様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　１２　日現在

機関番号：　６２６１５
研究種目：　基盤研究(B)（一般）
研究期間：　2015～2017
課題番号：　１５Ｈ０２６９９

研究課題名（和文）アンビエントDNSセンサーに関する研究

研究課題名（英文）A study on ambient DNS sensor

研究代表者

福田　健介（Fukuda, Kensuke）

国立情報学研究所・アーキテクチャ科学研究系・准教授

研究者番号：９０４３５５０３

交付決定額（研究期間全体）：（直接経費）　　14,100,000円

研究成果の概要（和文）：本研究では，インターネットワイドで生じる大きなネットワークイベントを，DNS権威サーバを用いた集権的なネットワークセンサー(DNSバックスキャッター)によって検出する手法に関して研究開発を行った．DNSバックスキャッターはイベントの発生源のIPアドレスの名前を他のホストがクエリすることで発生する．個々のバックスキャッターの情報量は小さいものの，多くのクエリが集まる大きなイベントは，機械学習を用いてそのイベントタイプを同定することが可能となった．

研究成果の概要（英文）：We design and evaluate a new type of network event sensor, called DNS backscatter, in order to detect network-wide events such as benign (CDN, mail, web crawler) and malicious (spam, scan) activities. DNS backscatter is a reverse DNS query generated at caching resolvers (queriers) close to a target for resolving IP address of an originator. We leverage on machine learning technique to identifying the type of such events. We demonstrate the effectiveness of DNS backscatter with two root DNS servers and one ccTLD.

研究分野：　インターネット工学

キーワード：　インターネット　DNS　セキュリティ

１．研究開始当初の背景

インターネットの構造がより複雑になるに連れ，インターネット上で生じている様々なイベントを知ることは難しくなっている．例えば，CDN(コンテンツデリバリーサービス)やサービスプロバイダ等のハイパージャイアントと呼ばれるプレイヤーは，地球規模のスケールでサービスを提供しているが，少数の観測点でトラフィックを見ているだけではその振る舞いを知ることは難しい．また，上記の正常なネットワークイベントの他に，異常なネットワークイベントが問題となってきている．例えば，新たなホストの脆弱性が明らかとなると，脆弱性を持ったホストを探すために大規模なネットワークスキャンが行われるが，これらのスキャンを小数の観測点から検出することは簡単ではない．

従来から取られているアプローチは，ネットワークの観測点のデータに頼るところが大きいが，インターネット規模で何かが起きていることを知ることが困難である．

２．研究の目的

本研究では，上記の正常・異常なネットワークワイドのイベントを効率良く検出する手法を確立することにある．スケーラブルに状態を監視するには，大規模分散したデータ収集を行う必要があるが，本研究では，集中的に集まる効率的なデータ収集を目指す点が，従来のアプローチとは異なる点となる．

３．研究の方法

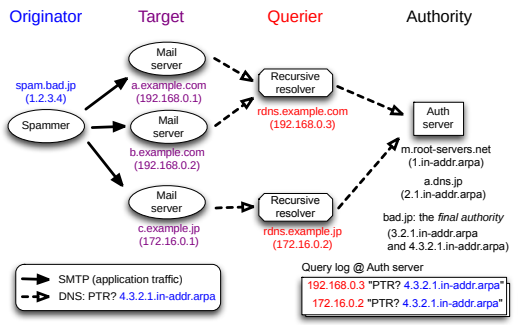


図 1: DNS バックスキャッター

本研究では，インターネット上で日常的に使われているサービスであるホスト名・IP アドレス変換を司る DNS (ドメインネームサービス)を利用する．DNS はサーバ・クライアント型のサービスであるが，クライアント(キャッシュリゾルバ)がサーバ(権威サーバ)へ問い合わせを行う際の，逆引き IP アドレスクエリに着目する．例えば，大規模な迷惑メール送信(スパム)が生じる際には，スパムの受取ホスト(ターゲット)がスパムの送信ホスト(オリジネータ)の IP アドレスからホスト名を DNS を用いて検索する．ターゲットは通常，その名前解決をキャッシュリゾルバ(クエリア)へ依頼し，キャッシュリゾルバは，名前が登録されている DNS 権威サーバ(オーソリティー)に DNS のツリー階層を辿りながら検索を行い，最終的に名前を得ることができる(図 1 参照)．キーアイディアである DNS バックスキャッターは，この DNS クエリを DNS 階層の上位の権威サーバで観測することにある．DNS 階層の最上位である，ルート DNS サーバでは，原理的には，世界中で起こっている大きなイベントを捉えることが可能である．ただし，DNS にはキャッシュ機能があるため，到達する DNS クエリ数が減少する問題がある．また，一つのクエリの持つ情報量は少ない点も問題となる．

そこで，本研究では，多数のクエリアからオリジネータの分別に必要な特徴量を抽出し，機械学習の技術を用いることで，オリジネータがどのようなイベントに関与しているかを明らかにする．

上記の DNS バックスキャッターのイベント検知能力を明らかにする上で以下の項目に関する研究を行った．

(1) DNS データ収集では，ルート DNS サーバ，および".jp"を管理する JP DNS サーバのログを収集した．これは，DITL (Day in the life of the Internet)と呼ばれる，インターネット上のトラフィック測定イベントにあわせて計測・収集されたものである．

(2) 検出イベントのクラス分け，およびそのラベル付データの構築．インターネット上の様々なデータソースを収集することで，IP アドレスとイベントを結びつけることが可能となる．例えば，メールサーバは，メーリングリストのサーバの IP アドレスを収集することでリストが得られる．同様にウェブクローラーはウェブサイトのアクセスログ，スパムはスパムアーカイブに現れる IP アドレス，スキャンはダークネットと呼ばれる経路広報を行うがホストが存在しないネットワークに到着したパケットの IP アドレス等となる．

(3) クエリアの特徴量抽出では，クエリアのホストタイプ(例えばネームサーバや ISP(インターネットプロバイダ)がカスタマに割り当てている IP アドレス)をホスト名から静的に推定する技術を確立する．さらには，クエリアの地理的多様性を IP アドレスのエントロピーからモデル化したものや，クエリ速度等の動的な特徴量を確立する．

(4) DNS で得られたデータの特徴量抽出を行ったデータとそのホストイベントのラベルを用いて教師付機械学習を行う．これにより，ラベルのついていないオリジネータに対してもラベルを推定することが可能となる．また，その検出精度を知ることができる．

４．研究成果

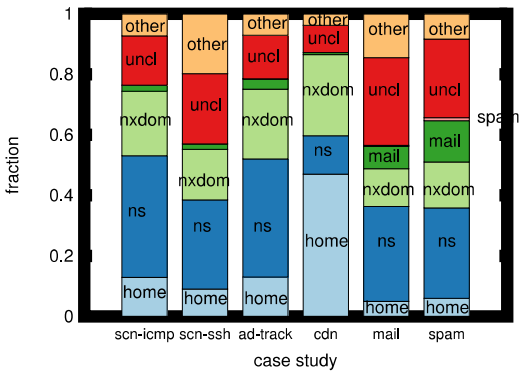


図 2: ネットワークイベント

(1) クエリアの静的タイプに基づくオリジネータの分別可能性を明らかにした．図 2 は，ネットワークイベント(スキャン，ブルートフォース攻撃，広告サーバ，CDN，メール，スパム)に対応するホストのクエリアの静的特徴量を示したものである．これらを比較すると，それぞれのネットワークイベントに関連する特徴が現れていることがわかる．例えば，CDN はホームユーザの比率が大きい，メールやスパムではメールサーバの比率が大きいことがわかる．

(2) ネットワークイベントのラベルデータを構築した．ウェブアクセスログ，スパム，ダークネット，CDN，クラウド，NTP，DNS 等のデータを収集・解析を行った．さらにネットワークスキャンデータを構築するために，バックボーントラフィック向けネットワークスキャン検出アルゴリズムを開発した．このアルゴリズムを 10 年にわたるバックボーントラフィックデータに適用し，その検出精度の確認およびスキャンの傾向を明らかにした．

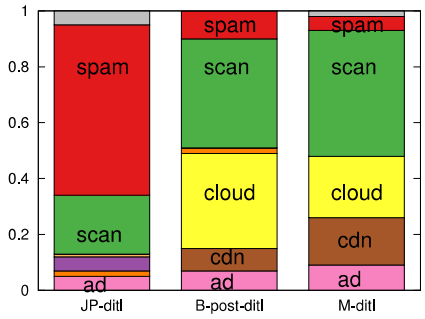
(3) 機械学習による DNS バックスキャッターを用いたネットワークイベント分別手法を開発した．上記の特徴量抽出を行ったラベルつきデータに対して，3 種類の機械学習アルゴリズム (CART(Classification And Regression Tree)，RF(Random Forest)，SVM(Kernel Support Vector Machine))を適用し，その精度を明らかにした．原理的には，機

表 1: 分類性能

dataset	algorithm	accuracy	precision	recall	F1-score
JP ditl	CART	0.66 (0.05)	0.63 (0.08)	0.60 (0.06)	0.61 (0.06)
	RF	**0.78** (0.03)	**0.82** (0.05)	**0.76** (0.06)	**0.79** (0.05)
	SVM	0.73 (0.04)	0.74 (0.05)	0.71 (0.06)	0.73 (0.05)
B post-ditl	CART	0.48 (0.05)	0.48 (0.07)	0.45 (0.05)	0.46 (0.05)
	RF	**0.62** (0.05)	**0.66** (0.07)	**0.60** (0.07)	**0.63** (0.07)
	SVM	0.38 (0.11)	0.50 (0.14)	0.32 (0.13)	0.39 (0.13)
M ditl	CART	0.53 (0.06)	0.52 (0.07)	0.49 (0.06)	0.51 (0.06)
	RF	**0.68** (0.04)	**0.74** (0.06)	**0.63** (0.05)	**0.68** (0.05)
	SVM	0.60 (0.08)	0.68 (0.10)	0.52 (0.08)	0.59 (0.09)
M sampled	CART	0.61 (0.03)	0.65 (0.04)	0.58 (0.04)	0.61 (0.04)
	RF	**0.79** (0.02)	**0.82** (0.02)	**0.77** (0.03)	**0.79** (0.02)
	SVM	0.72 (0.02)	0.76 (0.03)	0.70 (0.03)	0.73 (0.02)

械学習アルゴリズムによる差異は大きくないと予想されるが，表 1 にあるように，Random Forest アルゴリズムが最も高い精度を達成している．最も重要となる指標である，F1-score では約 80%の精度となった．この値は，元となるデータセットにも大きく依存することが明らかとなったことから，データおよびラベル情報の収集方法に関して，さらに精度向上の余地があると言える．また，ラベルデータはイベントごとに同数ではなく，大きな偏りがあるため，このアンバランスなデータから精度の向上をはかるための手法が必要である．

(4) 上記推定手法を用いて，どのようなイベントが多くのバックスキャッターデータを生成するかを明らかにした．バックス



(a) Top 100

図 3: トップ 100 オリジネータ

キャッターを生成するトップ 100 のオリジネータについて調査したところ(図 3)，JP-DNS ではスパムやスキャンが多く含まれるが，ルート DNS サーバでは CDN やクラウドと言ったインフラストラクチャに関するイベントも多く含まれることがわかった．この理由の一つは，これらのインフラストラクチャに関する IP アドレスは海外の IP アドレスであり，国内で使われているサービスであっても，国内の DNS サーバでは検出することができない

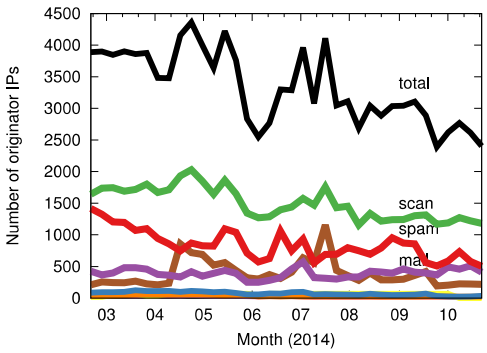


図 4: オリジネータの時間変化

ためである．しかしながら，このような大きなイベントに対応するホストが数多く検出できていることから，当初の目標

である，DNS バックスキャッターによるイベント検出は十分に行えていると言える．
さらに 7 ヶ月にわたるバックスキャッターデータを用いて，バックスキャッターで検出できたイベント数を推定することで，大規模イベントの時間推移を明らかにした(図 4)．

このように従来は大規模に分散した測定点でのトラフィック収集を必要とした大規模イベントの推定が，局所的な DNS 権威サーバへのクエリログを用いることで実現可能となったことから，本研究の目標は実現できたと考える．しかしながら，データ収集とりわけラベル情報の収集手法や，機械学習による精度の向上等，解決していくべき問題もまた明らかとなった．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　7 件）

(1) Kensuke Fukuda, John Heidemann, Abdul Qadeer, “Detecting Malicious Activity with DNS Backcatter”, IEEE/ACM Transactions on Networking, pp.3203-3218, vol.25, no.4, 2017. DOI:10.1109/TNET.2017.2724506 （査読あり）
(2) Johan Mazel, Romain Fontugne, Kensuke Fukuda, “Profiling Internet Scanners: Spatiotemporal Structures and Measurement Ethics”, Proceedings of IEEE/IFIP TMA 2017, p.1-9, 2017. DOI:10.23919/TMA.2017.8002909 （査読あり）
(3) Romain Fontugne, Patrice Abry, Kensuke Fukuda, Darryl Veitch, Kenjiro Cho, Pierre Borgnat, Hendit Wendt, “Scaling in Internet Traffic: a 14 year and 3 day longitudinal study with multiscale analyses and random projections”, IEEE/ACM Transactions on Networking, pp.2152-2165, vol.25, no.4, 2017. DOI:10.1109/TNET.2017.2724506 （査読有り）
(4) 風戸雄太，福田健介，菅原俊治，“DNS グラフ上でのグラフ分析と脅威スコア伝搬による悪性ドメイン特定”，コンピュータソフトウエア，pp.16-28，vol.33，2016．DOI:10.11309/jssst.33.3_16（査読あり）
(5) Romain Fontugne, Johan Mazel, Kensuke Fukuda, “Characterizing Roles and Spatio-Temporal Relations of C&C Servers in Large-Scale Networks”, Proceedings of WTMC 2016, pp.12-23, 2016. DOI:10.1145/2903185.2903192 (査読あり)
(6) Johan Mazel, Romain Fontugne, Kensuke Fukuda, “Identifying Coordination of Network Scans Using Probed Address Structure, Proceedings of TMA 2016, pp.1-8, 2016. http://tma.ifip.org/2016/papers/tma2016-final13.pdf（査読あり）
(7) Kensuke Fukuda, John Heidemann, “Detecting Malicious Activity with DNS Backscatter” Proceedings of ACM IMC 2015, pp.197-210, 2015. DOI:10.1145/2815675.2815706（査読あり）

〔学会発表〕（計　3 件）

(1) Kensuke Fukuda, “Tracking the evolution in residential and mobile traffic in Japan”, AINTEC 2016 (招待講演)
(2) Kensuke Fukuda, “Internet traffic anomalies and their detection techniques”, NOLTA 2016 (招待講演)
(3) 福田健介，“インターネット計測・解析”，将来のコミュニケーションクオリティ研究に向けた提言，依頼シンポジウム，電子情報通信学会ソサイエティ大会，2015 (招待講演)

〔図書〕（計　0 件）

〔産業財産権〕

○出願状況（計　0 件）

○取得状況（計　0 件）

〔その他〕

(1) 福田健介，“インターネットの諸々を計測する”，ウェブサイエンスとその周辺，研究 100 連発，サイエンスアゴラ，ニコニコ学会β実行委員会，科学技術振興機構，2015，https://www.dialogue-for-social-inclusion.com/%E3%82%A6%E3%82%A7%E3%83%96%E3%82%B5%E3%82%A4%E3%82%A8%E3%83%B3%E3%82%B9%E3%81%A8%E3%81%9D%E3%81%AE%E5%91%A8%E8%BE%BA/（アウトリーチ活動）

６．研究組織

(1)研究代表者

福田 健介（Fukuda, Kensuke）
国立情報学研究所・アーキテクチャ科学研究系・准教授
研究者番号：90435503

(2) 研究分担者

(3) 連携研究者
Johan Mazel
国立情報システムセキュリティ庁(フランス)・研究部門・研究員

(4)研究協力者
John Heidemann
南カリフォルニア大学(アメリカ合衆国)・情報科学研究所・教授
加藤 朗 (Kato, Akira)
慶應義塾大学大学院・メディアデザイン研究科・教授