2012年度 修士論文

DNS情報による
悪意のあるサイトの検出法

提出日：2012年8月2日

指導：後藤滋樹教授

早稲田大学大学院 基幹理工学研究科 情報理工学専攻
学籍番号：5110B140-4

劉　亦晨

目次

図一覧

表一覧

第 1 章

序論

1.1 研究の背景

インターネット上の悪意のあるサイトによる被害が拡大・深刻化している。悪意のあるサイトの中には、サイトを訪れるだけてウイルスや悪質なプログラムを強制的にダウンロードさせて、ユーザの個人情報の不正入手を狙うサイトが存在している。またユーザーのパソコンのシステムファイルを削除し、変更を加えて不具合を引き起こすサイトもある。いずれもブラウザや OS のセキュリティ上の弱点を利用するものである。ユーザが悪意のある意図に気付くことがなく、防御が難しい。さらに、ボットネット (Botnet) によるマルウェアの感染やフィッシング詐欺の被害が深刻になっている。

2010 年上半期だけで新種のマルウェアが 1 億 2400 万件も出現したという報告 [1] がある。また、2011 年上半期だけでフィッシング詐欺サイトが 195,901 件も発見されたという報告もある [2]。

このような問題に対して、従来からファイアウォールや IDS (侵入検知システム) などの技術が研究開発されている。ただし、既知の悪意のあるサイトや通信に対して有効な対策法は、今後さらに複雑になり多様化する新種の悪意のあるサイトや通信に対応できない恐れがある。既知の悪意のあるサイトや通信だけでなく、新しいタイプの悪意のあるサイトや通信に対しても有効なリアルタイムで軽量な検出手段が必要とされている。

1.2　研究の目的

すでに 1.1 節で述べたように、多様化する悪意のあるサイトに対してリアルタイムで検出できる方法を確立することは大きな意味を持つ。本研究ではドメイン名に注目して、リアルタイムで悪意のあるサイトを検出することを目的とする。具体的にはドメインの特徴パターンを用い、高精度な検出を実現する。

1.3　本論文の構成

本論文は以下の章により構成される。

第 1 章 序論

本研究の概要について述べる。

第 2 章 DNS: Domain Name System

DNS: Domain Name System について解説する。

第 3 章 BlackList の例

BlackList の具体的な例を紹介する。

第 4 章 提案手法

提案手法を説明する。

第 5 章 実証実験

実験により本提案を検証する。

第 6 章 結論

本論文の結論を述べるとともに、残された課題を示す。

第 2 章

DNS: Domain Name System

2.1 ドメイン名空間

DNS (Domain Name System) は、インターネットにおけるホスト名と IP アドレスの対応関係を効率よく管理するための分散データベースである。単なる数字から構成されている IP アドレスは、人間にとって使い易いとはいえない。IP アドレスの代わりにドメイン名を使うことができる。

ドメイン名とは、人間にとって覚えやすい英文字、数字、ハイフン (-) から構成される「名前」のことである。DNS のドメイン名空間は、逆木の形をしている。木構造の一番上の根をルートと呼び、他の節点をノードと呼ぶ。各レベルのドメイン名の空間もまたドメインと呼ばれる。図 2.1 参照。

ドメイン名における制限

各ノードには、最大 63 文字のドットを含まない単純なラベルが付けられる。ラベルの英文字は大文字と小文字が区別されない。ルートドメインはヌルラベルを持つ特殊なノードである。木構造の深さは最大 127 レベルに制限されている。

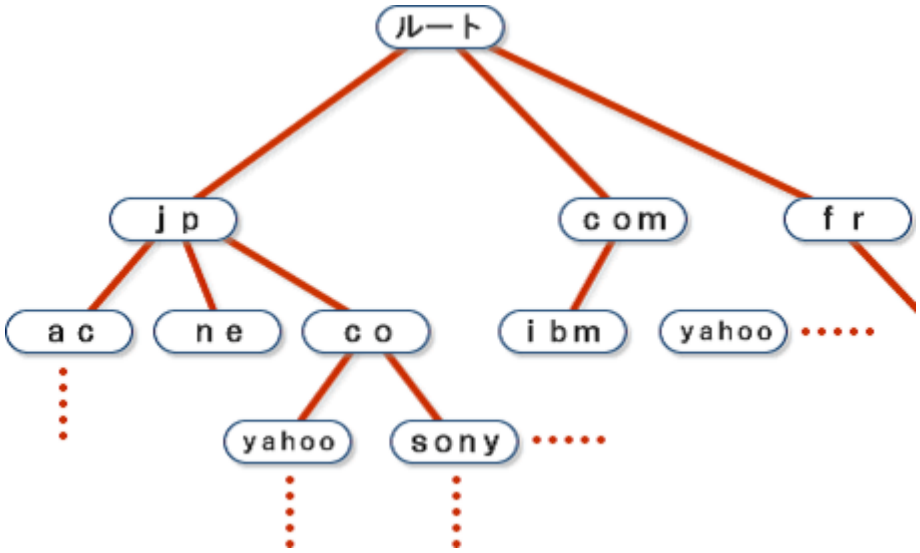

図 2.1: DNS の木構造

2.2　委任

インターネットの初期の時代には、ホストと IP アドレスとの対応関係を一つのテキストファイルで管理していた。その後に、インターネットが急激に発展してホストの数が膨大となったために、当初のように単一のファイルに全部のホストを収容することが難しくなった。DNS では、管理の分散化を計るために委任 (delegation) という考え方を採用している。

ドメインを管理する組織は、ドメインを管理しやすいように、ドメインを複数のサブドメインに分割する。その分割したサブドメインの管理を他の組織に任せることができる。このような方法を委任と言う。委任された組織はそのサブドメインの管理の全責任を持つ。委任された組織はサブドメインのデータを維持管理するだけではなく、サブドメインをさらに複数のサブドメインに分割し、それぞれの管理を他の組織や人間に委任することもできる。

2.3　ネームサーバとゾーン

ドメイン名空間に関する情報を管理するプログラムとマシンはネームサーバと呼ばれる。また、2.2 節で述べたように分割されて委任されたドメインの単位をゾーンと呼ばれる。ネームサーバには、担当するゾーンについての完全な情報が格納されている。

ゾーンは、他のサーバに権威が委任されているサブドメインを除き、同じドメイン名を持つホストのすべてのドメイン名を含む。ネームサーバはゾーンの情報のみを扱い、権威が他のサーバに委任されている情報については管理しない。

2.4　ネームサーバの種類

DNS のネームサーバにはプライマリ DNS とセカンダリ DNS の 2 種類が定義されている。プライマリ DNS は、自ホストの設定ファイルからゾーンのデータを読み込む。セカンダリ DNS は、ゾーンの権威を持っているプライマリ DNS からゾーンのデータを取得する。セカンダリ DNS は起動時にプライマリ DNS に問い合わせて、必要に応じてゾーンデータを取得する。

2.5　リゾルバと名前解決

リゾルバとは、ネームサーバに問い合わせて、ホスト名に基づいて IP アドレスの検索を依頼したり、その逆方向の検索を依頼したりするクライアントである。IP アドレスやホスト名を必要とするインターネットの応用プログラムは、通常はリゾルバを介して名前解決を行う。リゾルバは以下の仕事を行う。

- ネームサーバへの問い合わせ
- 応答の解釈
- 仕事を要求したプログラムへの情報の返送

リゾルバはその処理のほとんどをネームサーバに頼っている。ゾーンからデータを検索する処理のことを、上記のように名前解決と呼ぶ。

2.6　再帰と反復

ネームサーバへの問い合わせの方法には、次の二つの種類がある。リゾルバからネームサーバへ問い合わせを行う際に使われる方式を再帰問い合わせという。ネームサーバは、リゾルバから受け取った問い合わせに対する回答を自分の管理するデータの範囲で行うことができなければ、他のネームサーバに問い合わせを行い最終的に解決しなければならない。これに対して

反復問い合わせを受けたネームサーバは、自らが持つ情報の範囲で名前解決に適したネームサーバを次の参照先として返す。

図 2.2 に再帰問い合わせを図解する。

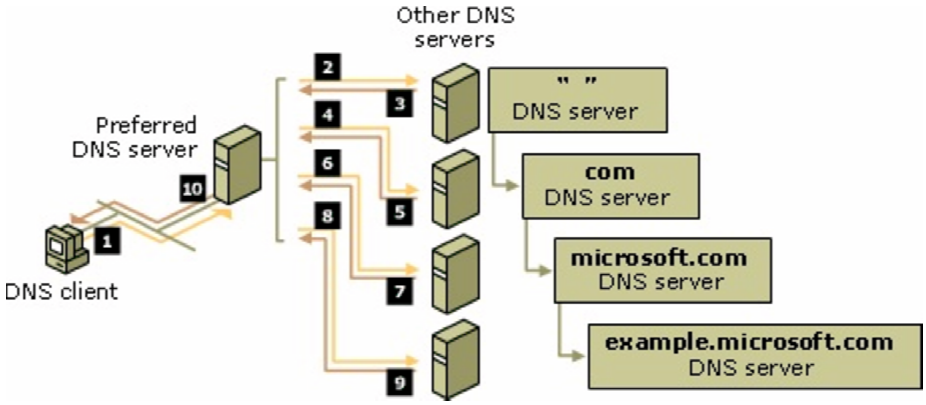

図 2.2: DNS の再帰問い合わせ

2.7　キャッシュ

ネームサーバは再帰的な問い合わせに対する回答を得るまでに、多数の問い合わせを他のネームサーバに送る。この過程で入手した情報を、後に関連した問い合わせ受ける場合に備えて、データとしてキャッシュする。DNS のキャッシュは名前解決を高速化し、ネームサーバへの問い合わせの回数を減らすことで負荷を抑える機能を持つ。

2.8　生存時間 (TTL: Time To Live)

DNS キャッシュには生存時間が設定されている。これはゾーンの管理者がゾーンのデータに対して設定するものである。生存時間を超えると、該当する DNS キャッシュのデータを保持しているネームサーバは DNS キャッシュを破棄する。古いデータを破棄することによって、データベースの整合性を維持している。キャッシュの生存時間をどのくらいの長さにするかは、問い合わせの処理効率とデータの一貫性とのバランスによる。生存時間を短くすると、キャッシュがすぐに無効になり、最新のデータを頻繁に問い合わせることになる。ネットワーク上に分散したデータの一貫性が高くなる一方で、上位のネームサーバに対する負荷が上昇する。ま

た、生存時間が短いと DNS キャッシュを更新する頻度が多くなるので、ネームサーバに不正な名前解決情報をキャッシュさせる DNS キャッシュポイゾニング攻撃 [4] の危険性が高まる。

DNS キャッシュポイゾニング攻撃に対する有効な対策として、DNSSEC (DNS Security Extentsins Securing the Domain Name System) [5] があるが、まだ完全に普及していない状況である [6]。また、DNSSEC の導入によりデータ量が増加するために DNS サーバへの負荷が高まることが指摘されている [7]。

2.9　DNS サーバの負荷分散

2.9.1　DNS ラウンドロビン

DNS ラウンドロビン (DNS round Robin) とは、一つのドメイン名に複数の IP アドレスを割り当てる負荷分散技術の一種である。トラフィックの負荷を複数の IP アドレスに振り分けることにより、一つサーバに対するアクセスをほぼ同量ずつ複数のサーバマシンに分配することができる。これは BIND[8] 等の DNS サーバのゾーン設定により容易に実現できる。

DNS ラウンドロビンでは TTL を短めに設定した運用が一般的であるため、2.8 節で述べたように、DNS キャッシュポイゾニング攻撃される危険性が高い。図 2.3 に DNS ラウンドロビンを図解する。

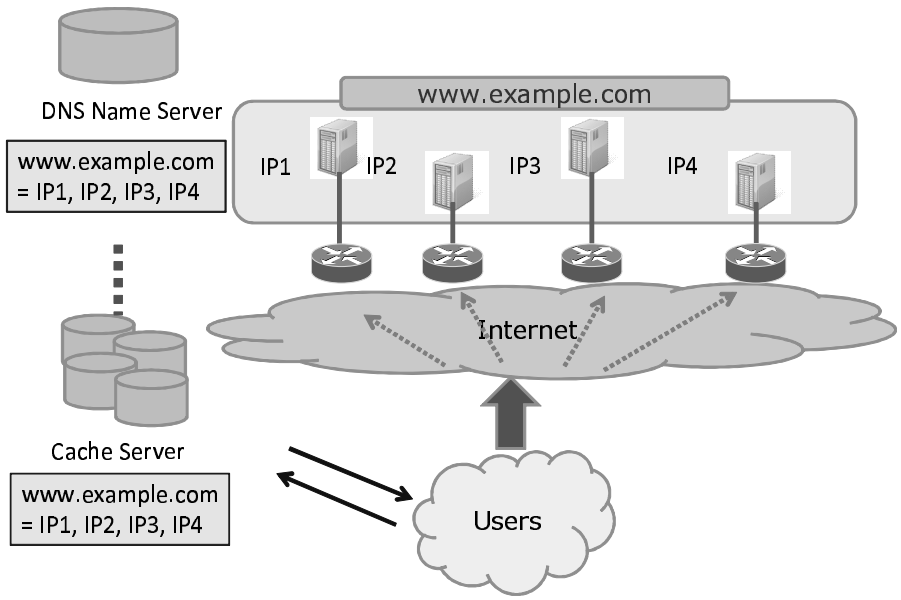

図 2.3: DNS ラウンドロビン

2.9.2　DNS の IP Anycast

インターネットでは、通常は一つの IP アドレスを送信先として指定する。この場合には一つの機器とだけ接続する。このような一対一の通信をユニキャスト (Unicast) と呼ぶ。インターネットにおいてもっとも一般的な通信方式である。

一つの IP アドレスをインターネット上の複数の機器に割り当てて、アドレスを共有する技術を IP Anycast (エニーキャスト) という。IP Anycast では同じ IP アドレスを複数の機器に重複して割り当てる。機器をノード (Node) と呼ぶ。各ノードには、個別に割り当てた IP アドレスとは別に、IP Anycast で使用する IP アドレスを追加で割り当てる。送信元が送り出したパケットは、anycast のアドレスを持つノードのうちのどれか一つに到達して、処理される。なお同じアドレスを持つすべてのノードは、受信したデータに対して同じサービスを提供する。

IP Anycast 技術を導入すると、DNS サーバの負荷を分散することができる。また、DDoS 攻撃 (Distributed Denial of Service attack) のような攻撃に対する耐性を強化することができる。

第 3 章

BlackListの例

BlackList (ブラックリスト) とは悪意のある活動を行うホストのドメイン名やIP アドレスを一覧にしたリストのことである。具体的に公開されているBlackList が複数存在する。Blacklist はプロバイダや公共機関によって広く用いられているが、情報が必ずしも正確とは限らず、個々のBlackList によっても正確性に差異があると言われている [9]。

3.1 RBL.JPプロジェクト

RBL はReal-time Blackhole List の略である。日本独自のブラックリストデータベースシステムで、スパムメールを送信、転送するやフィッシング詐欺と思われるホストのドメインを収集し、各種のRBL を立ち上げている [10]。

3.2 SORBS

SORBS はSpam and Open Relay Blocking System の略である。スパムメールを送信する、または転送すると思われるホストのドメインを収集し、公開している。S ORBS は2002 年11 月に設立、2003 年1 月までには、私的リストとして運営されていた。今は12,000,000 超の悪意のあると思われるホストドメインが登録されている [11]。

3.3　The Spamhaus Project

Spamhaus は 1998 年に設立された非営利団体であり、スパムメールの送信などに関連のあるドメインや IP の BlackList を管理している。 スイスのジュネーブに本部があり、38 の調査チームと解析チームが世界の 10 カ所で運営を行っている。Spamhaus は複数の DNS BlackList を管理している。その中で主要な Domain Black List として DBL がある。本論文では DBL を用いて提案手法を検証する。

The Domain Block List (DBL)

フィッシング詐欺、スパムメールを送信するホスト、またはマルウェアやウイルスなどに感染したと思われるホストのドメインを収集したものである。専門チームによってリストが最新に保たれている。

3.4　SpamCop

SpamCop はフィルタリング兼スパム報告支援システムである。ユーザはこのシステムを利用することで、受信したスパムに関係するネット業者に対し、スパム受信報告を簡易に行うことができる。このような自動処置依頼システムをユーザに提供する一方で、そこで集めた情報により作られたフィルタリングシステムを有料サービスとして提供している [13]。

第 4 章

提案手法

本論文は DNS 情報を用いて悪意のあるサイトを検出する手法を提案する。

4.1 提案手法の概要

提案手法は DNS 情報を用いて悪意のあるサイトを検出する。従来の悪意のあるサイトの検出法には、ファイアウォールや IDS (Intrusion Detection System) で IP アドレスや URL の情報に基づいて検出する手法がある。ただし IP アドレスや URL 情報に基づいた検出法は、処理するための負担が大きい。さらに IP アドレスや URL 情報は DNS 情報を問い合わせた後に分析することになるため、DNS 情報を使う方がリアルタイム性の面で優位性がある。このため、より高速に悪意のあるサイトを検出するには、DNS 情報を用いて悪意のあるサイトを検出する手法に利点がある。

4.2 提案手法の詳細

悪意のあるサイトの DNS 情報には次のような三つの特徴がある。この特徴のそれぞれを本論文で検証する。最終的に、提案手法は三つの特徴を用いて総合的に分析できることを実証する。

- 英数字が混在するドメイン
  悪意のあるサイトのドメインは英文字と数字が混在するものが多い [14]。
- 10 文字以上で構成されるドメイン
  悪意のあるサイトのドメインは Fast-Flux[15] などの技術を用いて自動生成されることが

多いため、人間にとって覚えにくくても 10 文字以上の長い文字列で構成されるものが多い。

- TTL 値が 300 以下のドメイン
  有名なサイトでも TTL 値を小さく設定することがあり得るが、悪意のあるホストでは TTL 値を 300 以下に設定する傾向がある [16]。これはドメイン名を捕捉されにくくするためである。

以下の本論文では三つの特徴を用いて、総合的に分析する手法を検証する。

第 5 章

実証実験

本章では提案手法の実証実験を行う。

5.1 評価に用いるデータ

本研究で用いたデータは、早稲田大学の対外接続回線において、早稲田大学内に入る方向の DNS のデータを、tcpdump を用いてポート番号を 53 番に指定して、パケットとして収集したものである。2012 年 6 月 3 日の 00:00:16 ～ 2012 年 6 月 5 日 00:00:16 の三日間に観測された 91,527,142 個のパケットを利用する。データの解析には、著者が作成したシェルスクリプトと本研究のために用意した正規表現を用いる。

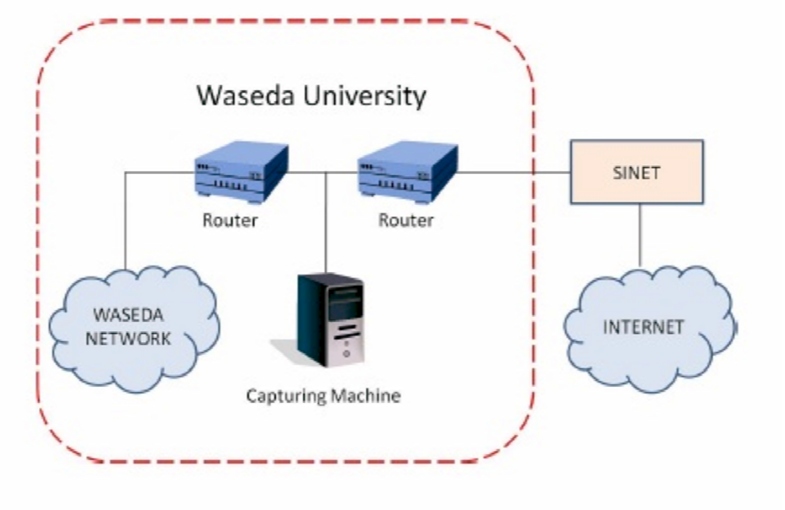

図 5.1: ネットワークの構成図

5.2　実験

5.2.1　特徴パケットの抽出

最初に、収集したデータから DNS クエリを抽出する。DNS クエリと DNS パケットの統計を表 5.1、図 5.2 に示す。

表 5.1: DNS クエリと DNS パケットの統計

DNS クエリ	DNS パケット全体	割合
32,551,179	91,527,142	35.6%

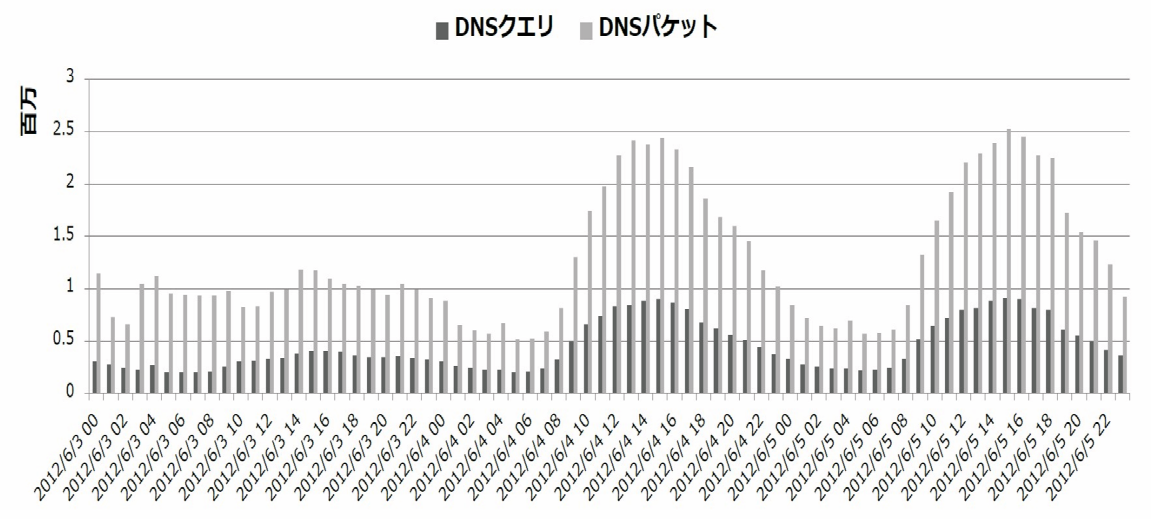

図 5.2: DNS クエリと DNS パケットの統計

DNS クエリは DNS パケットに占める割合が 35.6% である。

次に抽出した DNS クエリと、前述の三つの特徴パターンをそれぞれ照合して、一致したドメインを抽出する。

- 英数字が混在するドメイン
- 10 文字以上で構成されるドメイン
- TTL 値が 300 以下のドメイン

結果として抽出したドメインと DNS クエリの統計を図 5.3、図 5.4、図 5.5 に示す。

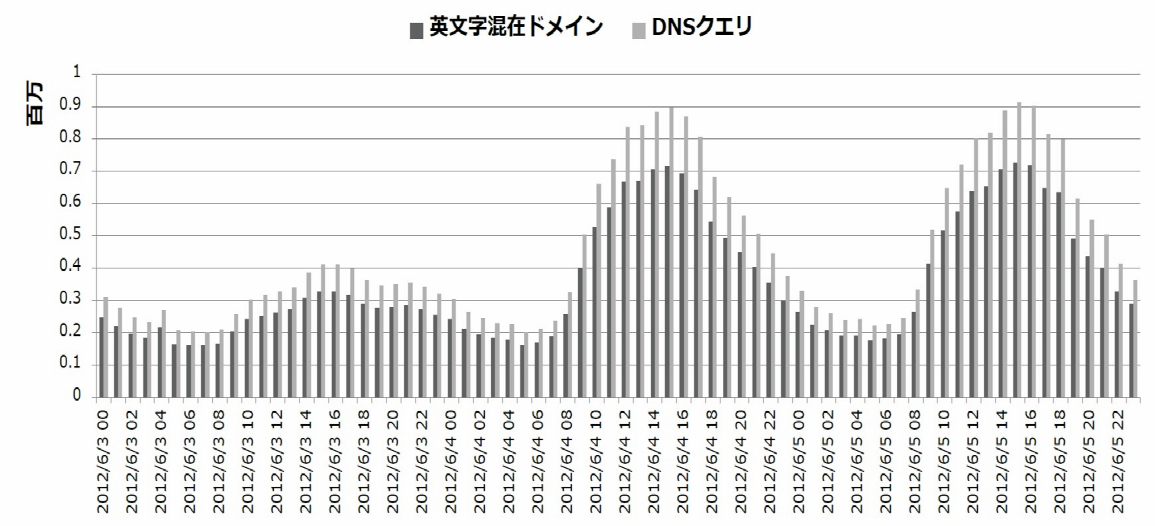

図 5.3: 英数字が混在するドメインと DNS クエリの統計

英数字が混在するドメインは、DNS クエリの約 79.7% を占めている。

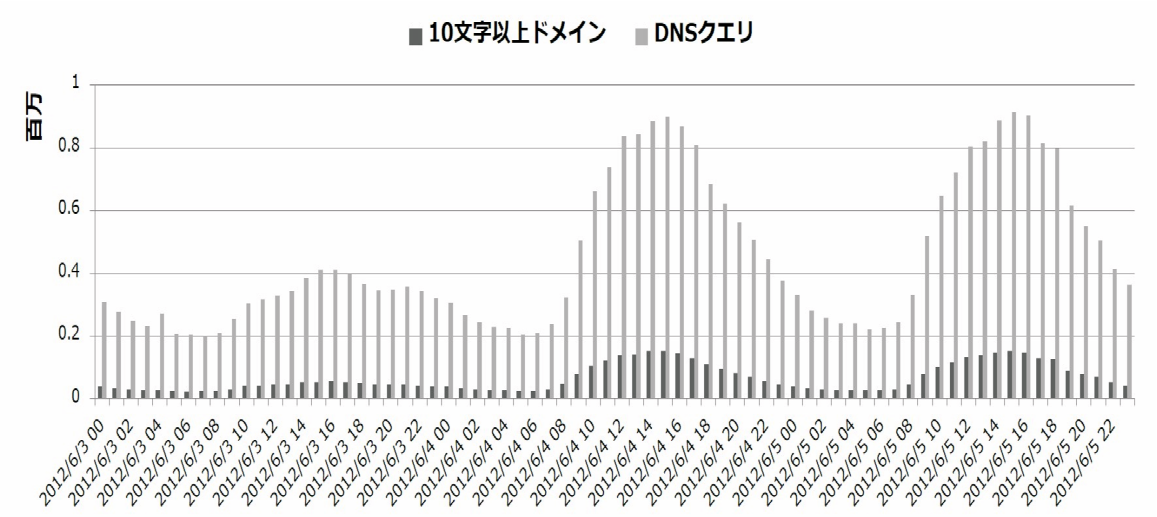

図 5.4: 10 文字以上で構成されるドメインと DNS クエリの統計

10 文字以上で構成されるドメインは、DNS クエリの約 13.9% を占めている。

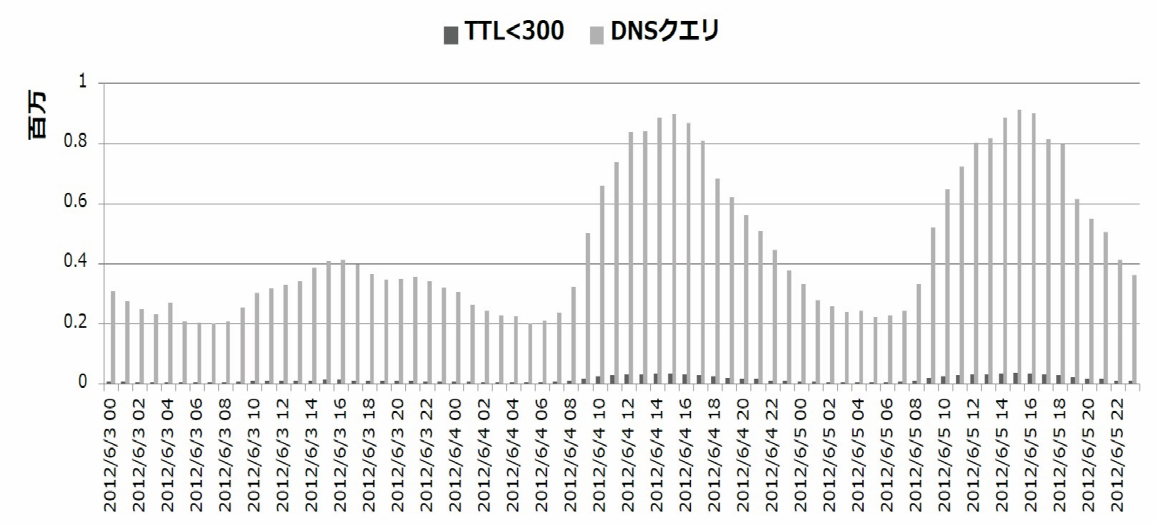

図 5.5: TTL 値が 300 以下のドメインと DNS クエリの統計

TTL 値が 300 以下のドメインは、DNS クエリの約 3.3% を占めている。

全体的に、英数字が混在するドメインが多いことが分かる。また、10 文字以上で構成されるドメインと TTL 値が 300 以下のドメインが DNS クエリに占める割合が低いことも分かる。

最終的に DNS クエリに三つの特徴パターンを全部用いて、適合したドメインを抽出する。抽出したドメインと DNS クエリの統計を図 5.6 に示す。

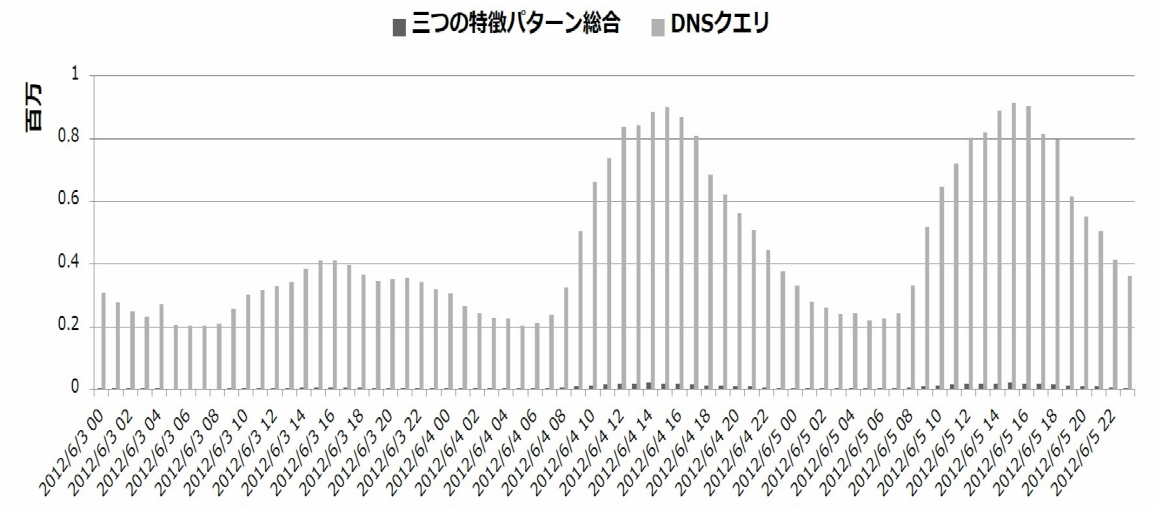

図 5.6: 総合して抽出したドメインと DNS クエリの統計

総合して抽出したドメインは、DNS クエリの約 1.9% を占めている。総合して抽出したドメ

インは、DNS クエリに占める割合が最も低い。

抽出したドメインの数を表 5.2 にまとめた。ただし、重複ドメインを除いてユニークなドメイン名を数えている。

表 5.2: ドメイン数のまとめ

特徴パターン	ドメイン数
英数字が混在するドメイン	21,288,471
10 文字以上で構成されるドメイン	1,144,394
TTL 値が 300 以下のドメイン	926,959
三つの特徴パターンの総合	595,108

5.2.2　実験結果と考察

これまでに得られた結果は、第 4.2 節で述べた三つの特徴パターンを個別にあるいは総合的に DNS クエリと照合して、抽出したものである。ここで得られたドメイン名を BlackList と比較して適合率 (Precision) を求めた。本論文では、Spamhaus の DBL を BlackList として使用する。BlackList 適合率の計算は、抽出したドメイン名の中の BlackList に登録されているドメイン名の数を分子とした。また抽出したドメイン名の中で、著名なドメインを集めたリストである Alexa Top Global Sites[17] に登録されているドメイン名の数を分母に加えた。その分数の値に 100 をかけて算出した。以下の式 5.1 のように表せる。計算の結果をまとめて表 5.3 に示す。

$$\text{適合率}\,(Precision) = \frac{\text{抽出ドメイン}\,(DBL)}{\text{抽出ドメイン}\,(DBL) + \text{抽出ドメイン}\,(Alexa)} \tag{5.1}$$

表 5.3: 特徴パターンの適合率

特徴パターン	適合率
英数字が混在するドメイン	20.4%
10 文字以上で構成されるドメイン	50.9%
TTL 値が 300 以下のドメイン	53.7%
三つの特徴パターンの総合	93.9%

この結果をみると、本論文の提案手法は DNS 情報だけを比較対象として分析しているが、高い確率で悪意のあるサイトを検出することが可能であることが分かる。

第 6 章

結論

6.1　まとめ

本研究は、DNS 情報を用いて悪意のあるサイトの検出を行った。悪意のあるサイトのドメイン名には三つの特徴パターンが観測されるという現象に着目した。そこから「三つの特徴パターンを満たすドメイン名は悪意のあるサイトである」という仮説を立て、実証実験を行った。その結果、三つの特徴パターンを総合的に適用して、抽出したドメイン名と BlackList との適合率が高いということがわかった。これによって悪意のあるサイトの検出法の一つとして利用することができる。

6.2　今後の課題

本論文で残された今後の課題を以下にあげる。

6.2.1　新たな特徴パターン

悪意のあるサイトに使われる技術は日々に変わっていく。今後新たな悪意のあるサイトや通信に対しては、新たな特徴パターンを洗い出す必要がある。その特徴が悪意のあるサイトや通信の検出手法として利用できるかどうかを調査することが必要である。

6.2.2　IP アドレスとの統合

今回は DNS 情報のドメイン名の特徴に着目して悪意のあるサイトの検出手法を実証した。今後ドメインと対応した IP アドレスの情報を組み合わせて、IP アドレスの特徴を分析し、またドメイン名と IP アドレスとの関連を用いて、さらに的中率の高い悪意のあるサイトの検出手法を提案する予定である。

謝辞

本修士論文の作成にあたり日頃より御指導を頂いた早稲田大学理工学部の後藤滋樹教授に深く感謝致します。本研究を進めるに上で貴重なアドバイスを頂いた千葉大紀氏、高田雄太氏、野間敬太氏、高田和也氏、胡曜氏に心より感謝致します。最後に、多大なる御協力を頂いた後藤研究室の諸氏に感謝致します。

参考文献

[1] 新種ウイルスが半年で 1 億 2400 万件「従来の対策では不十分」 – ニュース：Itpro, September 2010. http://itpro.nikkeibp.co.jp/article/NEWS/20100902/351743/

[2] Phishing Activity Trends Report 1st Half 2011, July 2012. http://www.antiphishing.org/reports/apwg_trends_report_h1_2011.pdf

[3] M. A. Rajab, L. Ballard, N. Jagpal, P. Mavrommatis, D. No- jiri, N. Provos, and L. Schmidt, “Trends in circumventing web-malware detection,” Google, Google Technical Report, Jul. 2011.

[4] Joe Stewart, GCIH, “DNS Cache Poisoning The Next Generation”, January 2003. http://www.ouah.org/DNScp.htm

[5] DNSSEC: DNS Security Extensions Securing the Domain Name System, http://www.dnssec.net/

[6] Status map of DNSSEC deployment in ccTLD and gTLD, http://www.ohmo.to/dnssec/maps/

[7] INTERNET Watch, DNSSEC 導入の負荷実験データに大きな関心, http://internet.watch.impress.co.jp/docs/event/janog26/20100714_380523.html

[8] Cricket Liu, Paul Albitz, 小柏伸夫訳, DNS & BIND 第 5 版, オライリー・ジャパン, December 2008.

[9] Intra2net, Blacklist Monitor, http://www.intra2net.com/en/support/antispam/

[10] RBL.JP, http://www.rbl.jp/

[11] SORBS, http://www.au.sorbs.net/

[12] The Spamhaus Project, http://www.spamhaus.org/

[13] SpamCop, http://www.spamcop.net/

[14] 浅見 秀雄, 阻止率 99%のスパム対策方式の研究報告, November 2009.
http://www.gabacho-net.jp/anti-spam/anti-spam-system.html

[15] Fast-Flux 攻撃とは, July 2011.
http://securityblog.jp/words/2898.html

[16] Ricardo Villamarin-Salomon and Jose Carlos Brustoloni,
“Identifying Botnets Using Anomaly Detection Techniques Applied to DNS Traffic”,
IEEE CCNC 2008, pp. 477 ∼ 480.

[17] Alexa Top Sites, http://www.alexa.com/topsites

[18] Malware domain list,
http://malwaredomainlist.com/

[19] Microsoft Technet Library: DNS,
http://technet.microsoft.com/ja-jp/library/cc779380

[20] Philip Miller, 苅田幸雄監訳『マスタリング TCP/IP 応用編』オーム社, 1998.

[21] W.Richard Stevens, 井上尚司監訳, 橘康雄訳『詳解 TCP/IP プロトコル Vol.1 プロトコル』, ピアソン・エデュケーション, 2000.

[22] 山田亮, DNS クエリーパターンを用いた OS の測定,
早稲田大学基幹理工学部情報理工学科 2011 年度卒業論文, February 2012.

[23] 民田雅人, これでいいのか TTL 短い DNS TTL のリスクを考える, JANOG 19,
http://www.janog.gr.jp/meeting/janog19/files/DNS_Minda.pdf

[24] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker,“Be- yond blacklists: learning to detect malicious web sites from suspicious urls,” in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ser,
KDD ’09. New York, NY, USA: ACM, 2009, pp. 1245 ∼ 1254.

[25] M. Akiyama, T. Yagi, and M. Itoh, “Searching structural neighborhood of malicious urls to improve blacklisting,” in Applications and the Internet (SAINT), 2011 IEEE/IPSJ 11th International Symposium on, Jul. 2011, pp. 1 $\sim$ 10.