4Z-6

ボットネットの C&C サーバ特定手法の経年変化データを用いた評価

岡安翔太† 佐々木良一†
東京電機大学†

1 はじめに

近年ボットネットによる被害が増加しており問題となっている．ボットネットとは悪意を持った攻撃者の命令に基づき動作するプログラムに感染した PC(以下，ボット PC)及び攻撃者の命令を送信する指令サーバ(以下，C&C サーバ)からなるネットワークであり，中には数万規模の PC などからなるボットネットもあると言われている[1]．攻撃者が C&C サーバに命令を送ることで，ボットネットに接続されたボット PC はフィッシング目的などの SPAM メールの大量送信や，特定サイトへの DDos(Distributed Denial of Service)などに利用され，非常に大きな脅威となりうる[2]．これらのボット PC を用いた攻撃の，攻撃元の特定方法として IP トレースバックなどの手法を用いることで，攻撃元を偽装した場合でも検出可能である．しかし，対策が不十分であれば PC が容易に感染するおそれがある為，根本的な解決とはならない．

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，攻撃者の特定を目的とする，多段追跡システムを構成した[3]．本論文は，このうち第二段において C&C サーバ・ダウンローダ(以下，第二追跡対象)を検知する方式に関するものであり，過去数年に渡るデータの経年変化の調査と，先に提案された検知方式[3]が最新のデータに対して有効であるか再評価を行い，結果を報告するものである．

2 第二段トレースバックシステム

第二追跡対象の特定には，数量化理論２類を用いた検知方式を行う．数量化理論２類を用いて分類を行う為に，マルウェア対策研究人材育成ワークショップより得られたボットネットに関するデータセット，CCCDatasets2011 と PRACTICE Datasets2013 のデータを使用する．

2.1 数量化理論を用いた検知方式

数量化理論は，元統計数理学研究所所長の林知己夫教授らにより開発されたデータ分析手法である[4]．この内，数量化理論２類ではダミー変数の導入による質的データの数量化を行うことで，判別分析に相当する処理を可能にする．

例えば，分析対象データの集合に 1 群と 2 群が混在するとき数量化理論２類を用いて 1 群と 2 群に判別するケースを考える．初めに，1 群と 2 群，それぞれを特徴付けるデータをパラメータとして設定した判別式と 1 群と 2 群に判別するための基準となる境界値の 2 種を設定する．この判別式に集合の各要素を入力し出力される判別値を，先に求めた境界値と比較する．この際に，判別値が境界値より高い値ならば 1 群，低い値ならば 2 群と判別を行う．また，使用した集合外の要素についても同様に境界値を比較することで，1 群・2 群どちらに属する可能性が高いかを推定することが可能である．

2.2 データ解析手法

既存方法に倣い，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，ボットネットに関連のあるドメイン(以下，B ドメイン)と，比較対象としてボットネットに関係の無いドメイン(以下，N ドメイン)の 2 種類のデータを用いる．B ドメインは 2011 年度のデータである CCCDatasets2011 と，2013 年度のデータである PRACTICE Datasets2013 に含まれる攻撃通信データから取得できる C&C サーバの DNS(Domain Name System)のドメイン情報を用いる．DNS のドメイン情報より，「逆引き」，「TTL」，「A レコード」，「minimum 値」，「MX レコード」，「登録期間」，「NS レコード」，「CNAME レコード」の 8 つのパラメータ候補の調査を行う．なお，調査対象は各ボット PC が接続する第二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ものではない．N ドメインに対しても，DNS のドメイン情報より上記 8 つのパラメータ候補の調査を行う．次章で判別における評価を行う．

3 実験による検証と評価

２章でのパラメータを用い，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[5]を使用した．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パラメータ設定実験では二種類の組み合わせと，経年変化の与える影響について検証を行う．

3.1 実験における検知率

本実験における B ドメイン検知判定の組み合わせを表１に示す．正しく B ドメインと判別されることを True Positive．正しく N ドメインと判別されることを True Negative．この 2 つを合わせたものが検知率である．また，B ドメインが N ドメインと判別されることを False Negative．N ドメインが B ドメインと判別されることを False Positive．この２つを合わせたものが誤検知率である．ボットネットであると疑わしいドメインを検知するため False Negative の値が低いことが望ましい．

表１． 検知判定組み合わせ

	検知結果が真	検知結果が偽
B ドメイン	True Positive (TP)	False Negative (FN)
N ドメイン	False Positive (FP)	True Negative (TN)

検知率 = (TP + TN) / ドメイン総数
FNR = FN / ドメイン総数
FPR = FP / ドメイン総数

Evaluation of Method for Detecting C&C Server of Botnet using Annual Change Data
SHOTA OKAYASU† RYOICHI SASAKI†
Tokyo Denki University†

3.2 2011 年度実験のパラメータ設定による検証

2011 年度データによる実験に使用したパラメータ値，及びパラメータの組み合わせを用いて 2013 年度のデータを用いて検証を行った．2011 年度データでは最適パラメータの数は 4 個で，最適な組み合わせは複数存在する．特に「逆引き」，「登録期間」がボットネットと特徴づけるパラメータとしてあがった．実験に使用する組み合わせは「逆引き」，「登録期間」，「MX レコード」「NS レコード」の４つで行う．表２に実験結果を示す．

表２． 2013 年度データ検証実験

	検知率	FNR	FPR
2011 年度データ	97.62%	0.00%	2.38%
2013 年度データ	80.00%	17.50%	2.50%

2013 年度では 2011 年度に比べ検知率が下がった．そこで 2013 年度の B ドメインの特徴を調査したところ「逆引き」，「A レコード」，「登録期間」にボットネットの特徴が見受けられる．さらに 2011 年度と 2013 年度の B ドメインの特徴を比較してみたところ，引き続き「逆引き」，「登録期間」は B ドメインを特徴付けるデータとしての有効性があると見受けられた．「A レコード」は 2011 年度から 2013 年度にかけて特徴付けのデータとしての有効性が上がった．しかし，「minimum 値」に関しては B ドメインと N ドメインの数値が近づく傾向があり，特徴付けるデータとしての有効性は下がった．これらの調査結果より B ドメインの特徴が時間経過とともに変動することで，従来用いていたパラメータの組み合わせと値では十分に対応できなくなってしまったと言える．

3.3 組み合わせデータによる検証

3.2 節と比較するため，パラメータの設定実験を 2013 年度と 2011 年度のデータで行う．2013 年度のみのデータでは取得ドメイン数が少なかったため結果を得ることができなかった．故に組み合わせデータでの実験を行う．

実験に使用するデータは 2013 年度 B ドメインが 18 個，2011 年度 B ドメインが 42 個の計 60 個と N ドメイン 66 個で行う．パラメータ設定実験と検証実験ではドメイン数が均等な数が望ましい．そのため，パラメータ設定実験に 2013 年度 B ドメインが 9 個，2011 年度 B ドメイン 21 個の計 30 個と N ドメイン 33 個で行う．同様に，検証実験では，残りの半数である B ドメイン，計 21 個と N ドメイン 33 個で行う．パラメータ設定実験の結果，最適パラメータ数は 4 個であり，最も検知率が高いパラメータの組み合わせは複数発見された．最適組み合わせパターンを表３に示す．

表３． 組み合わせデータの最適パラメータ

	パラメータ			
①	逆引き	A レコード	登録期間	CNAME レコード
②	逆引き	A レコード	登録期間	NS レコード
③	逆引き	A レコード	登録期間	MX レコード
④	TTL	A レコード	登録期間	minimum 値

解析結果，「逆引き」，「A レコード」，「登録期間」をボットネットと特徴付けるパラメータとしてあがった．表３での組み合わせ検証実験の検知率を表４に示す．表４の①～④は表３の①～④に対応している．

表４． 組み合わせデータの検知率

	パラメータ設定実験			検証実験		
	検知率	FNR	FPR	検知率	FNR	FPR
①	88.71%	6.45%	4.84%	90.32%	4.84%	4.84%
②	88.71%	6.45%	4.84%	90.32%	4.84%	4.84%
③	88.71%	6.45%	4.84%	90.32%	4.84%	4.84%
④	85.48%	6.45%	8.07%	93.55%	1.61%	4.84%

3.4 組み合わせデータの設定による検証

3.3 節でのパラメータの組み合わせにおける 2011 年度，2013 年度データでの有効性を検証した．検証結果を表５に示す．表５の①～④は表３での組み合わせに対応している．3.2 節と比較した結果，②の「逆引き」，「A レコード」，「登録期間」，「NS レコード」での組み合わせにおいて，2011 年度の検知率を維持したまま，2013 年度の検知率を向上することができた．新たなデータを追加することで，最適パラメータの組み合わせを検証できたと言える．

表５． 最適パラメータ検証実験

	2011 年度データ			2013 年度データ		
	検知率	FNR	FPR	検知率	FNR	FPR
①	85.72%	7.14%	7.14%	87.50%	5.00%	7.50%
②	97.62%	0.0%	2.38%	87.50%	5.00%	7.50%
③	95.24%	2.38%	2.38%	87.50%	5.00%	7.50%
④	76.19%	16.67 %	7.14%	82.50%	15.00 %	2.50%

４ 終わりに

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，2011 年度から 2013 年度にかけて検知率は若干低下した．今後も時間経過により B ドメインの特徴が変動すると考えられる．そのため，現在用いているパラメータの組み合わせと値では，今後も検知率の低下が懸念される．

従って，一定期間ごとに最適なパラメータの組み合わせと値に更新する必要がある．パラメータの決定には数量化理論を用いる方法以外に，機械学習による検証方法がある．今後，機械学習での検証を試みるとともに，学習データのパラメータを自動更新かつ，検証データを基に B ドメインをフィルタリングできるシステムを検討する．

５ 参考文献

[1]警視庁 情報セキュリティ広場
http://www.keishicho.metro.tokyo.jp/haiteku/haiteku/haiteku409.htm
[2]サイバークリーンセンター
https://www.ccc.go.jp/bot/
[3] 三原元，佐々木良一，”数量化理論と CCCDATAsets2009 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”，情報処理学会論文誌 VOL.51, No.9，pp1579-1590
[4]林知己夫,”数量化−理論と方法”, 朝倉書店(1993)
[5]株式会社エスミ http://www.esumi.co.jp/