CCC からの累積データを用いたボットネットの C&C サーバ検知システムの開発と評価

中村 暢宏 †　　佐々木 良一 †

†東京電機大学
〒120-8551 東京都足立区千住旭町 5
{nakamura_n, 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，3年分の累積したデータを併用する検証を行い，検出漏れへの一定の効果を示すと同時に，パラメータ候補の項目を追加する事で検出精度の改善を行い，システム有効性の継続検証をした．本論文では，先に検証した第2段追跡方式をSquid上に実装し，検出精度の検証をすると共に処理性能の評価を検証し報告を行う．

Development and Evaluation of the System to Detect C&C Server in Botnet Using Accumulated Data from CCC

Nobuhiro Nakamura†　Ryoichi Sasaki†

†Tokyo Denki University
†5, Senju-Asahi-Cho, Adachi-ku,
Tokyo, 120-8551 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 on a proxy server program named Squid. We also evaluation second stage trace back method which consists of black list and Quantification methods No. 2 with CCCDataSet2009, 2010, 2011 and new parameters. This paper reports at first the system implemented for the 2nd Stage Trace Back, and then the results of functional experiment and performance experiment.

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，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，たとえボット PC を発見して対策を施しても，対策が不十分であれば，新たなボットに感染したり，他の PC が容易に感染したりするなどの恐れがある為，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，先に累積したデータを用いて継続検証を行ってきた検知方式[4]について, 検出機能を持つシステムをフリーのプロキシサーバ「Squid」[5]上に試作開発し, 機能ならびに性能実験を行った結果を報告する.

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 章で時間経過による検出精度の変動について述べ，4 章で試作したシステムの構成について述べ，5 章でシステムの性能評価について結果を報告する.

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 MWS2012 実行委員会より得られた，CCC DATAset 2012 [6]に含まれる,攻撃通信データ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.

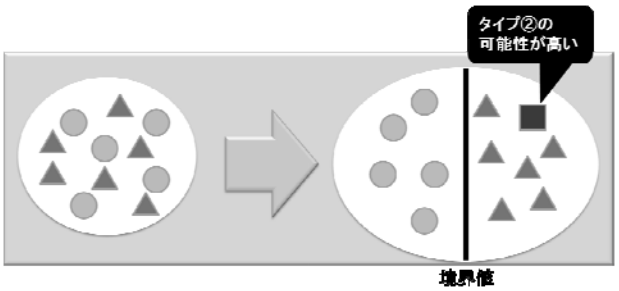

図1. 数量化理論 2 類概要

3 検出精度変動の検証

本章では，過去の検証で求めた数量化理論 2 類検知方式の検出条件に関して，2011 年度の時点で収集されていた各ドメインについて，2012 年 7 月末時点の時点における検出率の変化を報告する.

3.1 検証に用いるデータ

先の研究[4]で，前 2 章で述べた様に，数量化理論２類を用いた検知方式の検知率に関する検証を行うに当たり，適用するパラメータ候補が必要である．その為，CCCDATAset2011 に含まれる，攻撃通信データの解析を行い，ドメイン情報を収集

した[6].

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．ただし，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．これらのドメインは以降，攻撃通信データの解析を通して取得した第二段追跡対象と思われるドメインを"ボットネットドメイン"と, ボットネットとは関係しないドメインを"ノーマルドメイン"とする．

取得したドメインに含まれる各種ドメイン情報の内，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，以下の項目を用いる．

1. **逆引き**
2. **TTL値**
3. **SOAレコード**
4. **WHOIS情報**
5. **Aレコードの個数**
6. **MXレコードの有無**
7. **NSレコードの有無**
8. **CNAMEレコードの有無**

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, 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. 5 は, 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．6 は，ドメイン宛てのメールと IP アドレスの関係付けを行う MX レコードの有無の調査．7 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 レコードの有無の調査．8 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 レコードの有無の調査である．

各項目の調査手法に関しては, 4 は WHOIS サービスを用いて調査を，残りの項目に関しては，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．これら取得したドメイン情報を基に，数量化理論 2 類による検知を行う．パラメータ設定には，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．

これらのドメイン情報の解析と，パラメータ設定を通して得たカテゴリースコアと境界値を用いて，次項検証を行う．

3.2 比較検証

本項では，先の検証[4]において，数量化理論を用いた検知方式における実験を通して設定したカテゴリースコアと境界値を用いて，2012 年 7 月末時点の時点における検知率を，過去の検知率と比較し，時間経過による影響を検証する．この検証で用いるデータとして，CCCDATAset2009～2011 の解析データから 81 個のボットネットドメインを， alexa．com[7]より日本国内アクセスランキング上位 50 サイトをノーマルドメインとして用いた．また検知率は，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．

表 1 に先の研究[4]で求めた検知率と 2012 年 7 月末時点で行った検証における検知率の結果を示す．項目は，それぞれ①は 2009 年度から 2011 年度に取得した 3 年分のドメインを併用し，先行研究[3]において用いられたパラメータ候補である, 前項で示したパラメータ候補 1～6 を用いて検証した設定値, ②は 2011 年度に取得したドメインを用いて，①におけるパラメータ候補に新たに前項で示したパラメータ候補 7, 8 を追加して検証した設定値，③は②におけるノーマルドメインを当研究室のネットワークから取得したドメインから, alexa．com の日本国内アクセスランキング上位 50 サイトへ変更して検証した設定値である．

なお，これら設定値は，時間経過が与える影響を調べるため，現在のデータに再設定した物では無く,先の研究[4]で設定した値を用いた．

表 1 検知率比較

	①	②	③
2011年度	76.20%	77.20%	97.60%
2012年度	31.30%	58.78%	87.79%

表の結果からも解るように，3 パターン全て検出率が低下していることが解る．①は，2012 年度分のデータを加えて無いため，大幅な検出率の低下が起きたと考えられる．②においても①と比べ

小さくはあるが減少している．①と比べ減少が抑えられた理由として，新たに採用したパラメータ候補が 2012 年度時点においても特徴としての有効性を持って為では無いかと考えられる．一方で，③は検出率の低下は在るが，一定の値を保っている．これは，検証時にノーマルドメインとしてアクセスランキング上位に来る大手サイトと比較している為，中小規模で在ることが多いボットネットドメインのホストとの差が大きく，検出を容易にしていると考えられる．ただし，上記の理由により，個人運営のサイトにおける誤検知が多く見られるなど，定期的な更新が必要と考えられる．

このように，時間経過による検知率の低下は，昨年検証した 3 種の組み合わせ全てに見られた．しかし，特徴変動への対応など，一定の対策を行った場合，その低下を防ぐことに成った．これらから，システムにおいて，時間経過による特徴変動への対応は不可欠で在ると考えられる．

4　システムの実装

4.1　システムの構成

先行研究[3]において，あるネットワーク上における数量化理論検知方式とブラックリスト検知方式を併用して用いた，第二段トレースバックシステムが提案されている．今回は，フィルタリング部に注目し，実装を行った．下記の図 2 にシステム図を示す．

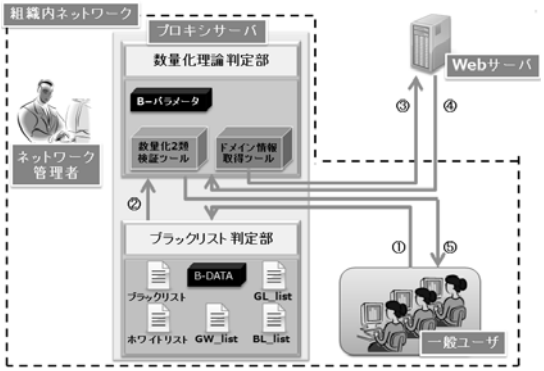

図 2　提案システム構成図

3 章で述べたように，時間経過による特徴変動への対応は不可欠で在る．そこで，通信データの累積を用いて情報の更新を行い対応する形へ，提案されたシステムの改良を行った．

システムを構成する要素に関しては、以下のとおりである．

- ブラックリスト
  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
- ホワイトリスト
  ボットネットに関係の無いホストのリスト
- BL_list
  ブラックリストで検出したホストのリスト
- GL_list
  第 2 段追跡対象と疑わしいホストのリスト
- GW_list
  数量化判定を通過したホストのリスト
- B-DATA
  ブラックリスト,BL_list,GL_list に登録されたホストとの通信データを記録するログ
- B-パラメータ
  数量化判定処理に用いる，カテゴリースコアと境界値

下記の図 3 に，これらの要素を用いたリスト判定のアルゴリズムを示す．

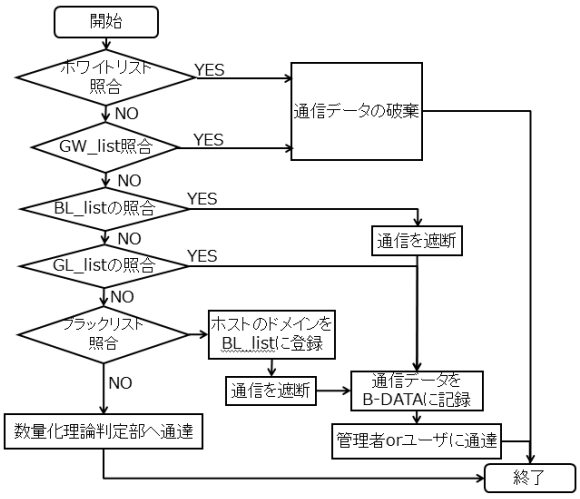

図 3　リスト判定部アルゴリズム

(1) 要求された URL が，ホワイトリストに記録されていた場合, 通信データを破棄し, 判定を終了.
(2) 要求された URL が，BL_list に記載されていた場合，通信を遮断し，通知を行う．
(3) 要求された URL が，GL_list に記載されてい

た場合，通信の遮断はせず，通知を行う．
(4) 要求された URL が，ブラックリストに記録されていた場合，BL_list に登録し，通信の遮断を行い，通知を行う．
(5) いずれのリストにも検出されなければ，数量化理論判定部へ移行する．
また, (2)~(4)においては通知前に B-DATA へ通信ログを記録する．
図 4 に，数量化判定アルゴリズムを示す

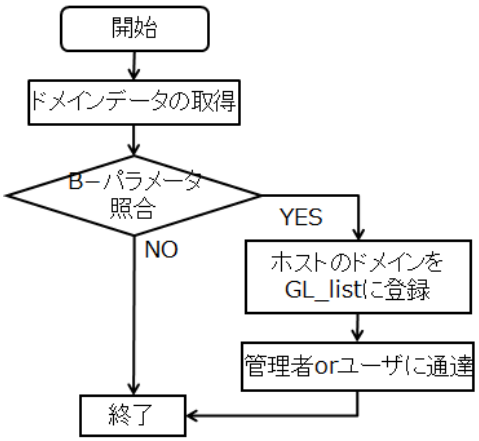

図 4 数量化判定部アルゴリズム

(1) 要求された URL のドメインについて，3 章で述べた 8 項目のドメイン情報を取得する．
(2) 各項目について，事前に求めたカテゴリースコアと境界値である B-パラメータと照合し，検出された場合，概要ドメインを GL_list に登録し，通知を行う．
また，検出判定を行った通信データを記録・蓄積し，定期的に数量化理論 2 類パラメータ設定処理を行う際のデータとする．順次設定値を最適化していく．これにより特徴変動への対応を行っていく．
次項にて，プロキシと各判定部の実装の説明を行う．

4.2 実装

プロキシにはフリーで公開されているプロキシサーバ Squid[5]を使用した．また，Squid には redirect_program というオプションがあり，プログラムを指定することで URL の判定・書き換え処理を行う事が出来る．この機能を用いて，図 3，図 4 の判定処理の実装を行った．ただし，現時点ではフィルタリング処理のみ実装しており，数量化理論 2 類を用いた設定値の自動更新処理の機能は未実装である．実装に使用した言語は Perl 言語を用いており，約 640 ステップと成っている．

5 性能評価

実装した，数量化理論検知方式とブラックリスト検知方式を併用して用いた，第二段トレースバックシステムの処理性能を調べるために，システムを適応した場合と，適応しない場合の 2 通りで Web ページが表示されるまでの時間を計測し検証する．検証に用いる Web ページとして Yahoo! JAPAN と当大学のホームページを用い，計 20 回アクセスし，その平均を比較する．Web ブラウザは Firefox14.0.1 を使用し, 計測には Firebug を使用した．事前条件として，Firefox は，履歴を無効，キャッシュを 0MB に設定し，Squid のキャッシュ機能は無効とした．また，実装したシステムは，一度計測して検出されなかった場合リストに登録され，以降の検証でリスト判定の時点で検出する仕組みになっている為，リストのリセットする．回線速度は, BNR スピードテスト[8]を用いて計測し，下りが約 36.17Mbps 上りが約 12.17Mbps と成っている．計測の結果を表 2 に示す．

表 2表 3 接続速度 計測結果

	適応しない	適応する
Yahoo! JAPAN	1.04	3.16
当大学HP	3.33	8.55

システムを適応しない場合，Yahoo! JAPAN では 1.04 秒，当大学の HP では 3.33 秒となり，システムを適応した場合，Yahoo! JAPAN では 3.16 秒，当大学の HP では 8.55 秒となった．この結果から，処理時間に，約 2〜5 秒かかる事が分かる．処理時間に幅が在る理由として, HP の要素数に関係が在ると考えられる．これは，要求のあった URL に対して，システムが逐次対応をするため，要素数が多いほど処理数が増えるためである．

次に, 処理に対する要素数の影響を調べるため, alexa. com の日本国内アクセスランキング上位 24 のサイトに対して計測を行った．この計測では, 要素数の指標としてページサイズを求め比較する．計測結果を，図 5 に示す.

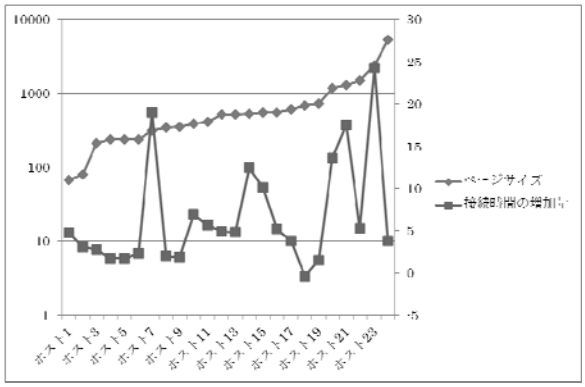

図 5 ページサイズと増加時間

図からも分かるように，ページサイズの増加に伴い，接続時間の差も増加している．一部飛び出している部分については，ページサイズに対して要素数が多いポータルサイトであった.

これらの計測結果から，第二段トレースバックシステムの処理時間は，大型サイトなど一部利用に影響を与えるが，運用は可能なレベルの負荷だと考えられる.

6 おわりに

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について，時間経過による検出率の低下を確認し，特徴変動への対応の必要性を再確認した．併せて，フリーのプロキシサーバ「Squid」上にシステムの試作開発を行い，性能評価を行った．検証により大型サイトなどページ内の要素数により負荷の変動は在るが，実際の運用が可能な範囲の負荷であった.

今後は，実際に運用を行い，問題点の発見を行うと同時に，通信負荷の軽減を図る．また，自動的に数量化理論 2 類と累積した通信データを用いて，カテゴリスコアと境界値の更新を行う仕組みを作り，より特徴変化への対応を行う，動的に対応可能なシステムの運用を検討する.

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, 佐々木良一, 「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」, マルウェア対策研究人材育成ワークショップ 2011 (MWS2011), 2A4-1, 2011 年 10 月.

[5] squid : Optimising Web Delivery
http://www.squid-cache.org/

[6] MWS2012 実行委員会, 研究用データセット MWS 2012 Datasets について,
http://www.iwsec.org/mws/2012/about.html#datasets

[7] Alexa the Web Information Company
http://www.alexa.com/

[8] BNR スピードテスト
http://www.musen-lan.com/speed/