情報処理学会論文誌　Vol. 51　No. 9　1579–1590 (Sep. 2010)

数量化理論と攻撃データ（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[†1]　　佐々木　良一[†1]

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 2 段での追跡方式として，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方式を考案した．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ともに，2 つの方式を用いた，多段追跡システム第 2 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 Attack Data (CCCDATAset 2009) and Quantification Methods Type II

Hajime Mihara[†1] and Ryoichi Sasaki[†1]

Recently, damage caused by the botnet has become 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, the authors proposed the multistep trace back system. In this paper, the authors deals with the second step system which consists of two methods. Here, in first method, the black list of the C&C server is used and in the second one, analytical result using quantification methods No.2 and CCCDATAset2009 is used. In this paper, we report the proposed 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．現在では，ボットネットを構成するボット PC が数百台から数万台のものが確認されている[2)]．ボット PC は，C&C サーバと呼ばれる中継サーバを介し，ボットネットを操作する攻撃者（以下，ハーダとする）からの命令を受け取り，DDoS（Distributed Denial of Service）攻撃に参加したり，スパムメールを送ったりするなど，様々な活動を行う．ここでボット PC は，ハーダからの命令によって，ダウンローダから不正プログラムをダウンロードすることにより，時間とともに変化するなどということも行っている．

ボット PC が，送信元を詐称して攻撃を行った場合，特定は困難である．そのような攻撃に対しては，IP トレースバック[3)]などの方法を用いることで，発見が可能である．しかし，たとえボット PC を発見して対策を行っても，対策が不十分だと，その PC がすぐにボットに感染したり，別の PC がボットに容易に感染したりするため，本来的な対策とはならない．

このような問題に対しては，ハーダを見つけて対処を行わない限り本格的に解決されることはない．そこで現在著者らは，企業や大学などの，ネットワーク管理者同士が情報共有を行うことにより，(1) ボット PC の特定だけでなく，(2) C&C サーバやダウンローダ，(3) そしてハーダの操作する PC の特定を目的とした，多段追跡システム[4)]を構想している．

本論文は，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関するものであり，C&C サーバとダウンローダの 2 つ（以下，第 2 段追跡対象とする）の特定手法を提案する．この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 つの方式を組み合わせたものである．ここでは，数量化理論 2 類を用いるため，CCCDATAset2009[5)]の一部を実験に使用する．CCCDATASet2009 とは，サイバークリーンセンター（CCC）で収集しているデータをもとに作成した，研究用データセットである．このデータセットは，「マルウェア検体」，「攻撃通信データ」，「攻撃元データ」の 3 つのデータから構成される．「マルウェア検体」とは，ハニーポットで収集したマルウェア検体のハッシュ値（MD5，SHA1）10 個を，テキスト型式で記載したファイルである．「攻撃通信データ」とは，2 台のハニーポットの通信を，2009 年 3 月 13 日と 2009 年 3 月 14 日の

†1 東京電機大学
Tokyo Denki University

© 2010 Information Processing Society of Japan

2 日間でパケットキャプチャした，lippcap 形式のファイルである．「攻撃元データ」とは，94 台のハニーポットが記録したマルウェア取得時のログであり，マルウェア取得時刻，送信元/先 IP アドレス，送信元/先ポート番号などで構成された，csv 形式のファイルである．

今回は実験に用いたのとは別の CCCDATASet2009 のデータを適用して行った提案方式の有効性の評価結果もあわせて報告する．

2. 関 連 研 究

ボットネットにおける，第 2 段追跡対象の特定を目的とした研究は，特定方式により次の 2 種類に分類できる．

（1）C&C サーバとボット PC 間で行われる制御通信の検知による特定方式

C&C サーバとボット PC 間で行われる制御通信を利用し，制御通信中に含まれる特徴的な文字列を検知して，特定を行う方式[6),7)] である．

しかしこの方式は，暗号化された制御通信が行われたり，C&C サーバに IRC プロトコル以外が使われたりしているような場合の検知は非常に困難である．したがって，本研究では，制御通信を利用した特定方式は使用していない．

（2）第 2 段追跡対象のドメイン名を用いて特定を行う方式

ボットネットの中には，ボット PC が，DNS サーバに対して第 2 段追跡対象の名前解決を行うものがある．そこで，朝長ら[8)] や Schonewille ら[9)] は，第 2 段追跡対象のドメイン名をブラックリストとして使用し，ブラックリストのドメイン名と対象となったドメイン名のマッチングを行うことで検知を行う方式を提案している．

本論文で提案する方式は，（a）ブラックリストと（b）数量化理論 2 類による解析を組み合わせる方式であり，このうちブラックリスト方式は，朝長らや Schonewille らの方式と基本的に同じである．しかし，第 2 段追跡対象の，ドメインの登録期間やドメインの逆引きの結果といった，ドメインの特徴に関し，数量化理論 2 類を用いて解析する方式は従来提案されていない．したがって，このような，ブラックリスト方式とともに，数量化理論 2 類を組み合わせて用いる方式は，既存方式にはなく，新しい方式であるといえる．

3. 組み合わせる 2 つの検知方式

3.1 ブラックリストを使用した検知方式

現在，ボットネットに関するホストの IP アドレスやドメイン名を公開しているサイトが複数存在する．今回著者らは，これらのサイトで公開しているドメイン名の一覧を，一定時間ごとに取得し，ブラックリストを作成する．そして，作成したブラックリストとマッチングを行うことで，第 2 段追跡対象の検知を行う方式（以下，ブラックリスト方式とする）を使用する．

表 1　ブラックリスト方式の検知結果
Table 1　Detection result of blacklist method.

	検知	無検知	合計
ドメイン数(%)	15(80%)	4(20%)	19(100%)

3.1.1 ブラックリスト方式の検証

ブラックリスト方式の検知率の検証を行った（**表 1**）．検証に使用したブラックリストは，SRI Malware Threat Center[10)] と DNS-BH[11)] の 2 つのサイトから，2009 年 6 月 20 日に取得したドメインのリストである．また，ブラックリストとマッチングをさせる対象のドメインには，CCCDATAset2009 の攻撃通信データより，以下の手順で第 2 段追跡対象と考えられるドメイン（以下，ボットネットドメインとする）を 19 個取得した．

① 通信データ中の DNS クエリを取得
② 取得した DNS クエリのうち，a)–c) を除外
　a) IP アドレスのもの
　b) Hotmail など明らかに正規サービスのもの
　c) 自 NW ドメインの問合せのもの
③ 重複したドメイン名を除外

19 個のボットネットドメインは，攻撃通信データ中の DNS クエリから取得したドメインのうち，Microsoft や Google といった，明らかにボットネットとは関係がないと判断可能なドメインを除外したものである．

表 1 より，全 19 個のドメインのうち，15 個のドメインがブラックリスト方式により検知でき，4 個のドメインが検知できなかった．

3.1.2 ブラックリスト方式の評価

3.1.1 項より，ブラックリスト方式の検知率は 80%で，けっして高いとはいえない．また，ブラックリスト方式は，リストにないものは検知できない．しかし，ボットネットは，その特徴として，第 2 段追跡対象に用いるホストの変更頻度が高い．よって，ブラックリスト方式は，検知率だけでなく，ボットネットに対する検知方式としての適合性という観点から見ても，信頼性に欠ける．

そこで，今回著者らは，第 2 段トレースバックシステムにおける，第 2 段追跡対象の検

© 2010 Information Processing Society of Japan

知方式として，ブラックリスト方式とともに，数量化理論 2 類を用いた検知方式を組み合わせて用いる．そして，2 つの検知方式を用いて，第 2 段追跡対象の検知率の向上を目指す．

次節で，数量化理論 2 類を用いた検知方式に関して述べる．

3.2　数量化理論 2 類を使用した検知方式

3.2.1　数量化理論 2 類

数量化理論 2 類とは，分析対象データが数値化不可能な質的データにおいて，判別分析と同様の処理を行うものである[12]．数量化理論 2 類を使用するためには，まず与えられたデータからカテゴリスコアと境界値を設定するフィッティング段階と，フィッティングにより求められた 2 つの値を用いて，判別および予測の段階に分けられる．

フィッティング段階では，まず，サンプル i が j 番目の要因における k 番目の項目に反応するとき，

$$\delta_i(jk) = 1 \tag{1}$$

とし，反応していないとき，

$$\delta_i(jk) = 0 \tag{2}$$

とする．j 番目の要因における k 番目の項目に与える数値（以下，カテゴリスコアとする）を X_{ij} とすると，各サンプルスコア α_i は

$$\alpha_i = \sum_{j=1}^{J} \sum_{k_j=1}^{K_j} X_{jk_j} \delta_i(j, k_j) \tag{3}$$

ただし，$i = 1, \ldots, n,$　　n：サンプル数

$j = 1, \ldots, J,$　　J：要因数

$k_j = 1, \ldots, k_j,$　　k_j：j 番目の要因の項目数

と表せる．なお，この X_{ij} の求め方は以下のとおりである．

いま，α_i のサンプル全体の分散を σ^2 とする．そして α_i のサンプルは，いくつかのグループに属しているとし，このグループ間のサンプルスコア α_i に対する級間分散を σ^{b2} とすると，

$$\eta^2 = \sigma^{b2}/\sigma^2 \tag{4}$$

で表される相関比が最大となるように，式 (3) のカテゴリスコア X_{jk} を定める．

また，式 (3) で求めた各サンプルスコアが，どのグループに属するかの判別や推定を行うための比較値（以下，境界値とする）は，式 (3) から求められる各サンプルスコア α_i をもとに，最も判別率が高い値を設定する．

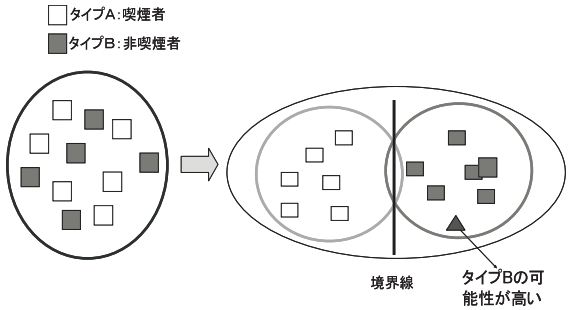

図 1　数量化理論 2 類の概要
Fig. 1　Outline of second quantification methods.

表 2　各サンプルの要因
Table 2　Sample factor.

	性別		飲酒	年齢		喫煙
	男	女		40以上	40未満	
a	0	1	1	0	1	0
b	0	1	0	1	0	0
⋮						
c	1	0	1	0	1	1
d	1	0	1	1	0	1

次に，判別および予測の段階では，式 (3) を用いてそれぞれのサンプルスコアを求め，サンプルスコアの値と境界値を比較する．そしてこのとき，サンプルスコアが境界値より高い値ならばタイプ A，低い値ならばタイプ B というように判別を行う．

たとえば，**図 1** のように，喫煙者（以下，タイプ A とする）と非喫煙者（以下，タイプ B とする）がバラバラに存在する集団に対し，数量化理論 2 類を用いてタイプ A とタイプ B に判別する試行を考える．

まず，カテゴリスコアと境界値を求めるために，タイプ A を特徴付ける要因を調査する（**表 2**）．そして，式 (4) で表される，相関比が最大のときの各カテゴリスコア（**表 3**）と境界値を求めることで，判別を行う．

このとき，要因 J は「性別群」「飲酒群」「年齢群」であり，カテゴリ k_i は「性別群」中の「男」「女」に該当する．また，α_i は集団中の各個人 a，b，c，…である．

© 2010 Information Processing Society of Japan

表 3　カテゴリスコア
Table 3　Category score table.

要因	カテゴリー	カテゴリースコア
性別	男	0.12
	女	-0.32
飲酒		0.11
年齢	40才未満	-0.14
	40才以上	1.55

よって，境界値が 1.55 ならば，a のサンプルスコア α_{a} は，式 (3) と表 2，表 3 より，

$$\alpha_{\mathrm{a}} = -0.32 + 0.11 - 0.14 = -0.35 < \text{境界値 } 1.55 \tag{5}$$

となり，a はタイプ B と判別することができる．

また，集団以外の未知の人物に対しても，同様にサンプルスコアと境界値を比較し，タイプ A，タイプ B どちらに属する可能性が強いかの推定を行うことが可能である．

3.2.2　数量化理論 2 類を用いた検知方式

提案方式では，数量化理論 2 類を用いて，第 2 段追跡対象かどうかの検知を行う．そのためには，フィッティング段階として，カテゴリスコアと境界値を求める．この 2 つを求めるために，あらかじめ第 2 段追跡対象の特徴を要素群とし，数値化した値（以下，パラメータとする）を設定する必要がある．そこで今回著者らは，CCCDATAset2009 の攻撃通信データの一部を利用し，数量化理論 2 類を用いて第 2 段追跡対象の特徴を解析する．そして，解析の結果得た第 2 段追跡対象の特徴をパラメータとして設定することで，カテゴリスコアと境界値を求める．

次章で，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．

4.　CCCDATAset2009 の解析

数量化理論 2 類を用いた検知方式に用いる，パラメータの候補を求めるため，CCCDATAset2009 のうち，攻撃通信データの解析を行う．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 種類のデータを用いる．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，各ボット PC が接続する，第 2 段追跡対象のドメインに関して調査を行う．

ボットネットドメインは，3.1.1 項と同様の手法で 19 個取得した．

表 4　ノーマルドメイン取得元通信データ
Table 4　Data obtained from normal domain.

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

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 4**）のうち，任意に選択した DNS 通信から，40 個のドメイン（以下，ノーマルドメインとする）を取得した．

今回著者らは，既存研究[2),5)] を参考として，数量化理論 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
(2)　SOA レコード
(3)　WHOIS
(4)　Mail，Web サービスの登録の有無
(5)　TTL 値

(1) は，DNS サーバに対して IP アドレスからドメイン名を問い合わせる調査である．(2) は，各 DNS サーバが管理する，ドメインの設定情報の調査である．(3) 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 番号の管理者情報の調査である．(4) は，同一ドメイン上の，Mail サーバ，Web サーバのドメイン名登録の有無の調査である．(5) は，DNS サーバから取得したドメイン情報の有効期間である TTL 値の調査である．

また，各項目の調査手法に関しては，(1)，(2)，(4)，(5) は調査対象となるドメイン情報を保持する DNS サーバに対し，dig コマンドを用いて各項目の調査を行った．(3) はレジストリである JPRS の WHOIS サービス[13)] を用いて調査を行った．

次節で，5 項目の調査結果に関して述べる．

4.1　逆　引　き

逆引き情報の調査の結果，**表 5** に示すように，ボットネットドメインでは逆引きの返答がない割合が多く，7 割を占めた．これに対し，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多く，7 割を占めた．

以上の結果より，逆引きの結果は，ボットネットドメインを特徴付けるデータとしての有用性が期待できる．

© 2010 Information Processing Society of Japan

表 5　ドメインの逆引きの調査結果
Table 5　Result on survey of reverse-lookup of domain.

	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	8個　（20%）	2個　（11%）
逆引き結果が正しくない	29個　（72%）	4個　（21%）
返答なし	3個　（8%）	13個　（68%）
合計	40個（100%）	19個（100%）

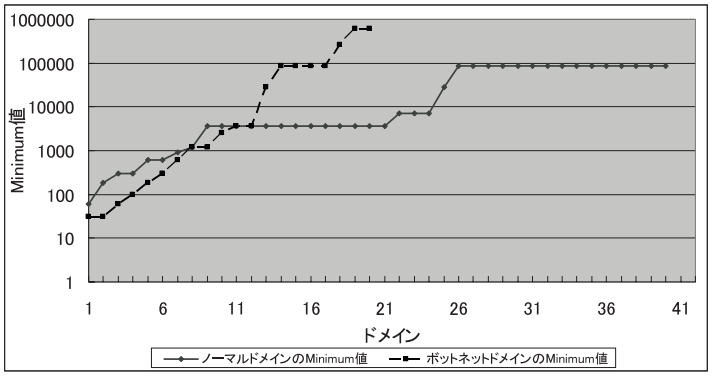

図 2　ドメインの Minimum 値の調査結果
Fig. 2　Result of survey on minimum value of domain.

4.2　SOA レコード

SOA レコードには以下の情報が含まれる．

(1)　ゾーン名
(2)　ネームサーバホスト名
(3)　管理者メールアドレス
(4)　シリアル番号
(5)　Refresh 値
(6)　Retry 値
(7)　Expire 値
(8)　Minimum 値

今回著者らは，SOA レコードのうち，(8) Minimum 値を調査した（**図 2**）．Minimum 値とは，ネガティブキャッシュとも呼ばれる値である．クライアントが DNS サーバにドメイン情報の問合せを行い，DNS サーバから対象のドメイン名が存在しないという返答の際，その情報を保持する期間が Minimum 値である．今回著者らは，ボットネットの登録ホスト更新頻度が，ノーマルドメインより高いため，(8) の値に特徴が見られると考えて調査を行った．

調査の結果，ノーマルドメインの Minimum 値は，3600 および 86400 という値に集中している傾向が分かった．この結果より，Minimum 値はボットネットドメインを特徴付けるデータとしての有用性が期待できる．

4.3　WHOIS

WHOIS からは，以下の情報を得ることができる．

(1)　登録ドメイン名
(2)　レジストラ名
(3)　ドメインが登録されている DNS サーバ名
(4)　ドメインの登録開始年月日
(5)　ドメインの登録有効期限
(6)　ドメイン名登録者の名前・住所
(7)　技術的な連絡の担当者連絡先
(8)　登録に関する連絡の担当者連絡先
(9)　登録者への連絡窓口の連絡先

今回著者らは，上記のうち，(4) と (5) の差分である (10) ドメインの登録期間に関して調査を行った（**図 3**）．調査の結果，(10) からは，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分かり，ボットネットドメインを特徴付けることが可能と思われる結果を得られた．

4.4　Mail，Web サーバの有無

同一ドメイン上における，Mail サーバと Web サーバの，ドメイン名登録の有無を調査した（**表 6**）．

調査の結果，ノーマルドメインとボットネットドメインともに，Mail サーバ，Web サーバが登録されている割合が多いということが分かった．しかし，ボットネットドメインを特徴付けるような，2 種類のドメインデータにおける差異は，今回の調査では見つけることはできなった．

© 2010 Information Processing Society of Japan

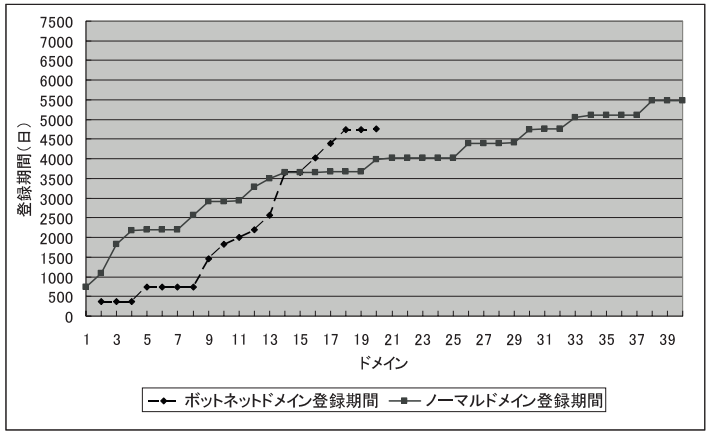

図 3　ドメインの登録期間の調査結果
Fig. 3　Investigation result on registration period of domain.

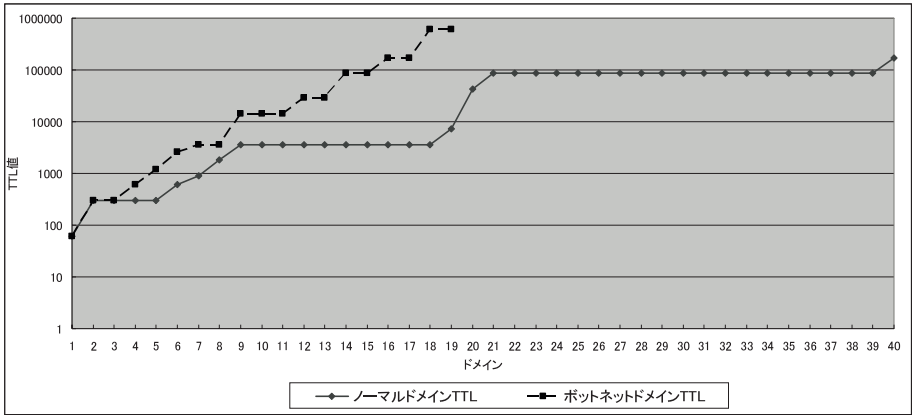

図 4　ドメインの TTL 値の調査結果
Fig. 4　Result of the survey on TTL value of domain.

表 6　Web，Mail サーバの登録の有無の調査結果
Table 6　Result on survey of registration existence of Web/Mail server.

	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	35個　（87%）	11個　（57%）
ウェブサーバ無し	2個　（5%）	3個　（16%）
返答なし	3個　（8%）	5個　（27%）
合計	40個（100%）	19個（100%）

	ノーマルドメイン	ボットネットドメイン
メールサーバ有	28個　（70%）	12個　（58%）
メールサーバ無し	8個　（20%）	2個　（15%）
返答なし	4個　（10%）	5個　（27%）
合計	40個（100%）	19個（100%）

4.5　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 4**）．調査の結果，ノーマルドメインの TTL 値は，3600 および 86400 という値に集中している傾向が分かった．この結果より，TTL 値は，ボットネットドメインを特徴付けるデータとしての有用性が期待できる．

5.　実験による最適パラメータの設定と評価

5.1　赤池情報量基準

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ための指標である[14]．AIC を用いることで，モデルの複雑さと，データとの適合度のバランスをとることが可能となる．

たとえば，ある測定データを統計的に説明する数式を作成する．数式は，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．だがその反面，無関係なノイズなどの影響が大きくなり，信頼性が低下する．この問題は，パラメータの数を抑えることで解決ができる．しかし，実際にどの数まで抑えればよいのかの判断は困難である．このような問題に対し，AIC は式 (6) によって求められる最小 AIC 時のパラメータ数を選択すれば，多くの場合，良いモデルを選択することが可能となる．

$$AIC = -2\ln L + 2k \tag{6}$$

ここで，L は最大尤度，k は自由パラメータである．今回の例においては，k は各要素数に相当し，L は，各パラメータ数での，数量化理論 2 類を用いて導き出した判別結果と，判別結果の正答との，乖離の最小二乗和に相当する．

5.2　実 験 方 法

フィッティング段階として，5.4.1 項のパラメータ数設定実験により数量化理論 2 類に用いる，パラメータの数を設定する．次に，5.4.2 項のパラメータ設定実験により，求めたパラメータ数での組合せのうち，最も検知率の高い組合せを求める．

そして，判別および予測の段階として，5.4.3 項の検証実験により，求めた組合せの検知率の検証を行う．

© 2010 Information Processing Society of Japan

表 7　パラメータ設定値
Table 7　Parameter setting value.

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

(a)

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

(b)

Mailサーバ	設定値
あり	1
なし	2

(c)

ウェブサーバ	設定値
あり	1
なし	2

(d)

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

(e)

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

(f)

最後に，パラメータ設定実験と検証実験の結果を比較し，最も検知率の高いパラメータの組合せを，3.2.2 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合せとする．

5.3　実験データ

5.3.1　実験に用いるパラメータの設定値

4 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 類のパラメータ設定値を設定した（**表 7**）．ただし，パラメータの設定にあたり，調査した 5 項目のうち，Web サーバと Mail サーバの調査結果を個別にパラメータとしたため，パラメータの総数は 6 つとなっている．

5.3.2　実験に用いるドメインのデータ

パラメータ設定実験と検証実験の，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．2 つの実験には，均等な数のドメインを使用することが望ましい．よって，各ドメインから任意に選択し，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 個用いる．同様に，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 20 個を用いることとする．

5.4　実験と結果

5.4.1　パラメータ数設定実験の結果

AIC を用いた，最適なパラメータ数の比較結果を**表 8** に示す．この結果より，最適なパラメータ数は，AIC 値が最も小さい，パラメータ数 3 であることが分かった．よって，以後のパラメータ設定実験，検証実験は，パラメータ数を 3 で行う．

表 8　AIC を用いた最適パラメータ数比較結果
Table 8　Compared result for number of best parameters using AIC.

パラメータ数	最大対数尤度	AIC値
5	100.78067	−191.56135
4	615.04532	−1222.09065
3	**728.83046**	**−1451.66091**
2	520.02917	−1036.05834

表 9　パラメータ設定実験の結果
Table 9　Result of parameter setting experiment.

パラメータの組み合わせ						検知率
TTL値	mailサーバの有無	ドメインの登録期間				83.3
TTL値	mailサーバの有無		webサーバの有無			56.7
TTL値		ドメインの登録期間	webサーバの有無			83.3
	mailサーバの有無	ドメインの登録期間	webサーバの有無			80
TTL値	mailサーバの有無			逆引き		76.7
TTL値			webサーバの有無	逆引き		76.7
		ドメインの登録期間	webサーバの有無	逆引き		80
TTL値		ドメインの登録期間		逆引き		83.3
	mailサーバの有無	ドメインの登録期間		逆引き		83.3
	mailサーバの有無		webサーバの有無	逆引き		70
TTL値	mailサーバの有無				Minimum値	0
TTL値		ドメインの登録期間			Minimum値	83.3
	mailサーバの有無	ドメインの登録期間			Minimum値	83.3
TTL値			webサーバの有無		Minimum値	0
	mailサーバの有無		webサーバの有無		Minimum値	26.7
		ドメインの登録期間	webサーバの有無		Minimum値	80
TTL値				逆引き	Minimum値	0
	mailサーバの有無			逆引き	Minimum値	76.7
			webサーバの有無	逆引き	Minimum値	80
		ドメインの登録期間		逆引き	Minimum値	83.3

5.4.2　パラメータ設定実験の結果

パラメータ数が 3 における，全数量化理論 2 類への試行結果を**表 9** に示す．

この結果より，最も検知率の高い値は 83.3%であり，この値を示したパラメータの組合せは 7 組存在した．よって，次項の検証実験で，7 組すべての検証を行い，パラメータ設定実験結果と検証実験結果それぞれの検知率を比較する．そして，最も検知率の高いパラメータの組合せを求める．

5.4.3　検証実験の結果

検証実験の結果，**表 10** に示すように，5.4.2 項で求めた，3 つのパラメータの組合せ 7 組の中で，最も検知率の高いものは「逆引き」「Minimum 値」「ドメインの登録期間」の，3 つ

© 2010 Information Processing Society of Japan

表 10　検証実験の結果
Table 10　Result of verification experiment.

パラメータの組み合わせ			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
TTL値	mailサーバの有無	ドメインの登録期間	83.30%	16.70%	79.31%	31.03%
TTL値	webサーバの有無	ドメインの登録期間	83.30%	16.70%	79.31%	31.03%
TTL値	逆引き	ドメインの登録期間	83.30%	16.70%	86.20%	13.79%
逆引き	mailサーバの有無	ドメインの登録期間	83.30%	16.70%	68.96%	24.13%
TTL値	Minimum値	ドメインの登録期間	83.30%	16.70%	75.86%	24.13%
Minimum値	mailサーバの有無	ドメインの登録期間	83.30%	16.70%	86.20%	13.79%
逆引き	Minimum値	ドメインの登録期間	83.30%	16.70%	96.55%	3.44%

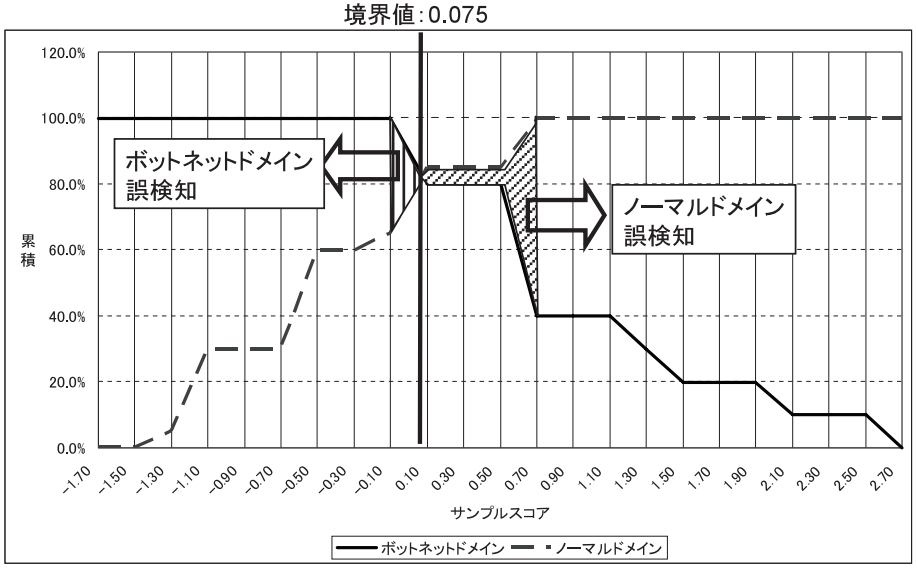

図 5　最適 3 パラメータの検知数の累積グラフ
Fig. 5　Accumulated distribution graph using three best parameters.

のパラメータの組合せ（以下，最適 3 パラメータとする）であることが分かった．

また，この最適 3 パラメータにおける検知率は 96.5%であり，False Negative と False Positive の値の合計は 3.5%であった．

表 11　最適 3 パラメータのカテゴリスコア
Table 11　Category score using three best parameters.

項目名	カテゴリー名	分類数	カテゴリースコア
Mininum	1-100	3	0.4363
	101-1000	2	-0.1565
	1001-100000	23	-0.0839
	NA	2	0.4664
ドメイン登録期間	1-2500	8	0.4559
	2501-5000	10	-0.1808
	5001-8000	9	-0.8256
	NA	3	1.8638
逆引き	リストなし	13	0.3064
	応答が正しくない	15	-0.3749
	正常な応答	2	0.8201

5.4.4　最適 3 パラメータにおける各種データ

最適 3 パラメータの，パラメータ設定実験における検知数の累積グラフを**図 5** に示す．

2 つのグラフが交わる点が境界値を表し，その値は 0.075 である．また，ボットネットドメインのグラフでは，サンプルスコアが境界値よりも高い領域は検知に成功したドメインを表し，ノーマルドメインのグラフでは，サンプルスコアが境界値よりも低い領域が検知に成功したドメインを表す．

対して，ボットネットドメインにおける，境界値よりも低い領域は False Negative を，ノーマルドメインにおける境界値よりも高い領域は False Positive を表し，合わせた値が 16.7%となる（図 5 中斜線部）．また，この値を 1 から引いた値が，先に述べた検知率 83.3%に相当する．

最適 3 パラメータのカテゴリスコアを**表 11** に示す．表 11 を使用し，各ドメインのカテゴリスコアを加算することにより，サンプルスコアを求めることができる．

たとえば，未知のドメイン α を予測する場合を考える．まずドメイン α に対し，各要素ごとに該当するカテゴリを調査し，表 11 からカテゴリスコアを求める．

そして，各カテゴリスコアを加算し，ドメイン α のサンプルスコアを求める．このとき，求めたサンプルスコアの値が境界値より高ければボットネットドメイン，低ければノーマルドメインと予測される．

5.5　数量化理論 2 類とブラックリスト方式を組み合わせて用いた検知率の評価

提案する 2 つの検知方式を組み合わせて用いた，検知率の評価を行った（**表 12**）．この結果より，提案する 2 つの検知方式が検知に失敗したボットネットドメインに，重複はない．

© 2010 Information Processing Society of Japan

表 12　提案する 2 つの検知方式を組み合わせた場合の検知結果
Table 12　Detected result using proposed method.

	ブラックリスト方式	数量化理論2類を用いた方式	
		パラメータ設定実験	検証実験
ボットネットドメイン1	×	○	
ボットネットドメイン2	○	○	
ボットネットドメイン3	○	○	
ボットネットドメイン4	○	○	
ボットネットドメイン5	○	○	
ボットネットドメイン6	×	○	
ボットネットドメイン7	○	○	
ボットネットドメイン8	○	○	
ボットネットドメイン9	○	○	
ボットネットドメイン10	○	×	
ボットネットドメイン11	○		○
ボットネットドメイン12	×		○
ボットネットドメイン13	○		○
ボットネットドメイン14	○		○
ボットネットドメイン15	○		×
ボットネットドメイン16	○		○
ボットネットドメイン17	○		○
ボットネットドメイン18	○		○
ボットネットドメイン19	×		○

つまり，提案する 2 つの検知方式を組み合わせて，どちらかが怪しいとしたものを怪しいドメインであるとすることにより，見落としがないことが分かる．よって，第 2 段トレースバックシステムに，提案する 2 つの検知方式を組み合わせた方式を実装することで，実環境でも同様の，高い検知率が期待できる．

6. 提案するシステム

6.1 多段追跡システム概要

著者らが構想している多段追跡システムは，**図 6** のように 3 つの段階で構成される．ネットワーク β にあるボット PC が，ネットワーク γ 内の PC に対して攻撃を行っていると仮定し，以下に各段の追跡について説明する．

(1) 第 1 段：第 1 段トレースバックは，IP トレースバックシステムを用いて，ボット PC が存在するネットワーク（以下ネットワーク β）まで追跡する．

(2) 第 2 段：第 2 段トレースバックでは，ネットワーク β から第 2 段追跡対象が存在するネットワーク（以下ネットワーク α）までを追跡する．

(3) 第 3 段：第 3 段トレースバックは，ネットワーク α からハーダの操作する PC までを追跡する．

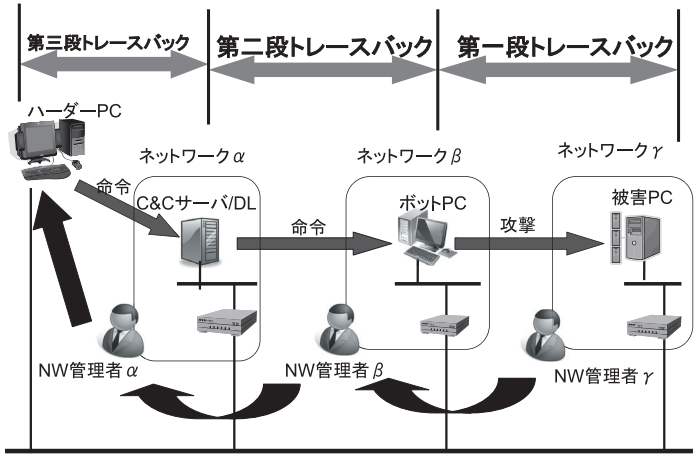

図 6　多段追跡システム概要
Fig. 6　The overview of multistep pursuit system.

6.2 提案する第 2 段トレースバックシステム概要

6.2.1 第 2 段トレースバックシステムの概要

提案する 2 つの検知方式を組み合わせて用いる，第 2 段トレースバックシステムのシステム構成を**図 7** の下部に示す．

第 1 段トレースバックにより，第 2 段トレースバックシステムの NW 管理者 β が，NW 管理者 γ から攻撃の通達（図 7 (a)）を受けた場合を想定する．その際 NW 管理者 β は，後述する B-DATA から，該当する自 NW 内攻撃ホストと通信を行った第 2 段追跡対象と疑わしきホスト（以下，対象ホストとする）を検索し，後述する Bot-GL と Bot-BL と照合することで，以下の処理を行う．

(1) 対象ホストが Bot-GL に登録されている場合

NW 管理者 β は，対象ホストの所属する NW 管理者 α に確認を依頼する（図 7 (b)）．確認の結果，第 2 段トレースバックシステムの誤検知ならば，Bot-GL より対象ホストを削除する．対して，確認により対象ホストが第 2 段追跡対象の可能性があるならば，NW 管理者 α に対処を依頼するとともに，対象ホストを Bot-GL から Bot-BL に移動させる．

(2) 対象ホストが Bot-BL に登録されている場合

NW 管理者 β は，対象ホストが所属する NW 管理者 α に強い警告（図 7 (b)）をし，対

© 2010 Information Processing Society of Japan

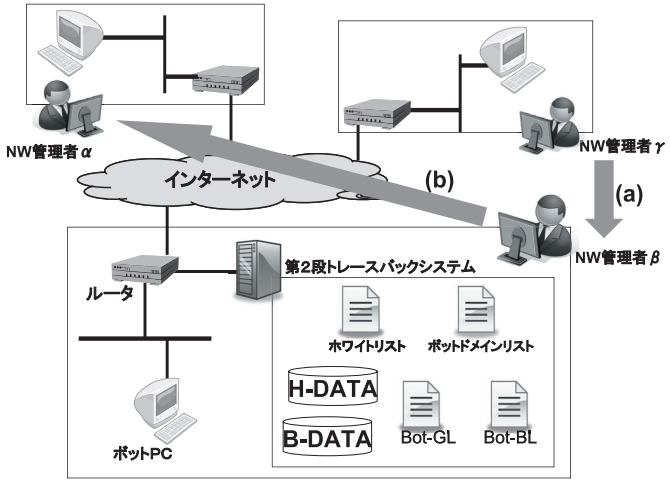

図 7　第 2 段での追跡のためのシステム構成
Fig. 7　System structure of the second step trace back.

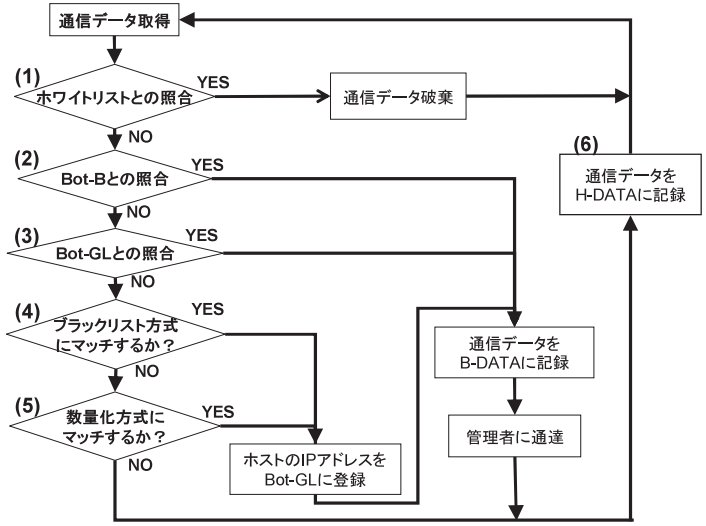

図 8　第 2 段での追跡のための処理の流れ図
Fig. 8　Flowchart of processing for second step trace back.

象ホストの対処を促す．警告の結果，NW 管理者 α より対象ホストの対処が完了したと通達を受ければ，対象ホストを Bot-BL から削除する．

ここで，第 2 段トレースバックシステムを構成する各要素に関して，以下に述べる．

・**Bot-GL**
第 2 段追跡対象と疑わしきホストの IP アドレスリスト

・**Bot-BL**
第 2 段追跡対象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第 2 段追跡対象ドメインのリスト

・**H-DATA**
取得したすべての通信データを一定時間記録する媒体

・**B-DATA**
Bot-GL，Bot-BL に登録された IP アドレスの通信データを記録する媒体

また，第 2 段トレースバックシステムは，第 1 段からの通達により，第 3 段への連絡を行うが，この処理とは別に，常時ルータを通過する通信データを取得し，以下の内容で監視を行う（**図 8**）．

ただし，第 2 段トレースバックシステムは，ポートミラーリング機能などによって，ルータを通過するすべての通信データのコピーを取得し，監視を行うものとする．

(1)　送信元 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，リストになければ (2) に移行する．

(2)　送信元 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．リストになければ (3) に移行する．

(3)　送信元 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，リストになければ (4) に移行する．

(4)　DNS クエリ内のドメイン名をブラックリスト方式で検知処理する．検知すればその IP アドレスを Bot-GL に登録し，NW 管理者 β に通達するとともに，B-DATA に通信を記録する．検知しなければ (5) に移行する．

(5)　DNS クエリ内のドメイン名を数量化理論 2 類を用いた検知方式で検知処理する．検

© 2010 Information Processing Society of Japan

知すればその IP アドレスを Bot-GL に登録し，NW 管理者 β に通達するとともに，B-DATA に通信を記録する．検知しなければ (6) に移行する．

(6)　すべての通信を一定時間 H-DATA に記録する．

なお，ここでは，第 2 段トレースバックシステムは，NW 管理者 β にはただちに異常を報告するが，第 3 段の NW 管理者 α には連絡をせず，第 1 段の NW 管理者 γ から連絡が来て，はじめて対応するとして説明をした．しかしこれは，異常を検知したらただちに管理者 α に連絡をし，対応を依頼することも可能である．

6.2.2　第 2 段トレースバックシステムの課題

ボットネットは更新の頻度が高いため，ブラックリスト方式でインターネット上から取得したドメインのリストが永続的に存在することはなく，更新が必要になる．同様に，数量化理論 2 類を用いた検知方式においても，今回の調査結果が長く適応可能であるという保障はない．そこで，検知方式それぞれで使用するデータの更新を考慮する必要がある．ブラックリスト方式におけるデータの更新は，インターネット上から定期的に行うことが可能である．数量化理論 2 類を用いた方式では，解析データ元に CCCDATASet2009 を用いており，新たに CCCDATASet を入手できた時点でデータの更新を行うことで対応をすることとする．

7. おわりに

本論文では，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を提案するとともに，第 2 段追跡対象の特定方式として，ブラックリスト方式と数量化理論 2 類を組み合わせて用いる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 2 段トレースバックシステムの実装を目指すとともに，ボットネットの特徴が変更された際に，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（コンピュータセキュリティ），Vol.47, No.8, pp.2512–2523 (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（コンピュータセキュリティ），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* (2008).
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* (2009).
6) Goebel, J. and Rishi, T.H.: Identify bot contaminated hosts by irc nickname evaluation, *1st Workshop on Hot Topics in Understanding Botnets* (*HotBots'07*), Cambridgs (2007).
7) Gu, G., Porras, P., Yegneswaran, V., Frog, M. and Lee, W.: BotHunter: Detecting malware infection through ids-driven dialog correlation, *Proc. 16th UNISEX Security Symposium* (*Security'07*), Boston (2007).
8) 朝長秀誠，田中英彦：Botnet の命令サーバドメインネームを用いた Bot 感染検知方法に関する研究，情報処理学会研究報告（コンピュータセキュリティ），Vol.35, No.3, pp.13–18 (2006).
9) Schonewille, A. and van Helmond, D.: The domain name service as an ids. How dns can be used for detecting and monitoring badware in a network (2006). http://staff.science.uva.nl/~delaat/snb-2005-2006/p12/report.pdf
10) SRI Malware Threat Center. http://www.mtc.sri.com
11) DNS-BH Malware Domain Blacklist.
http://www.malwaredomains.com/wordpress/
12) 林知己夫：数量化—理論と方法，朝倉書店 (1993).
13) WHOIS サービス．http://whois.jprs.jp
14) 赤池弘次，甘利俊一，北川源四郎，樺島祥介，下平英寿：赤池情報量基準 AIC，共立出版 (2007).

(平成 21 年 11 月 30 日受付)
(平成 22 年 6 月 3 日採録)

三原　元

平成 16 年 4 月東京電機大学工学部第一部情報メディア学科入学．卒業研究でネットワークセキュリティに関する研究を実施．平成 20 年 3 月同大学卒業．同年 4 月東京電機大学大学院入学．平成 22 年 3 月同大学院修了．同年 4 月株式会社 NTT PC コミュニケーションズ入社．

© 2010 Information Processing Society of Japan

佐々木良一（フェロー）

昭和 46 年 3 月東京大学卒業．同年 4 月日立製作所入社．システム開発研究所でシステム高信頼化技術，セキュリティ技術，ネットワーク管理システム等の研究開発に従事．平成 13 年 4 月より東京電機大学工学部教授，平成 19 年 4 月より未来科学部教授．工学博士（東京大学）．平成 10 年電気学会著作賞，平成 14 年情報処理学会論文賞，平成 19 年総務大臣表彰，平成 20 年情報処理学会功績賞等を受賞．著書に，『IT リスクの考え方』岩波新書，2008 年等．情報処理学会コンピュータセキュリティ研究会顧問．日本セキュリティ・マネージメント学会会長，情報ネットワーク法学会理事長，日本ネットワークセキュリティ協会会長．

© 2010 Information Processing Society of Japan