攻撃データ(CCCDataset2010)を利用したボットネットの C&C サーバ特定手法の再評価

中村 暢宏† 名雲 孝昭† 田中 達哉† 三原 元‡ 佐々木 良一†

†東京電機大学
〒101-8457 東京都千代田区神田錦町 2-2
{nakamura_n,nagumo,tanaka,sasaki}@isl.im.dendai.ac.jp
‡株式会社 NTTPC コミュニケーションズ
〒105-0003
東京都港区西新橋 2-14-1 興和西新橋ビル B 棟

あらまし 近年ボットネットによる被害が増加している. ボットPCの特定・隔離だけでは他のPCがボットとなり, ボットネットは攻撃者を特定しない限り解決に至らない. そこで著者らは, ボットネットを根源まで追跡する多段追跡システムの構想を示した。既存の多段追跡システム第2段追跡方式として, ①C&Cサーバに関するブラックリストと②CCCDataSet2009の解析結果を数量化理論2類に適用する検知方式を併用するC&Cサーバの特定手法を先に提案し有効性を評価した. 本論文では, CCCDataSet2010の解析結果を先の特定手法に適用し, 異なるデータにおいても同手法が有効かどうかを検証し, その評価結果の報告行う.

Revaluation of Technique to DetectC&C Server of Botnet Using CCC DATAset 2010

Nobuhiro Nakamura† Takaaki Nagumo† Tathuya Tanaka†

Hajime Mihara‡ Ryoichi Sasaki†
†Tokyo Denki University
2-2, Kanda-Nishiki-cho, Chiyoda-ku, Tokyo, 101-8457 JAPAN
{nakamura_n,nagumo,tanaka,sasaki}@isl.im.dendai.ac.jp
‡NTTPC Communications, Inc 105-0003
Nishishinjuku,Minato-ku,Tokyo2-14-1Kouwa-nishishinjuku- building-B

Abstract Recently, the damage caused by the bot net has been increasing. There exists a problem that the other bot PCs can be produced, even if one bot PC could be specified and removed. Therefore, we proposed the Multi Stage Trace Back system. We also developed second stage trace back method which consists of black list and Quantification methods No. 2 with CCCDataSet2009. As a result, we were able to know that the developed second stage trace back method could be useful. This paper reports the evaluated results of applying Quantification methods No. 2 with CCCDataSet2010 instead of CCCDataSet2009.

1　はじめに

近年ボットネットの被害が増加し問題になっている。ボットネットとは，ボットウイルスに感染したコンピュータ(以下，ボット PC とする)が複数組み合わさって構成されるネットワークであり，構成するボット PC の数は数百から数万台にのぼる.

ボット PC は C&C(Command and Control)サーバと呼ばれる中継サーバを介して，ボットネットを操作する攻撃者(以下，ハーダーとする)からの命令を受けることで，複数台のボット PC が一斉に SPAM メールの送信や DDos(Distributed Denial of Service)攻撃の実行など，様々な活動を行う.

ボット PC からの攻撃が，送信元を偽装していた場合，特定が困難である. この場合、IP トレースバックなどの方法を用いることで，発見が可能である. しかし対策が不十分であれば、ボットが容易に感染する恐れがある為、根本的な解決には至らない。

このような問題に対して著者らは，ネットワーク管理者が情報共有を行い，ボット PC や C&C サーバ，ハーダーの操作 PC の特定を目的とする，多段トレースバックシステムを構想している.

本論文では，多段追跡システムのうち，先に提案された第二段トレースバックシステムにおける C&C サーバとダウンローダ(以下，第二段追跡対象とする)の検知方式について，CCCDATAset2010 の解析結果を適応させ，CCCDATAset2009 の解析結果を適応させた場合との比較検証を行うと共に，検知手法の有効性について評価する.

2　第二段トレースバックシステム

2.1　用いる検知方式

第二段追跡対象の特定に用いる手法は，ブラックリストを用いる検知方式と，数量化理論 2 類を用いる検知方式の 2 つを組み合わせたものである.

ここでは数量化理論 2 類を用いる為に CCCDATAset2010 の解析結果を使用する.

2.2　ブラックリスト用いる検知方式

ボットネットに関する IP アドレスやドメイン名を公開している複数のサイトが存在する. これらのサイトから，ドメイン名一覧の取得し，ブラックリストを作成する. このブラックリストとマッチングを行うことで，第二段追跡対象の検知を行う.

2.3　数量化理論を用いる検知方式

数量化理論は，林知己夫教授らにより開発された日本独自のデータ分析手法である. ダミー変数の導入による質的データの数量化を行い，多変量解析を行うものである. 使用する数量化 2 類は，分析対象データが数量化不可能な質的データにおいて判別分析と同等の分析を行うものである.

数量化理論 2 類は 2 つのタイプ①と②があるとき，説明変数の係数を適切に設定することで，①は①毎に近い値を，②は②毎に近い値を取るようにしつつ，①と②は離れた値を取る用にするものである. これによって，明快な境界線の設定を可能にする. この係数は相関比を最大にすることにより得られる. 未知のデータを得た場合，この係数の値を用いて①，②どちらに属するか推定することができる.

数量理論 2 類を用いて第二段追跡対象の検知を行うにあたり，「与えられたデータからカテゴリースコアと境界値を設定する」段階と，「求められた 2 種類の値から判別と予測を行う」段階に分かれる.

3　CCCDATAset 解析結果

既存方式に習い，数量化理論 2 類を用いた検知方式に用いる，パラメータ候補を求めるため，CCCDATAset2010 に含まれる，攻撃通信データの解析を行う。この解析には，ボットネットに関連のあるデータと関連の無いデータの 2 種類のデータを用いる. 調査対象は，各ボット PC

が接続する第二段追跡対象のドメインであり，直接ボットネットについての特徴を調査したわけではない.

CCCDATAset2010 の攻撃通信データより下記の手順で第二段追跡対象と思われるドメイン(以下，ボットネットドメインとする)を 30 個取得した.

1. 通信データ内の DNS クエリを取得
2. 取得した DNS クエリ内の以下を除外
   (ア) IP アドレスのみ
   (イ) Yahoo など明らかな正規ドメイン
3. 重複した項目を除外

さらに，30 個のボットネットドメインとは別に，ボットネットとは関係しない通信データとして、当研究室のネットワークから通信データを取得し，任意の DNS 通信から 51 個のドメイン(以下，ノーマルドメインとする)を取得した.

取得した各ドメインから数量化理論 2 類を用いた検知方式に使用するパラメータとして，以下の項目を調査した.

1. 逆引き
2. TTL 値
3. SOA レコード
4. Web，Mail 各サービスの有無
5. WHOIS

これらの調査はそれぞれ、1は，DNS サーバに対して IP アドレスからドメイン名の問い合わせを行う調査. 2 は，DNS サーバから取得したドメイン情報の有効期限である TTL 値の調査. 3 は，各 DNS サーバから取得したドメインの設定情報の調査. 4 は，同一度名上の，Mail サーバ，Web サーバのドメイン名登録の有無を調査. 5 は，各レジストリ組織が管理している，ドメインや IP アドレス，管理者情報の調査である.

各項目の調査手法に関しては，1 は WHOIS サービスを用いて調査を、2.3.4.5 は各ドメイン情報を持つ DNS サーバに対して，dig コマンドを用いて調査を行った.

このうち、「逆引き」と「Web，Mail 各サービスの有無」，「WHOIS」については昨年度と比べ変化が見られず，「逆引き」と「Web，Mail 各サービスの有無」はボットネットドメインを特徴付けるデータとして有効性があると考えられる.「SOA レコード」は，値が集中する箇所の変化が見られるが，特徴付けるデータとしての有効性は失われていない. しかし，「TTL 値」に関しては，一部の値に集中する傾向が少なく，特徴付けるデータとして有効性が下がった.

4 実験による検証と評価

前 3 節の結果を基に，数量化理論 2 類による検知の実験を行った. 実験には，株式会社エスミ社のソフトウェア Excel 数量化理論 Ver2.0 を使用した. この際，判別式に用いるパラメータとその組み合わせを設定する必用があるが， 今回 3 パターンの検証を行う.

4.1 2009 年度の設定による検証

2009 年度データによる実験に使用したパラメータ設定値，及びパラメータの組み合わせを用いて 2010 年度データの検証を行う
各パラメータの設定値は、以下の表 1 に示す通りである.

表 1 パラメータ設定値

逆引き	値
返答なし	1
返答が不正	2
返答が正しい	3

Mailサーバ	値
有り	1
無し	2

TTL値	値
1-1000	1
1001-100000	2
100001-	3
N/A	4

Webサーバ	値
有り	1
無し	2

Minimum値	値
1-100	1
101-1000	2
1001-	3
N/A	4

ドメイン登録期間	値
1-2500(日)	1
2501-5000(日)	2
5001-	3
N/A	4

実験に使用する組み合わせは「Minimum 値」「ドメイン登録期間」「逆引き」の 3 つである. 表 2 に実験結果を示す. 昨年比で検出率が下がったが，検出精度は依然高い結果となった. また，以下の検知率はボットネットドメインにおける False Negative と，ノーマルドメインにお

ける False Negative を合わせた値を 1 から引いた値である.

表 2　検証実験結果

	検証実験	
	検知率	誤検知率
2009年度データ	96.55%	3.45%
2010年度データ	90.0%	10.00%

4.2　2010 年度データによるパラメータ設定実験を用いる検証

赤池情報量基準(以下, AIC とする)は, 統計モデルの良さを評価する為の指針である.AIC を用いる事で, モデルの複雑さとデータとの適合度のバランスを取る事が可能となる.AIC は式(1)で求められる. ここでは, 数量化理論 2 類試行時のパラメータ数の設定に使用する.L は最大尤度、k は自由パラメータである.

$$AIC=-2\ln L+2k \qquad 式(1)$$

今回の使用する値は, k が各要素数にあたり, L が各パラメータ数での数量化理論 2 類を用いて求めた判別結果と, 正答との乖離の最小 2 乗和に相当する.

パラメータ設定実験と検証実験, それぞれに使用するドメインのデータを用意する. ボットネットドメインは 30 個, ノーマルドメインは 51 個有り, それぞれ均等な数での実験が望ましい. よって各ドメインを任意に選択し, ボットネットドメインは 15 個ずつ, ノーマルドメインはパラメータ設定実験に 25 個, 検証実験に 26 個を用いる.

4.2.1　パラメータ数設定実験結果

AIC を用いた最適なパラメータ数の比較結果を表 3 に示す. この結果から, 最適なパラメータ数は AIC 値が最も低いパラメータ数 3 であると分かる.

表 3　AIC 実験結果

パラメータ数	AIC値
2	−833.6359774
3	−955.731048
4	−653.8421243
5	−202.6540491
6	−37.46076972

4.2.2　パラメータ設定実験結果

パラメータ数を 3 で試行した数量化理論 2 類の結果を表 4 に示す. この結果より, 最も検知率が高い値は 70.73%であり, この値の組み合わせは, 4 組存在した.

表 4　パラメータ設定実験結果

						検知率
逆引き	TLL	minimum				65.85%
逆引き	TLL		Mailサーバ			65.85%
逆引き	TLL			Webサーバ		65.85%
逆引き	TLL				登録期間	65.85%
逆引き		minimum	Mailサーバ			70.73%
逆引き		minimum		Webサーバ		68.29%
逆引き		minimum			登録期間	68.29%
逆引き			Mailサーバ	Webサーバ		68.29%
逆引き			Mailサーバ		登録期間	70.73%
逆引き				Webサーバ	登録期間	63.41%
	TLL	minimum	Mailサーバ			65.85%
	TLL	minimum		Webサーバ		70.73%
	TLL	minimum			登録期間	65.85%
	TLL		Mailサーバ	Webサーバ		53.66%
	TLL		Mailサーバ		登録期間	58.54%
	TLL			Webサーバ	登録期間	68.29%
		minimum	Mailサーバ	Webサーバ		60.98%
		minimum	Mailサーバ		登録期間	70.73%
		minimum		Webサーバ	登録期間	65.85%
			Mailサーバ	Webサーバ	登録期間	68.29%

4.2.3　検証実験結果

検証実験の結果を表 5 に示す.結果、最も検知率が高い組み合わせは,「Minimum 値」「Mail サーバの有無」「登録期間」であり, 検知率は 85.0%である. この値は, ボットネットドメインにおける False Negative 13.3%と, ノーマルドメインにおける False Negative 16.0%を合わせた値を1から引いた値であり, 検出精度は項 4.1 と比べ低い値となった.

表 5　検証実験結果

						パラメータ設定実験		検証実験	
						検知率	誤検知率	検知率	誤検知率
逆引き		minimum	Mailサーバ			70.73%	29.27%	80.00%	20.00%
逆引き			Mailサーバ		登録期間	70.73%	29.27%	80.00%	20.00%
	TLL	minimum		Webサーバ		70.73%	29.27%	65.00%	35.00%
		minimum	Mailサーバ		登録期間	70.73%	29.27%	85.00%	15.00%

4.3　パラメータ設定実験に2009年度データと2010年度データを合わせての検証

項 4.2.2 で行ったパラメータ設定実験では，2010 年度のデータから取得したボットネットドメインの半数を使用して実験を行った．ここでは，2009 年度データから取得したボットネットドメインの内，半数に当たる 9 個のドメインを，パラメータ設定実験に追加することで，継続的な調査の有効性について検証を行う．

4.3.1　パラメータ数設定実験結果

AIC を用いた最適なパラメータ数の比較結果を表6に示す．この結果から，最適なパラメータ数は項 4.2.1 同様，パラメータ数 3 である．

表 6　AIC 実験結果

パラメータ数	AIC値
2	−1298.5357
3	−1702.336867
4	−1278.003754
5	−499.5205291
6	−45.1028877

4.3.2　パラメータ設定実験結果

パラメータ数を 3 で試行した数量化理論 2 類の結果を表7示す.この結果より，最も検知率が高い値は 78.00%であり，この値の組み合わせは 1 組であった．

表 7　パラメータ設定実験

						検知率
逆引き	TLL	minimum				72.00%
逆引き	TLL		Mailサーバ			32.00%
逆引き	TLL			Webサーバ		72.00%
逆引き	TLL				登録期間	68.00%
逆引き		minimum	Mailサーバ			76.00%
逆引き		minimum		Webサーバ		74.00%
逆引き		minimum			登録期間	72.00%
逆引き			Mailサーバ	Webサーバ		70.00%
逆引き			Mailサーバ		登録期間	70.00%
逆引き				Webサーバ	登録期間	78.00%
	TLL	minimum	Mailサーバ			72.00%
	TLL	minimum		Webサーバ		74.00%
	TLL	minimum			登録期間	68.00%
	TLL		Mailサーバ	Webサーバ		32.00%
	TLL		Mailサーバ		登録期間	58.00%
	TLL			Webサーバ	登録期間	26.00%
		minimum	Mailサーバ	Webサーバ		66.00%
		minimum	Mailサーバ		登録期間	72.00%
		minimum		Webサーバ	登録期間	70.00%
			Mailサーバ	Webサーバ	登録期間	74.00%

4.3.3　検証実験結果

検証実験の結果を表 8 に示す.結果、パラメータ設定実験でもっとも検知率が高かい組み合わせ「逆引き」「Web サーバの有無」「登録期間」の検知率は 80.0%である．この値は，ボットネットドメインにおける False Negative 6.7%と，ノーマルドメインにおける False Negative 28.0%を合わせた値を 1 から引いた値であり，検出精度は前項までと比較して低下した結果となった．

表 8　検証実験結果

				検知率	誤検知率	検知率	誤検知率
逆引き			Webサーバ 登録期間	78.00%	22.00%	80.00%	20.00%

5　数量化理論 2 類とブラックリスト方式を組み合わせた方式

前項までの実験結果を表 8 に示し比較する．

表 8　ボットネットドメイン検知結果

	項4.1	項4.2	項4.3	ブラックリスト
1	○	○	○	×
2	○	○	×	○
3	×	×	○	○
4	○	○	○	○
5	○	○	○	×
6	○	○	○	○
7	○	○	○	○
8	○	○	○	○
9	○	○	○	×
10	○	○	○	×
11	○	○	○	○
12	○	○	○	○
13	○	×	○	×
14	○	○	○	○
15	○	○	○	○

結果，パラメータ設定実験において2009年度データを使用した項 4.1 では 1 カ所，2010 年度データを使用した項 4.2 では 2 カ所，データを組み合わせた項 4.3 では 1 カ所検出漏れが発生した．しかし，ブラックリストを併用した場合，項 4.1 と項 4.3 では検知漏れした箇所を補う形でブラックリストに検知された．

この結果と各検知率の比較から，継続的な調査により取得したデータを併用して用いること

は，ボットネットドメインの検知について有効であると言える．

6　終わりに

多段追跡システムのうち，先に提案された第2 段トレースバックシステムの解析手法について，新たなデータによる検知方式の検証を行った．実験により，検出率を確認したが昨年度と比べ全体的に若干，低下しているが，ブラックリスト方式と組み合わせたボットネットドメインの検知については引き続き有効であることが明らかになった．今後は，検出率の低下原因と考えられるボットネットの特徴の変動についての調査と共に，第2段トレースバックシステムの実装を目指す．

7　参考文献

[1] 三原元，佐々木良一，「数量化理論とCCCDATAset2009 を利用したボットネットのC&C サーバ特定手法の提案と評価」，マルウェア対策研究人材育成ワークショップ 2009 (MWS2009), A6-1, 2009 年 10 月．
[2] ボットネット概要，
http://www.jpcert.or.jp/research/2006/Botnet_summary_0720.pdf
[3]「ボットウイルスの脅威と対策」2010 年 7 月 総務省・経済産業省連携プロジェクト サイバークリーンセンター
[4]赤池弘次，甘利俊一，北川源四郎，樺島祥介，下平英俊「赤池情報量基準 AIC」
[5] 原口正行のホームページ EXCEL を使った多変量解析
http://gucchi24.hp.infoseek.co.jp
[6] 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 ～MWS 2010Datasets～，MWS2010（2010 年 10 月）