累積データを用いたボットネットの C&C サーバ特定手法の評価

中村 暢宏 †　　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,sasaki}@isl.im.dendai.ac.jp

あらまし 近年，ボットネットの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段追跡方式について，数量化理論2類に適用するデータとして，CCCDATAset2010と前年度のCCCDATAset2009を併用することで，依然として有効である事を再検証した．本論文では，CCCDATAset2011の解析結果を先の特定手法に適応し，有効であるか検証すると共に，3年分の累積したデータの有用性を検証し，その評価結果の報告を行う.

Evaluation of Technique to Detect C&C Server of Botnet Using Accumulated Data

Nobuhiro Nakamura†　Ryoichi Sasaki†

†Tokyo Denki University
†2-2, Kanda-Nishiki-cho, Chiyoda-ku,
Tokyo, 101-8457 JAPAN
{nakamura_n, sasaki}@isl.im.dendai.ac.jp

Abstract Recently, the damage caused by the botnet 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 and CCCDataSet2010.This paper reports the evaluated results of applying Quantification methods No. 2 with CCCDataSet2011 and accumulated of the three years minute.

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登る[1].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.これらのボット PC からの攻撃が送信元を偽装していた場合,特定が困難であるが,それでも 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,ボットが新たに感染する恐れがある為,根本的な解決には至らない[2].

このような問題に対して本研究室では,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た[3].

本研究では,多段追跡システムのうち, 第 2 段トレースバックシステムにおいて C&C サーバ・ダウンローダの 2 種(以下,第二段追跡対象とする)を検知する方式に関するものであり, 先に再検証した検知方式[4]が,新しいデータが得られた時点でも有効か再評価結果を報告すると同時に,3 年間に渡って収集したデータを併用することで,累積したデータの検知方式式への影響を評価するものである.

本稿では,2 章で第 2 段トレースバックシステムについて説明し,3 章で CCC2011 の解析結果について述べ, 4 章で解析結果と過去 2 年分の解析結果とを用いた検知方式の実験検証の結果を報告し,5 章で新たにパラメータ候補の検証結果を報告する.

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二つを組み合わせたものである.ここでは,数量化理論２類を用いて分類を行う為にサイバークリーンセンターより得られた,CCCDATAset2011[5]の解析結果を使用する.

2.2　ブラックリストを用いる検知方式

ボットネットにおいて,不正を働く可能性が高い C&C サーバの 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,ドメインの一覧を取得し,ブラックリストを作成する.このブラックリストとの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. データ分析手法において,分析対象のデータが数値化不可能な量的データである場合,多変量解析が使用できる.しかし,解析対象データが数値化不可能な質的データである場合,多変量解析が使用できない.これに対して数量化理論は,分析対象が質的データの場合でも, ダミー変数の導入による質的データの数量化を行い,多変量解析をする事で解析を行う事ができる.その中でも,今回使用する数量化 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,量的データの判別分析に相当する処理が可能である[4].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行う実験は,「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し,最適な組み合わせを仮定する」パラメータ設定実験と,「求められた組み合わせを,異なるデータを用いて検証する」検証実験の 2 段階に分かれる.

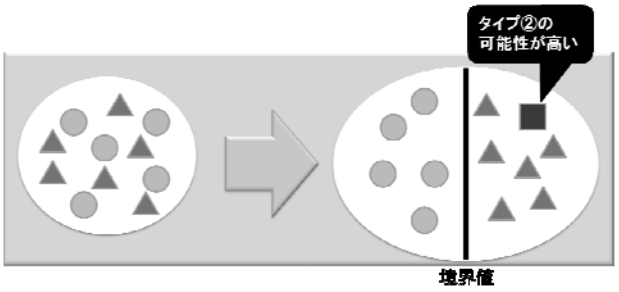

図1. 数量化理論 2 類概要

3　CCCDATAset2011 解析結果

前2章で述べた様に,数量化理論２類を用いた検知方式へ適用するパラメータ候補を求める為,CCCDATAset2011 に含まれる,攻撃通信データの解析を行った[5].この解析では,ボットネットに関係するデータ,比較対象としてのボットネットに関係のないデータ,これら2種類のデータを用いる.なお,これらの調査は,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関する特徴を調査したわけではない.

今回の解析で CCCDATAset2011 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 42 個取得した.

1. 受信データ内の 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

これら 42 個のボットネットドメインの他に,ボットネットとは関係しない通信データとして,当大学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42個のドメイン(以下,ノーマルドメインとする)を取得した.作業手順は,ボットネットドメインの取得時に準ずる.

取得したドメインに含まれる各種データの内,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サーバの有無
5. WHOIS情報

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.

これらはボットネットの特徴が見られるとして調査された項目である.昨年度(2010 年度)データと比較を行った.

その結果,「逆引き」「Whois 情報」「Web,Mail の各サーバの有無」に関して,大きな変化が見られなかった.この為,「逆引き」「Web,Mail の各サーバの有無」に関して依然としてボットネットドメインを特徴付けるデータとして一定の有効性が保たれていると考えられる一方で,「TTL 値」「minimum 値」には値の一定の値に集中する傾向が見られたが, 「TTL 値」に関しては,ノーマルドメインと比較した場合,比較対象として有効とは言えない結果であった.

4　併用実験による検証と評価

前 3 章の結果を基に,数量化理論 2 類による検知実験を行った.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 Ver3.0」を使用した.

ここで,判別式に用いるパラメータ数とその組み合わせを設定するパラメータ設定実験では,下記の 3 種の組み合わせを用いて検証を行った.

1. 2010 年度データを用いる検証
2. 2 年分のデータを併用する検証
3. 3 年分のデータを併用する検証

本章では,これらの組み合わせの実験で,累積したデータの検知方式への影響を検証する.

4.1　2011 年度データでの試行

検証実験に用いるパラメータ数と組み合わせを,2011 年度データを用いたパラメータ設定実験で求める.実験に使用したパラメータ設定値を以下の表 1 に示す.

表1. パラメータ設定値 1

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

パラメータ設定実験では,最適なパラメータの数を調べる為,赤池情報量基準(以下,AIC とする)[6][7]を用いる.

AIC は,元統計数理研究所所長の赤池弘次によって考案された, 統計モデルの良さを評価するための指標である.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.データを統計的に説明する数式では,パラメータの数を増やせば,測定データとの適合度が高くなる.しかし,無意味なノイズの影響を多く受

ける事に繋がり,信頼性が低下してしまう.このような問題に対して,AIC は式(1)によって求められる最小 AIC 時のパラメータ数を選択する事で,多くの場合,良いモデルを選択する事が可能となる.

$$AIC=-2\ln L+2k \quad 式(1).$$

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．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.

また,今回新たに比較対象としてベイズ情報量基準(以下,BIC とする)を,最適パラメータ数の選定に用いた.BIC は,AIC 同様,統計における情報量基準の一つである.BIC は AIC 同様,式(2)によって求められる最小 BIC 時のパラメータ数を選択する事で,多くの場合に良いモデルを選択する事が可能となる.

$$BIC=-2\ln(L)+k\ln(n) \quad 式(2).$$

使用するドメインのデータは,前 3 章で示したように,ボットネットドメインが 42 個,ノーマルドメインが 42 個有る.パラメータ設定実験,検証実験,それぞれの使用するドメイン数は均等な数が望ましい.よって各ドメインを任意に選択し,両ドメインを 21 個ずつ用いた.

なお,以降出てくる検知率は,ボットネットドメインにおける True Positive と,ノーマルドメインにおける True Positive を合わせた値とする.

4.1.1 パラメータ設定実験結果

パラメータ設定実験結果から AIC と BIC 双方を求めたところ,共に最適なパラメータ数 3 という結果を得られた.表 2 にパラメータ数 3 における試行結果を示す.この結果より,もっとも検知率が高い 6 組について検証実験を行った.

検証実験の結果を表 3 に示す.結果,最も検知率が高い組み合わせは,「逆引き」「TTL 値」「登録期間」で,76.2%となった.

表2. 節 4.1 パラメータ設定実験結果

逆引き	TLL	minimum				76.2%
逆引き	TLL		Mailサーバ			71.4%
逆引き	TLL			Webサーバ		76.2%
逆引き	TLL				登録期間	76.2%
逆引き		minimum	Mailサーバ			76.2%
逆引き		minimum		Webサーバ		69.0%
逆引き		minimum			登録期間	73.8%
逆引き			Mailサーバ	Webサーバ		76.2%
逆引き			Mailサーバ		登録期間	66.7%
逆引き				Webサーバ	登録期間	76.2%
	TLL	minimum	Mailサーバ			71.4%
	TLL	minimum		Webサーバ		69.0%
	TLL	minimum			登録期間	71.4%
	TLL		Mailサーバ	Webサーバ		73.8%
	TLL		Mailサーバ		登録期間	71.4%
	TLL			Webサーバ	登録期間	71.4%
		minimum	Mailサーバ	Webサーバ		33.3%
		minimum	Mailサーバ		登録期間	64.3%
		minimum		Webサーバ	登録期間	61.9%
			Mailサーバ	Webサーバ	登録期間	64.3%

表3. 節 4.1 検証実験結果

逆引き	TLL	minimum				76.2%	71.4%
逆引き	TLL			Webサーバ		76.2%	71.4%
逆引き	TLL				登録期間	76.2%	76.2%
逆引き		minimum	Mailサーバ			76.2%	71.4%
逆引き			Mailサーバ	Webサーバ		76.2%	69.0%
逆引き				Webサーバ	登録期間	76.2%	73.8%

4.2 データによる併用試行

検証実験に用いるパラメータ数と組み合わせを求める為,節 4.1 で用いたデータに加え,2010 年度取得したデータを併用する実験と，2009 年度取得したデータを併用する実験,それぞれを行った.この際に用いたデータは,節 4.1 で使用したドメインに,CCCDATAset2010 から取得したボットネットドメインの半数に当たる 15 個,CCCDATAset2009 から取得したボットネットドメインの半数に当たる 9 個をそれぞれ加えたモノである.節 4.1 同様 AIC と BIC を求めたところ,共に最適パラメータ数 3 という結果を得られた.

表 4 に 2 年分のデータ併用による試行結果を，表 5 にパラメータ数 3 における試行結果を示す.

表4. 節 4.2 パラメータ設定実験結果

逆引き	TLL	minimum				68.4%
逆引き	TLL		Mailサーバ			64.9%
逆引き	TLL			Webサーバ		66.7%
逆引き	TLL				登録期間	77.2%
逆引き		minimum	Mailサーバ			68.4%
逆引き		minimum		Webサーバ		63.2%
逆引き		minimum			登録期間	75.4%
逆引き			Mailサーバ	Webサーバ		64.9%
逆引き			Mailサーバ		登録期間	73.7%
逆引き				Webサーバ	登録期間	70.2%
	TLL	minimum	Mailサーバ			66.7%
	TLL	minimum		Webサーバ		70.2%
	TLL	minimum			登録期間	68.4%
	TLL		Mailサーバ	Webサーバ		70.2%
	TLL		Mailサーバ		登録期間	66.7%
	TLL			Webサーバ	登録期間	70.2%
		minimum	Mailサーバ	Webサーバ		29.8%
		minimum	Mailサーバ		登録期間	64.9%
		minimum		Webサーバ	登録期間	63.2%
			Mailサーバ	Webサーバ	登録期間	59.6%

表5. 節 4.3 パラメータ設定実験結果

逆引き	TLL	minimum				65.2%
逆引き	TLL		Mailサーバ			68.2%
逆引き	TLL			Webサーバ		72.7%
逆引き	TLL				登録期間	74.2%
逆引き		minimum	Mailサーバ			69.7%
逆引き		minimum		Webサーバ		69.7%
逆引き		minimum			登録期間	75.4%
逆引き			Mailサーバ	Webサーバ		64.9%
逆引き			Mailサーバ		登録期間	73.7%
逆引き				Webサーバ	登録期間	70.2%
	TLL	minimum	Mailサーバ			66.7%
	TLL	minimum		Webサーバ		70.2%
	TLL	minimum			登録期間	68.4%
	TLL		Mailサーバ	Webサーバ		70.2%
	TLL		Mailサーバ		登録期間	66.7%
	TLL			Webサーバ	登録期間	70.2%
		minimum	Mailサーバ	Webサーバ		29.8%
		minimum	Mailサーバ		登録期間	64.9%
		minimum		Webサーバ	登録期間	63.2%
			Mailサーバ	Webサーバ	登録期間	59.6%

この結果より,2 年分の併用では,項 4.1 と同様の結果となった.また,3 年分の併用では,最も検知率が高い組み合わせが「逆引き」「minimum 値」「登録期間」となり,検証実験を行ったところ,検知率は 71.4%となった.

4.3 併用実験に関する考察

3 種類の組み合わせで行った検証実験の比較結果を表 6 に示す. この際, 併せてボットネットドメインにおける False Negative の値を比較する. 多段追跡システムにおける第 2 段トレースバックシステムは, 前提としてボットドメインと疑わしいドメインの判別を行う. この為, ボットドメインの検出漏れを防ぐ為に False Negative の値が低いことが望ましい.

表6 実験結果比較

	設定実験	検証実験	FalseNegative
項4.1	76.2%	76.2%	33.30%
項4.2	77.2%	76.2%	33.30%
項4.3	75.4%	71.4%	30.4%

表から解るように, 節 4.2,4.3 とデータ数を増やしたところ, 検知率の招く結果となった. 一方で, 節 4.3 の False Negative は節 4.1, 節 4.2 と比べ改善した. これらから, 蓄積したデータが検知精度に及ぼす影響は, 安定していないが, 検出漏れの防止に一定の効果が見られる, しかしながら, 現状取得したデータを用いた実験では, 検知精度の向上が難しいと考えられる. そこで次章では, 従来使用してきたパラメータ候補について, 追加・変更することで, 検知精度の向上を計る.

5 新規パラメータの検証と評価

今回,従来使用してきた数量化理論 2 類を用いた検知方式に使用するパラメータに加えて,以下の項目を追加調査した.

1. NSレコード
2. CNAMEレコード
3. Aレコードに関する調査の変更

これらの調査はそれぞれ,1 は,DNS サーバから取得した,委任するドメインの情報を持つネームサーバの指定をする NS レコードの個数の調査.2 は,ホストに割り当てられた名前から正規の名前を取得する際に使用する CNAME レコードの個数の調査である.これらの調査は,従来収集していた SOA レコード同様に,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.また,3 に関しては,従来「Web サーバの有無」としてきた,IP アドレスとの関係付けを行う A レコードの個数に関しての調査へ変更した.これは,昨年度からであるが Web サーバの有無のみでは, 2 種のドメイン間に差が無いことが確認できる為,有効性がある候補とする為である.これらのパラメータ設定値は,以下の表 7 に示す通りである.

表7 パラメータ設定値 2

NSレコード	値
有り	1
無し	2

CNAMEレコード	値
有り	1
無し	2

Aレコード個数	値
0	1
1～2	2
3～	3

5.1 パラメータ設定実験結果

パラメータ設定実験から AIC を求めたところ, 最適なパラメータ数 4 という結果を得た. 従来の方式では,すべてパラメータ数3という結果が出ている.この変化を比較するために,パラメータス 4 で最も検知率が高い 6 組,並びに,パラメータ数 3 で最も検知率が高い1組について検証実験を行う.

5.2 検証実験結果

検証実験の結果を表 8・表 9 に示す.結果,パラメータ数4では,最も検知率が高い組み合わせは二組存在し,「逆引き」「minimum 値」「A レコード」

に,「TTL 値」又は「登録期間」を加えたモノであった.これらの検知率は,97.6%と非常に高い検知率であり,検知漏れを表すFalse Negativeの値は0%と,検知漏れが無いことが確認できる.一方で, パラメータ数 3 では, 最も検知率が高い組み合わせでも 90.5%とパラメータ数 4 には劣るが高い値を示し,False Negative の値も 9.5%と高く無い結果ではあるが,パラメータ数 4 には劣る結果となった.

表8 5 章検証実験結果 1

逆引き	TLL	minimum			Aレコード			97.6%
逆引き		minimum		登録期間	Aレコード			97.6%
	TLL	minimum		登録期間	Aレコード			95.2%
		minimum	Mailサーバ	登録期間	Aレコード			92.9%
		minimum		登録期間	Aレコード	NSレコード		90.5%
		minimum		登録期間	Aレコード		CNAMEレコード	90.5%

表9 5 章検証実験結果 2

		minimum		登録期間	Aレコード個数		90.5%

5.3 ブラックリストを併用した検証

次に,表 10 に示す比較結果は,前章までの実験結果に加えて,2010 年 1 月から 2011 年 4 月の期間中に公開されたブラックリストを用いたブラックリスト方式でのマッチング結果を併せた検出結果の比較である.

表10 ボットネットドメイン検知結果

	項4.1	項4.2	項4.3	5章	BL		項4.1	項4.2	項4.3	5章	BL
1	○	○	○	○	×	11	×	○	○	○	○
2	○	○	○	○	○	12	○	○	×	○	○
3	○	○	○	○	○	13	○	○	○	○	○
4	○	○	○	○	×	14	○	○	○	○	×
5	○	○	○	○	○	15	○	○	○	○	×
6	○	○	○	○	○	16	○	○	○	○	○
7	×	×	×	○	×	17	○	○	×	○	×
8	×	×	×	○	×	18	○	○	○	○	○
9	×	○	○	○	○	19	×	×	×	○	○
10	×	○	×	○	○	20	○	○	×	○	○
						21	×	○	○	○	○

結果, パラメータ設定実験において, 2011 年度データを使用した項4.1では7カ所, 2010年度データを併用した項4.2では3カ所, 2009年度データを併用した項4.3では7カ所検出漏れが発生した. ここに, ブラックリストによる検知を併用した場合でも, 項4.1で2カ所, 項4.2で2カ所, 項4.3では3カ所検出漏れが発生した.

これらの結果と各検知率の比較から, データの併用が必ずしも検出精度の向上に有効であるとは言えない結果となった. これは, 時間経過と共にボットネットの特徴自体が変化し, 従来収集してきたドメインデータだけでは, データの併用をしても対応が難しくなったと考えられる. この点, パラメータ候補の変更により, ボットネットドメインの検知について, 一定の成果が得られ, 改善した.

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, 3 年間に渡るデータの検知方式への影響の検証を行った. 実験により, 検出率を検証したが, データが増えるほど検出精度が不安定になり, 検知率が低下することが確認された.

しかし, パラメータ候補の追加を行い検証したところ, 検出精度は極めて高い値を得る結果となった. その理由は, ボットネットの特徴が変動し, 従来用いていたパラメータ候補では対応できなくなっているが, 変更する事で検出精度の改善につながったと考えられる.

今後, パラメータに用いる特徴ある項目の調査と共に, 特徴変動に対応できるシステムを検討する.

参考文献

[1] 「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

[2] ボットネット概要, http://www.jpcert.or.jp/research/2006/Botnet_summary_0720.pdf

[3] 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.

[4] 中村暢宏, 名雲孝昭, 田中達哉, 三原元,佐々木良一, 「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」, マルウェア対策研究人材育成ワークショップ 2010 (MWS2010), 3F2-3, 2010 年 10 月.

[5] 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 2011 Datasets～, MWS2011(2011 年 10 月)

[6] 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」

[7] 原口正行のホームページ EXCEL を使った多変量解析