.18.116.75	4112	128	32.13	6	75
16	JJJ.219.170.67	3963	106	37.39	56	72
17	KKK.16.245.53	3766	31	121.48	9	63
18	LLL.90.134.24	3723	66	56.41	15	83
19	MMM.180.151.74	3473	29	119.76	6	87
20	HHH.247.2.32	3368	29	116.14	11	82

表 4: 主要な感染フェーズ

フェーズ	期間	関係式
1	2008/5/1〜2008/7/19	$y_1 = -34.6y_2 - 2063$
2	2008/7/20〜2008/9/17	$y_1 = -60.8y_2 + 140.4$
3	2008/9/18〜2008/10/27	$y_1 = 55.7y_2 + 1933$
4	2008/10/28〜2008/12/26	$y_1 = -31.2y_2 + 525.5$
5	2008/12/27〜2009/4/30	$y_1 = 4.8y_2 + 1446$

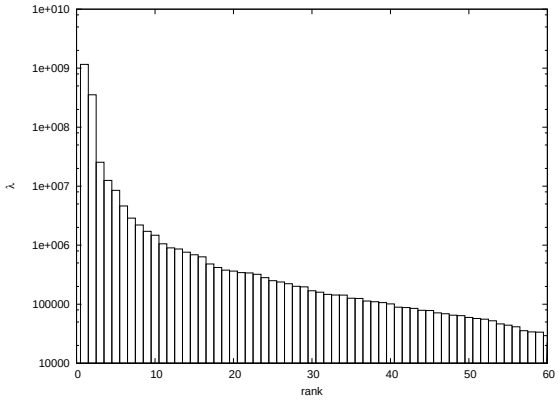

図 2: 上位 20 位の固有値 λ

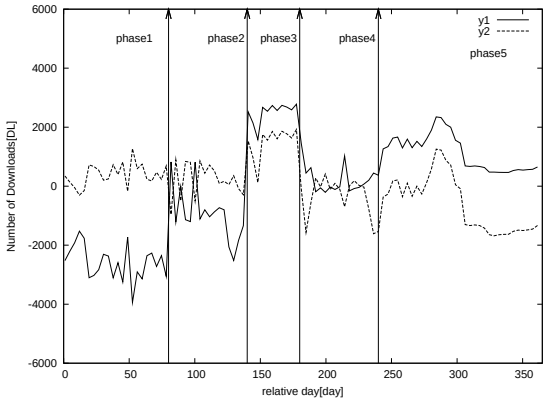

図 3: 第一，第二主成分の推移

ら独立な基底が 2 つはあり，少なくとも大規模な 2 つのボットネットがあると考える．

更に，7 より 4 つの明確な直線が観測できる．本実験ではこれを 4 つの独立したボットネットと考える．しかし，この個数に関する精度は十分とは考えていない．それは，主に上位 2 位までのソースアドレスの DL 数が異常なほど多かったことと，データ収集を行っているハニーポットのアドレス分布の偏りにも影響されているためである．

3.5.2 ボットネットの活動期間

図 4 で観測される y_2，y_1 の動きから，観測期間を主観的に図 3 の 5 つのフェーズに分けた．各フェーズを表 4 に整理する．フェーズ 1 では，y_2 は y_1 に対して負の線形関係にあるのに対して，phase3 と 5 では正の関係にある．

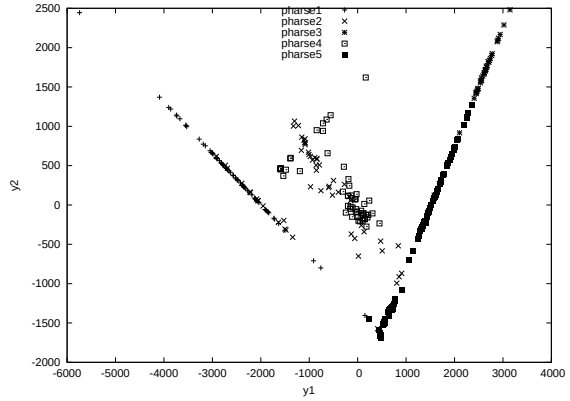

図 4: 第一，第二主成分の分布とフェーズ

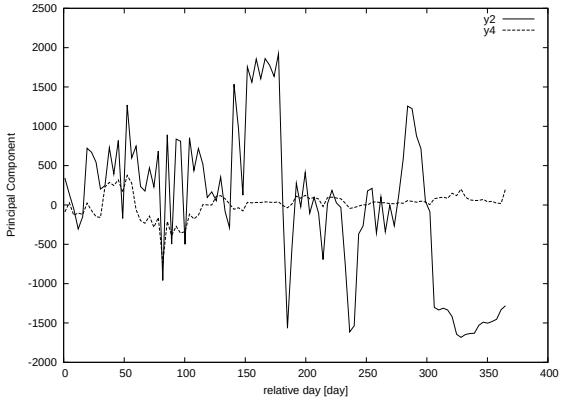

図 6: 第 2，第 4 係数推移

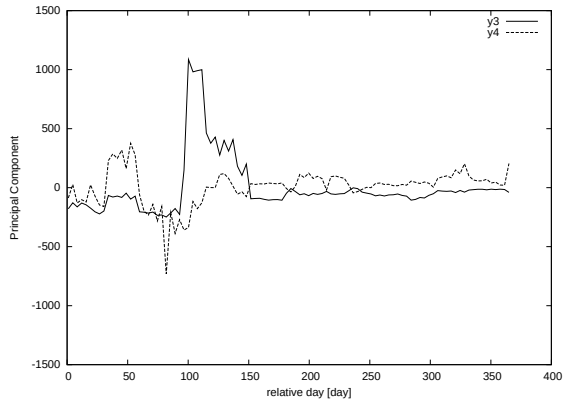

図 5: 第 3，第 4 係数推移

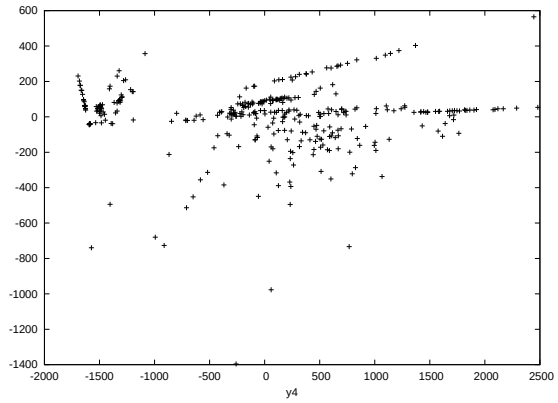

図 7: 第 2，第 4 係数散布

3.5.3 ボットネットの勢力変化

前節にて y_2 は y_1 に対して正と負の線形関係，二つの関係あることを述べた．この理由には，

1. 二つの巨大なボットネットが互いのホストの勢力を争っている．
2. 単一のボットネットが主要な DL サーバを切り替えて運用している．

の二つの説が考えられる．一つめの説は，y_1 と y_2 が負の線形の時のとき二つのボットネットが感染ホストを奪い合い，比例した動きをした時，一つのボットネットによって支配されていると解釈する仮説である．

二つ目の仮説は，単一のボットネットが主要な DL サーバを切り替えながら運用しているとみなすものである．y_1 と y_2 の間に相補的な関係が見えたのは，DL サーバの運用を二つの群単位で行っており，相補的な時はどちらか一つの群で，比例している時は二つの群で協調して運用をしているのではないかと考えられる．表 4 のフェーズ 2 や 4 のように，ある期間では相補的する形で増減を繰り返し，同フェーズ 3 や 5 のような期間では同じように増減している．

4 結論

本論文では，ダウンロードログに主成分分析を適用して独立したボットネットを判別する解析方式を提案した．実験結果から独立した 4 つのボットネットが識別できた．観測期間の一年間は，独立した 5 つのフェーズに分かれていた．

謝辞

本研究を遂行するにあたって，攻撃元データの解析方法について助言してくださった (株) 日立製作所の小堀智弘氏，鬼頭哲郎氏，中小路博史氏，藤原将志氏に深謝します．

表 3: 直交基底 (一部)

IP アドレス	主軸			
	$\boldsymbol{u}_1$	$\boldsymbol{u}_2$	$\boldsymbol{u}_3$	$\boldsymbol{u}_4$
A.10.167.74	**-0.83**	0.54	-0.02	0.08
A.10.166.195	0.55	**0.83**	-0.02	0.02
B.114.143.2	-0.05	0.02	-0.31	**-0.90**
B.114.141.207	-0.02	0.03	**0.94**	-0.27
C.215.1.206	0.01	-0.11	-0.03	0.28
D.95.79.6	0.02	-0.02	-0.04	0.11
A.10.169.26	-0.01	0.02	-0.04	-0.01
E.48.75.63	0.01	0.01	-0.03	0.04
C.18.161.250	0.01	-0.01	-0.02	0.05
A.8.143.164	-0.01	0.00	0.01	-0.09
F.202.252.41	0.01	0.00	-0.01	0.00
G.131.76.60	-0.01	0.01	-0.03	0.00
C.215.1.226	0.00	-0.01	-0.01	0.03
H.247.2.38	0.00	-0.01	-0.01	0.03
I.18.116.75	-0.01	0.01	0.00	-0.03
J.219.170.67	-0.01	0.00	0.04	-0.04
K. 16.245.53	0.00	0.01	0.03	0.01
L.90.134.24	0.00	0.00	0.00	0.01
M.180.151.74	0.00	0.00	-0.01	0.01

参考文献

[1] 野津誠, “ボットネットに関与した ISP の接続停止で世界のスパムが 38 %減少”, *Internet Watch*, 2009 月 3 日, http://internet.watch.impress.co.jp/docs/news/20090903_312659.html (2009 年 9 月参照).

[2] Shadowserver, “Botnet Charts”, http://www.shadowserver.org/wiki/pmwiki.php/Stats/BotnetCharts (2009 年 9 月参照).

[3] 高橋 睦美, “ボット対策、最後に頼りになるのは”, ITMedia エンタープライズ, 2006, http://www.itmedia.co.jp/enterprise/articles/0612/07/news028.html (2009 年 9 月参照).

[4] 大類他, “分散ハニーポット観測からのダウンロードサーバ間のアソシエーションルール抽出”, 情報処理学会, コンピュータセキュリティシンポジウム, CSS2009, 2009.

[5] Octavehttp://www.gnu.org/software/octave/

[6] 戸田, 他, “ISDAS: Internet Scan Data Acquisition System”, 情報処理学会, コンピュータセキュリティシンポジウム CSS 2004, pp. 199-204, 2004.

[7] 福野, 菊池, 寺田, 土居,“不正アクセスのトラフィックによるセンサの独立性”, CSEC36, pp.95-102,2007.

[8] 畑田 充弘, 中津 留勇, 寺田 真敏, 篠田 陽, “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” , MWS 2009, 2009