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

三原　元[†]　　佐々木　良一[‡]

[†,‡]東京電機大学大学院　工学研究科　情報メディア学専攻

あらまし 近年，ボットネットによる被害が問題になっている．ボットネットは，ボット PC を特定・隔離しても別の PC がボットになり，根本的な解決にならない．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．今回，多段追跡システムの第二段での追跡手法として，C&C(Command&Control)サーバに関するブラックリストを用いる検知方式と，CCC DATAset 2009 の解析結果を数量化理論２類に適用する検知方式の２方式を用いることで C&C サーバの特定を行う手法を考案した．本論文では，C&C サーバのブラックリストを用いた方式と数量化理論２類を用いた方式に関して述べると共に，２つの方式を用いた多段追跡システム第二段の提案とその評価結果の報告を行う．

Proposal and Evaluation of Technique to Detect C&C Server on Botnet Using CCC DATAset 2009 and Quantification Methods

Hajime Mihara[†]　Ryoichi Sasaki[‡]

[†‡]Information Systems and Multimedia Design,　Graduate School of Engineering,
Tokyo Denki University

Abstract　Recently, damage caused by the botnet becomes a big problem.　There exists a problem that the other bot PCs can be produced, even if one bot PC could be specified and removed.　Therefore,　it is not a fundamental solution to specify only the bot PC.　To solve this problem authors proposed the multistep trace back system.　In this paper, the authors developed the second step system which consists of two methods: first method is to use the black list of the C&Cserver and the second one is to use analytical result using quantification methods No. 2 and CCCDATAset2009. In this paper, we describe the second step system using two methods and the result on the evaluation using CCCDATAset2009.

1 はじめに

近年，ボットネット[1]による被害が問題になっている．ボットネットとは，ボットと呼ばれるマルウェアに感染した PC(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る．ボット PC は，C&C 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)からの命令を受け取り，様々な活動を行う．

現在では，数百台から数万台のボット PC から構成されるボットネットが確認されている[2]．このボットネットは，ハーダーからの命令により，複数のボット PC が一斉に DoS(Denial of service)攻撃を行う，DDoS(Distributed DoS)攻撃に利用される．

ボットからの攻撃パケットは送信元 IP アドレスが偽装されていることがあり，ボット PC の特定は困難である．その問題に対処するため，IP トレースバックシステム[3]が提案されている．しかしながら，IP トレースバックシステムだけでは，攻撃パケットを送信しているボット PC の特定はできても，ハーダーや C&C サー

バを特定することはできない.

そこで現在著者らは,ネットワーク管理者同士が情報共有を行い，ボット PC の特定だけではなく，C&C サーバやハーダーの操作する PC の特定を目的とした,多段階トレースバックシステム[4]を構想している.

本論文では,多段追跡システムのうち,第二段トレースバックシステムの提案を行うと共に，第二段トレースバックシステムにおける C&C サーバの特定手法として，C&C サーバのブラックリストを用いる検知方式と，CCCDATAset2009[5]の解析結果を,数量化理論２類に適用する検知方式の二方式について述べ,その評価を行う.

2 提案システム

2.1 数量化理論2類

数量化理論は,統計数理研究所出身の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.

数量化理論２類は，２つのタイプＡとＢがばらばらにあるとき,パラメータの係数を適切に設定することで，ＡはＡ同士，ＢはＢ同士で近い値を取るようにし,かつＡとＢは遠い値を取るようにしようとする．よって，A と B の明快な境界線の設定が可能となる．また，未知のデータに対しては,パラメータの係数の値を用いて，A，Bどちらに属する可能性が強いかを推定することが可能である.

2.2 提案システム概要

第二段トレースバックシステムの構成図(図1)と処理の流れ(図 2)を以下に示す.

ただし，第二段トレースバックシステムは,ポートミラーリング機能等によって,ルータを通過する全てのパケットを取得可能とする.

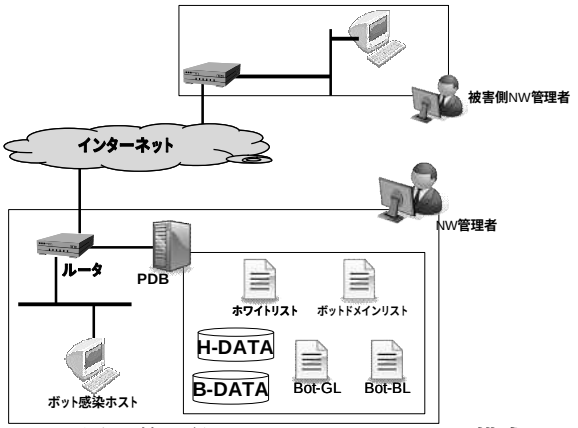

図 1 第二段トレースバックシステム構成

・**Bot-GL**
ボットと疑わしきホストの IP アドレスリスト

・**Bot-BL**
ボットと確定したホストの IP アドレスリスト

・**ホワイトリスト**
ボットネットに関係しないホストの IP アドレスリスト

・**ボットリスト**
インターネット上から取得したボットネットに関するドメインのリスト

・**H-DATA**
取得した全ての通信データを一定時間記録する

・**B-DATA**
Bot-GL の IP の通信データを記録する

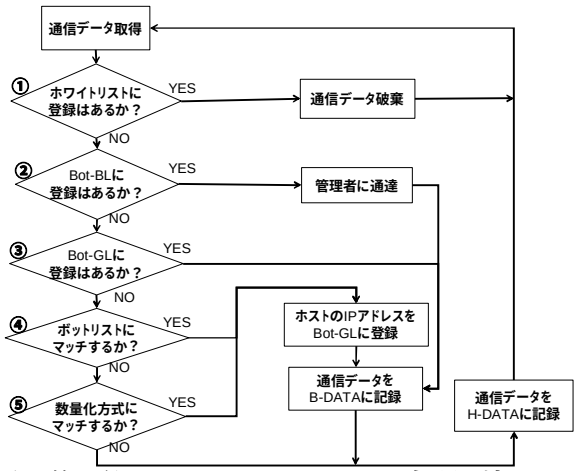

図 2 第二段トレースバックシステムの処理の流れ

① 通信ホストの IP アドレスがホワイトリストにあるか確認する．リストにあればその通信データを破棄し，リストに無ければ②に移行する

② 通信ホストの IP アドレスが Bot-BL にあるか確認する．リストにあれば管理者に通達し，その通信データを B-DATA に記録する．リストに無ければ③に移行する

③ 通信ホストの IP アドレスが Bot-GL にあるか確認する．リストにあればその通信データを B-DATA に記録し，リストに無ければ④に移行する

④，⑤に関しては以降の項にて述べる.

2.2.2 ボットリストを使用した検知方式

④では,通信ホストのドメインがボットリストにあるか確認する.リストにあればそのホストの IP アドレスを Bot-GL に登録し,通信データを B-DATA に記録する．ボットリストは，C&C サーバ等のブラックリストであり,インターネット上で公開されている C&C サーバ等のドメインのリストを，一定時間ごとに取得し，更新する．よって，この方式は，日々新たに C&C サーバ等が更新されるボットネットに対して有効である.

2.2.3 数量化理論2類を用いた検知方式

⑤では, 数量化理論２類を用いた検知を行う. 2.2.2 項のボットリストを使用した検知方式では, ボットリストに記載されていないホストは検知することができない. そのため, 今回著者らは, ボットリストを使用した検知方式に加え, 数量化理論２類を用いた検知方式を用い, さらなる検出精度の向上を目指す.

数量化理論２類を用いるためには, 判別式に入力するパラメータを決定する必要がある. そのため, 今回著者らは, CCCDATAset2009 の３種類のデータのうち, 攻撃通信データを解析し, ボットネットの特徴を調査する. そして, 調査の結果得たボットネットのデータを数量化理論２類のパラメータとして使用することを目指す.

次章にて, CCCDATAset2009 の解析を行い, 数量化理論２類を用いた検知方式のパラメータ候補を求める.

3 CCCDATAset2009 の解析

数量化理論２類を用いた検知方式のパラメータ候補を求めるため, CCCDATAset2009 のデータ解析を行った. 解析には, ボットネットに関する通信データと, ボットネットに関係しない通信データの２つを用いた. 今回著者らは, 通信データから直接ボットネットの特徴を調査するのではなく, 通信データ中の各ボットの接続先ホストのドメインに着目し, そのドメインに関する特徴の調査を行った.

まず, CCCDATAset2009 の攻撃通信データから, ボットネットに関するドメインを取得し, 以下の手順で２４個のドメイン(以下, ボットネットドメインとする)のデータを得た.

①通信データ中の DNS クエリを取得
②取得した DNS クエリの内, a)-c)を除外
  a)IP アドレスのもの
  b)Hotmail 等明らかに正規サービスのもの
  c)自 NW ドメインの問い合わせのもの
③重複したドメイン名を除外

さらに, ボットネットに関係しない通信データとして, 本研究室のネットワークから取得した通信データ(表 1)のうち, 任意に選択した DNS 通信から, ５０個のドメインデータ（以下ノーマルデータとする）を得た.

表１　ノーマルデータ取得元通信データ

PC台数	20台 (OS:WindowsXP)
通信データ取得時間	24時間
パケット数	約50万パケット

今回著者らは, 既存研究[2][5]を参考として, 数量理論２類を用いた検知方式のパラメータ候補として, 以下の 5 項目を選択した.

①逆引き
②SOA レコード
③WHOIS
④mail,www サーバの有無
⑤TTL 値

逆引きとは, DNS サーバに対して IP アドレスからホストのドメイン名を問い合わせることであり, SOA レコードとは, 各 DNS サーバが管理するドメインの, 設定情報である. また, WHOIS とは, 各レジストリ組織が管理する, ドメインや IP アドレス, AS 番号の管理者情報であり, TTL 値とは, 各 DNS サーバが管理するドメイン情報の, 取得した際の生存時間である. さらに, 今回著者らは, ボットネットに関するドメインには, 不必要なサービスは登録されていないのではないか, という推測に基づき, 同一ドメイン上での mail, www サーバの有無の調査を行う.

次節にて, 各パラメータ候補の調査結果に関して述べる.

3.1 逆引き

ボットネットドメインとノーマルドメインの逆引き情報を調査した(表 2).

表 2　ドメインの逆引き結果

	ノーマルドメイン		ボットネットドメイン	
逆引き結果が正しい	9個	(18%)	2個	(8%)
逆引き結果が正しくない	34個	(68%)	5個	(21%)
返答なし	7個	(14%)	17個	(71%)
合計	50個	(100%)	24個	(100%)

調査の結果, ボットネットドメインとノーマルドメインでは, 逆引きの結果に明らかな差異があることがわかった. 具体的には, ボットネットドメインでは逆引きができない割合が多く, ボットネットドメイン全体の７割を占めたのに対し, ノーマルドメインでは逆引きの結果が正しくない割合が高く, ノーマルドメイン全体の７割であった.

3.2 SOA レコード

SOA レコードには以下の情報が含まれる.

a) ゾーン名
b) ネームサーバホスト名
c) 管理者メールアドレス
d) シリアル番号
e) Refresh 値
f) Retry 値
g) Expire 値
h) Minimum 値

今回著者らは, ボットネットドメインとノーマルドメインの SOA 情報の内, h)の Minimum 値を調査した(図 3). これは, ボットネットドメインの更新頻度はノーマルドメインの更新頻度より多いと言われているため, ネガティブキャッシュとして機能する h)の値が, ノーマルドメインの値と比較して差異が期待できると考えたためである.

図 3 より, ノーマルドメインがある一定の値に集中して Minimum 値が設定されている傾向にあることがわかった. 対して, ボットネットドメインにはそのような傾向は見られなかった.

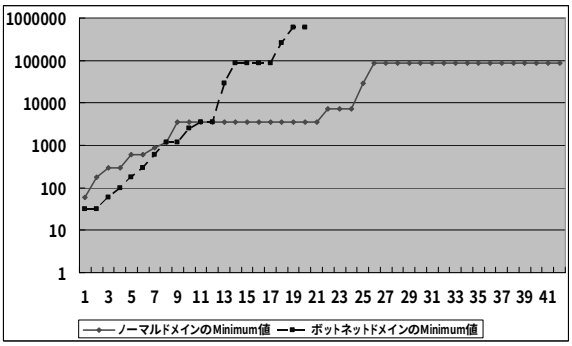

図 3 Minimum 値

3．3 WHOIS

WHOIS からは, 一般的に以下の情報を得ることができる.

a) 登録ドメイン名
b) レジストラ名
c) ドメインが登録されている DNS サーバ名
d) ドメインの登録年月日
e) ドメインの有効期限
f) ドメイン名登録者の名前や住所
g) 技術的な連絡の担当者連絡先
h) 登録に関する連絡の担当者連絡先
i) 登録者への連絡窓口の連絡先

今回著者らは, 上記の内, d)と e), そして d)と e)の差分である j)ドメインの登録期間に関して調査を行った(図 4, 図 5, 図 6).

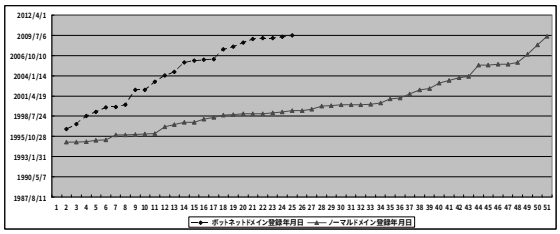

図 4 ドメインの登録年月日

調査の結果, d)と e)には差異が見られなかったが, j)に関しては, ノーマルドメインよりも, ボットネットドメインの登録期間が比較的短い傾向にあるということがわかった.

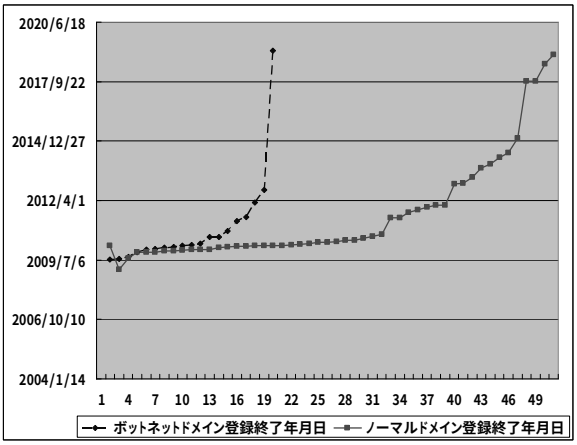

図 5 ドメインの登録終了年月日

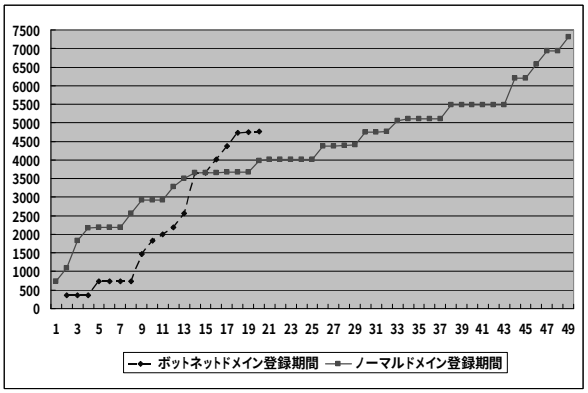

図 6 ドメインの登録期間

3．4 mail,www サーバの有無

ボットネットドメイン, ノーマルドメインに対し, 同一ドメインにおける mail サーバ, ウェブサーバの登録の有無の調査を行った(表 3, 表 4).

表 3 ウェブサーバの有無

	ノーマルドメイン		ボットネットドメイン	
ウェブサーバ有	45個	(90%)	13個	(54%)
ウェブサーバ無し	2個	(4%)	4個	(16%)
返答なし	3個	(6%)	7個	(30%)
合計	50個	(100%)	24個	(100%)

表 4 Mail サーバの有無

	ノーマルドメイン		ボットネットドメイン	
Mailサーバ有	36個	(72%)	14個	(59%)
Mailサーバ無し	10個	(20%)	3個	(12%)
返答なし	4個	(8%)	7個	(29%)
合計	50個	(100%)	24個	(100%)

調査の結果, ノーマルドメインとボットネットドメイン共に, 同一ドメイン上にウェブサーバおよびメールサーバが登録されている割合が高く, 2つのドメインにおける差異は見られなかった.

3.5 TTL 値

ノーマルドメインとボットネットドメインの TTL 値を調査した(図 7).

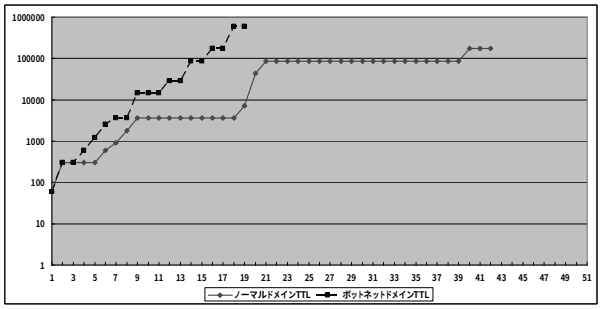

図 7　TTL 値

調査の結果, ノーマルドメインの TTL 値は, 特定の値に集中して設定されている傾向にあるのに対し, ボットネットドメインの TTL 値はノーマルドメインのような傾向が見られない, ということがわかった.

4 実験による最適パラメータの設定と評価

4.1 実験方法

まず, パラメータ設定実験により, 2.2.3 項の数量化理論２類に用いる, 最適なパラメータの組み合わせを選択する. パラメータ設定実験では, 3 章の各パラメータ候補に対し, 数値を設定することにより, 数量化理論２類に適用し, 試行を行う. 試行は, パラメータの全ての組み合わせに対して行う. そして, 試行結果の内, 判別率が高く, かつパラメータの組み合わせ数が最小のものを選択する.

次に, パラメータ設定実験にて選択した組み合わせに対して, 検証実験を行う. これは, パラメータ設定実験で出力された判別率が, 判別式に使用したドメインのデータを, 再度判別式に入力して得られた値であるためである. よって, 検証実験では, 判別式にパラメータ設定実験で使用したデータ以外のドメインのデータを入力し, 同様の判別率が得られるか, 検証を行う.

そして, パラメータ設定実験と検証実験の結果を比較し, 最も検出率の高いパラメータの組み合わせを選択し, 2.2.3 項の数量化理論２類を用いた検知方式に使用するパラメータの組み合わせとする.

ここで, 数量化理論２類の適応にあたり, 株式会社エスミ社のソフトウェア EXCEL 数量化理論[7]を使用した.

4.1.1 実験パラメータ設定値

3 章の調査結果を用いて, 実験を行う際に必要となる, 数量化理論２類の判別式のパラメータとする(表 5, 表 6, 表 7).

表 5 ドメイン登録期間, ドメイン登録終了年月日のパラメータ設定値

ドメイン登録期間	設定値
1-2500 (日)	1
2501-5000 (日)	2
5001-8000 (日)	3
NA	4

ドメイン登録終了年月日	設定値
-2010.09.13	1
2010.09.14-	2
NA	3

表 6　mail サーバ, ウェブサーバのパラメータ設定値

Mailサーバ	設定値
あり	1
なし	2

ウェブサーバ	設定値
あり	1
なし	2

表 7 逆引き, TTL 値, Minimum 値のパラメータ設定値

逆引き	設定値
返答無し	1
返答が正しくない	2
返答が正しい	3

TTL値	設定値
1-1000	1
1001-100000	2
100001-1000000	3
NA	4

Minimum値	設定値
1-100	1
101-1000	2
1001-100000	3
NA	4

4.2 実験データ

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 19 個あり, ノーマルドメインは 50 個ある. よって, パラメータ設定実験にはボットネットドメインを 10 個, ノーマルドメインを 20 個使用する. 検証実験には, ボットネットドメインは残りの 9 個, ノーマルドメインは残り 30 個から任意に 20 個選択し, 使用する.

4.3 実験結果

4.3.1 パラメータ設定実験結果

実験結果を表 8 に示す. ただし, 表 8 は以下の３つの条件を満たしたパラメータの組み合わせである. 今回の実験では, a), b)を満たす最小のパラメータの組み合わせ数は２つであった.

a) ボットネットドメインの検出精度が 80%以上
b) ノーマルドメインの検出精度が 80%以上
c) パラメータの組み合わせ数が最小

表 8 実験結果

パラメータの組み合わせ		ボットネットドメイン検出精度	ノーマルドメイン検出精度
ドメイン登録期間	ウェブサーバの有無	80	80
ドメイン登録期間	逆引き	90	80
ドメイン登録終了年月日	逆引き	80	80

表 8 より，a)，b)，c)を満たし，かつボットネットドメインの検出率が最も高いパラメータの組み合わせは,「ドメイン登録期間」と「逆引き」の２つのパラメータの組み合わせであることがわかった.

また, この２つのパラメータの組み合わせにおける, ボットネットドメインとノーマルドメインを合わせた全体の検出率は 83%であり, 誤検出率は 17%であった.

４．３．２ 検証実験結果

検証実験の結果(表 9)，4.3.1 項の a)，b)，c)の条件を満たす 3 つのパラメータの組み合わせの内, 最も高い検出率を示したパラメータの組み合わせは,「ドメイン登録期間」と「逆引き」の２つのパラメータの組み合わせであることがわかった.

また, この２つのパラメータの組み合わせにおける, ボットネットドメインとノーマルドメインを合わせた全体の検出率は 86%であり, 誤検出率は 14%であった.

表 9 検証実験結果

パラメータの組み合わせ		ボットネットドメイン検出精度 (%)	ノーマルドメイン検出精度 (%)
ドメイン登録期間	ウェブサーバの有無	100	35
ドメイン登録期間	逆引き	88	85
ドメイン登録終了年月日	逆引き	55	85

このパラメータの判別累積グラフを図 8 に示す.

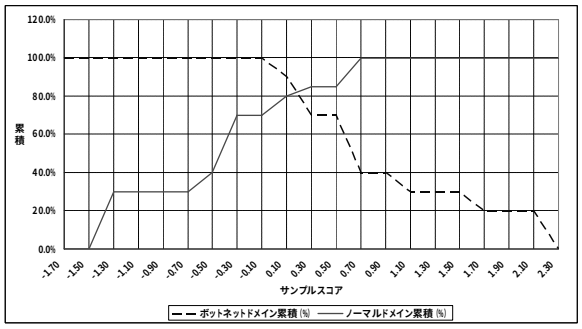

図 8　累積判別グラフ

検証実験では, パラメータ設定実験と同じパラメータの組み合わせが, 検出率, 誤検出率共に比較的良い値を示した. これらの値は, 一般には良好な値とは言えない. しかし, C&C サーバの候補としてチェックの対象とする場合には十分な値である考えられる.

以上より，2.2.3 項の数量化理論に用いる最適なパラメータは, 以下の 2 つのパラメータの組み合わせとなった.

・ドメイン登録期間
・逆引き

5 おわりに

本論文では, 多段追跡システムのうち, 第二段トレースバックシステムを提案し, 第二段トレースバックシステムにおける C&C サーバの特定手法として, ブラックリスト方式と数量化理論２類を用いた検知方式に関して述べた. また, 実験により, 数量化理論２類に用いる最適なパラメータの選定を行い, その有効性を確認した.

今後は, 第二段トレースバックシステムの実装を目指すと共に, ボットネットの特徴が変更された場合に対して, 動的に対応可能な検知システムの運用を目指す.

参考文献

[1]ボットネット概要
http://www.jpcert.or.jp/research/2006/Botnet_summary_0720.pdf
[2]高橋正和, 村上純一, 須藤年章, 平原伸昭, 佐々木良一,「フィールド調査によるボットネットの挙動解析」, 情報処理学会論文誌, Vol.47, No.8，2007
[3] 藩博文, 佐々木良一:IP トレースバックのための出国印方式の試作と評価, 情報処理学会論文誌, Vol.49, No.9, (2008).
[4]三原元, 名雲孝昭, 芦野祐樹, 上原哲太郎, 佐々木良一,「ボットネットの多段追跡システムの構想と CCCDATAset2008 の利用手法」MWS2008
[5] 畑田充弘, 中津留勇, 寺田真敏, 篠田陽一,「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」MWS2009
[6] Malware Threat Center
http://www.mtc.sri.com
[7]株式会社エスミ
http://www.esumi.co.jp
[8]林知己夫, [数量化-理論と方法], 朝倉書店, 1993 年