様　式　Ｆ－１９－２

科学研究費助成事業　　研究成果報告書

平成 ３０ 年　 ６ 月 １３ 日現在

機関番号：６２６１５
研究種目：国際共同研究加速基金（国際共同研究強化）
研究期間：2016～2017
課題番号：１５ＫＫ００１９

研究課題名（和文）アンビエントDNSセンサーに関する研究（国際共同研究強化）

研究課題名（英文）A study on ambient DNS sensor(Fostering Joint International Research)

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）　　7,200,000円
渡航期間：　10ヶ月

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ネットワークスキャンを，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を行った．DNSバックスキャッターをネットワークスキャン検出に用いる際の問題点である，ラベル付データセットの収集に関して，ダークネットタクソノミーを構築し，ネットワークスキャンの定量化を行った．また，機械学習を行う際のデータの鮮度に関して評価を行い，データの継続的な収集の必要性・アンバランスデータでの精度向上について知見を得た．

研究成果の概要（英文）：We evaluate DNS backscatter for detecting internet-wide network scannings. We first construct a darknet taxonomy that is used for labeled data of network scans, and evaluate the effectiveness of the taxonomy with over 10 years darknet traffic data. We also evaluate the dependency of the labeled data source on the classification accuracy, and demonstrate that two class classification with SMOTE achieves acceptable performance on the scan detection.

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つれ，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．とりわけ，安全・安心なネットワークの実現のためには，セキュリティ上の脅威を正しく・迅速に見つけ出す必要がある．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少数の観測点から検出することは容易ではない．

従来から取られているアプローチでは，ローカルに設置したファイアーウォールでのログ解析や，ダークネットと呼ばれる経路広報は行うがホストの存在しないネットワークに到着するパケットを解析することが行われてきた．しかし，両者は局所的なデータであり，地球規模のカバーを得ることは不可能である．

２．研究の目的

本研究では，ネットワークスキャンに特化した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を目指す．スケーラブルに状態を監視するためには，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，DNS を用いた集権型モデルによ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．

３．研究の方法

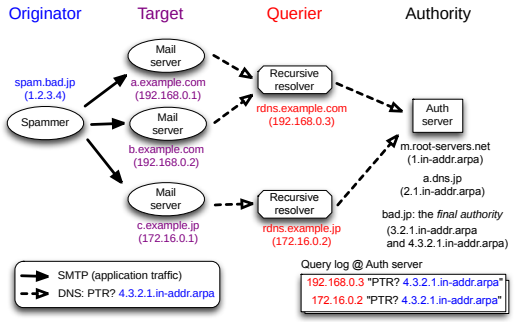

図 1: DNS バックスキャッター

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)を使用する．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，大規模なネットワークスキャンが生じる際には，スキャンを受けたホスト(ターゲット)もしくはその手前側に設置されているファイアーウォールでは，スキャン元(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，世界中で起こっているネットワークスキャンを原理的には補則可能である．ただし，DNS にはキャッシュ機能が存在するため，到達する DNS クエリが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，オリジネータであるネットワークスキャンを効率よく検出することを示す．そのために，具体的には 2 つの項目に関する研究を行った．

(1) DNS バックスキャッターでは，オリジネータの特徴量を抽出し，機械学習アルゴリズムを適用することから，ネットワークスキャンのラベル付データを作成する必要がある．そのため，ダークネットトラフィックタクソノミーによるネットワークスキャンの定量化を行った．

(2) DNS バックスキャッターを用いたネットワークスキャンの機械学習による検出手法を開発した．とりわけ，ラベル付データのデータ鮮度の影響を評価した．

４．研究成果

表 1: ダークネットタクソノミー

Anomaly	Category		Darknet Traffic Rule
Port Scan	TCP	Heavy	$(\#ipSrc == 1) \cap (\#ipDst \geq N_2) \cap (ScanFlagPktRatio \geq R\%) \cap (Avg\ \#Pkt\ per\ portDst > M)$
		Light	$(\#ipSrc == 1) \cap (\#ipDst == 1) \cap (\#portDst \geq N_2) \cap (ScanFlagPktRatio \geq R\%) \cap (Avg\ \#Pkt\ per\ portDst \leq M)$
	UDP	Heavy	$(\#ipSrc == 1) \cap (\#ipDst == 1) \cap (\#portDst \geq N_2) \cap (Avg\ \#Pkt\ per\ portDst > M)$
		Light	$(\#ipSrc == 1) \cap (\#ipDst == 1) \cap (\#portDst \geq N_2) \cap (Avg\ \#Pkt\ per\ portDst \leq M)$
Network Scan	TCP	Heavy	$(\#ipSrc == 1) \cap (\#portDst == 1) \cap (\#ipDst \geq N_1) \cap (ScanFlagPktRatio \geq R\%) \cap (Avg\ \#Pkt\ per\ ipDst > M)$
		Light	$(\#ipSrc == 1) \cap (\#portDst == 1) \cap (\#ipDst \geq N_1) \cap (ScanFlagPktRatio \geq R\%) \cap (Avg\ \#Pkt\ per\ ipDst \leq M)$
	UDP	Heavy	$(\#ipSrc == 1) \cap (\#portDst == 1) \cap (\#ipDst \geq N_1) \cap (Avg\ \#Pkt\ per\ ipDst > M)$
		Light	$(\#ipSrc == 1) \cap (\#portDst == 1) \cap (\#ipDst \geq N_1) \cap (Avg\ \#Pkt\ per\ ipDst \leq M)$
	ICMP	Heavy	$(\#ipSrc == 1) \cap (\#ipDst \geq N_1) \cap ((Type, Code) == (8, 0)) \cap (Avg\ \#Pkt\ per\ ipDst > M)$
		Light	$(\#ipSrc == 1) \cap (\#ipDst \geq N_1) \cap ((Type, Code) == (8, 0)) \cap (Avg\ \#Pkt\ per\ ipDst \leq M)$
One Flow	TCP		$(\#ipSrc == 1) \cap (\#ipDst == 1) \cap (\#portDst == 1) \cap (\#Pkt > N_3) \cap (Protocol == TCP)$
	UDP		$(\#ipSrc == 1) \cap (\#ipDst == 1) \cap (\#portDst == 1) \cap (\#Pkt > N_3) \cap (Protocol == UDP)$
Backscatter	TCP		$(\#ipSrc == 1) \cap (\#Pkt \geq 1) \cap (TCP\_Flags \in \{SA \cup A \cup R \cup RA\})$
	UDP		$(\#ipSrc == 1) \cap (\#Pkt \geq 1) \cap (\#portSrc \in \{53 \cup 123 \cup 137 \cup 161\}) \cap (Protocol == UDP)$
	ICMP		$(\#ipSrc == 1) \cap (\#Pkt \geq 1) \cap (((Type, Code) == (0, 0)) \cup (Type == 3) \cup ((Type, Code) == (11, 0)))$
IP Fragment			$(\#ipSrc == 1) \cap (\#FragmentPkt \geq 1)$
Small SYN			$(\#ipSrc == 1) \cap (\#ipDst < N_1) \cap (\#portDst < N_2) \cap (\#Pkt \leq N_3) \cap (TCP\_Flags == S)$
Small UDP			$(\#ipSrc == 1) \cap (\#ipDst < N_1) \cap (\#portDst < N_2) \cap (\#Pkt \leq N_3) \cap (Protocol == UDP)$
Small Ping			$(\#ipSrc == 1) \cap (\#ipDst < N_1) \cap (\#Pkt \leq N_3) \cap ((Type, Code) == (8, 0))$
Others			Including "Other TCP", "Other UDP", "Other ICMP" and "Other"

(1) ダークネットタクソノミーとは，ダークネットへ到着する正常ではないパケットのグループに対して，その解釈を与えたものである．本課題では，国内に設置されている 2 つのダークネットから得られた 10 年分のトラフィックデータを元に，ダークネットタクソノミーを構築した．タクソノミーの与える異常イベントとして，ポートスキャン，ネットワークスキャン，1 フロー，バックスキャッター，IP フラグメント，スモール Syn，ス

モール UDP，スモール ping，その他を定義した. 表 1 は, そのルールを示したものである. ポートスキャンおよびネットワークスキャンにおいては，そのパケット数によって，その大小をクラス分けしている.

定義されたタクソノミーはルール中のパラメータに依存することから，各パラメータの依存性および，ダークネットのネットワークサイズ，ウィンドウサイズに関する依存性について評価を行い，提案タクソノミーの妥当性を示した.

次に，現在，研究で広く使われているネットワークスキャナーである，Zmap および masscan がどの程度，ダークネットデータに含まれるのかについて調査を行った. その結果, Zmap は全ソース IP アドレス中の約 0.03%, masscan は約 0.01%を占めることを明らかにした. この数値は，全体に占める割合としては小さいものの，新たな脆弱性の発見において，今後も増加する可能性があると考えられる.

同様に，現在のスキャンを占める大きな原因である，ウィルス等の自動的なスキャンについて着目した. 世間的にも大きな影響を与えた Conficker ワームはホストの TCP ポート 445 に接続を試みることから，別設置のハニーポットデータおよびダークネットデータにより，ダークネット内の Conficker ワームの割合を算出した. 小規模 TCP ネットワークスキャンおよびスモール Syn に該当するソース IP アドレスは，2010 年には 30%以上を占めていたものの，2015 年には 10%程度と減少していることが明らかとなった. これは，ワーム自身の蔓延には歯止めがかかっているものの，未だに一定の割合で感染しているホストが存在することを示している. 同様に，近年話題となっている，IoT 機器に感染する Mirai ボットは脆弱性の存在するホストを探すために，TCP23 および TCP2323 ポートをスキャンする. この活動は，ダークネットタクソノミーでは，同様に，小規模 TCP ネットワークスキャンおよびスモール Syn に該当する. 2016 年 10 月には，全ソース IP アドレスのうち約 70%が Mirai ボットによるものと推定されており，このボットの影響が非常に大きいことがわかった(図 2). しかし，ネットワーク管理者やベンダの活動により，その割合は 2017 年 4 月には 20%程度まで減少している.

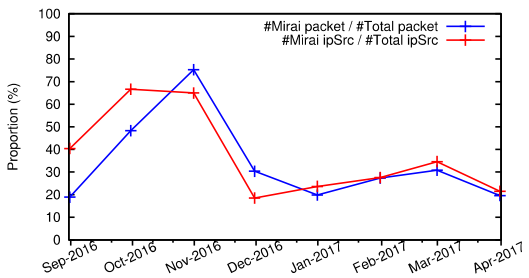

図 2: Mirai ボットの割合

このように，ダークネットタクソノミーはダークネットへ到着するフローの分類に有益であり，DNS バックスキャッターでのスキャン分類において，信頼性の高いラベル付データとなることを示している.

(2) DNS バックスキャッターにおけるラベル付データの学習に対する影響を評価した. DNS バックスキャッターでは，ラベル付データの取得がその検出精度に影響を受けると予想される. そのため，ある特定期間に収集された，正常・異常のラベル付データを用いて学習を行い，その期間より過去および未来のデータを用いた評価を行った. 図 3, 4 は，それぞれ正常ラベルおよび異常ラベルのついた IP アドレスがどの程度安定的に過去・未来のデータに現れるかをプロットしたも

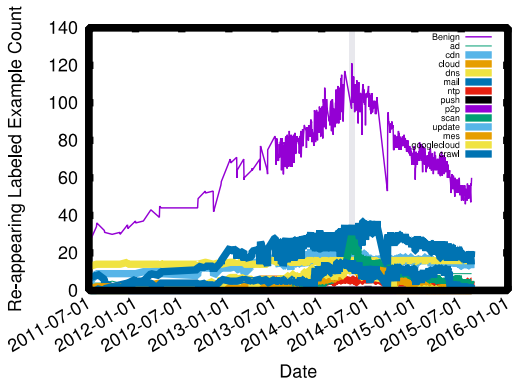

図 3: 正常イベントのオーバラップ

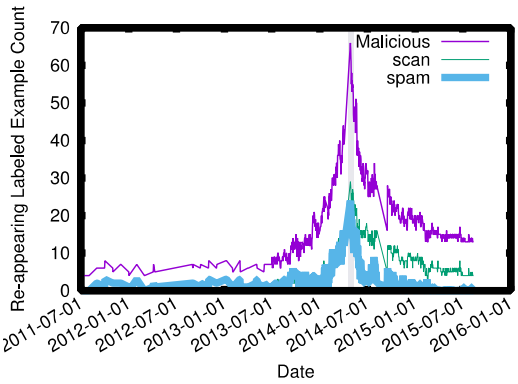

図 4: 異常イベントのオーバラップ

のである. 二つの図を比べてみると，正常なイベントでは，比較的プロットがフラットであるのに対して，異常なイベントでは大きな減衰が見てとれる. これは，正常なイベントは，インターネット上に定常的に存在しているが，異常なイベントは一度生じた後，すぐさま消え去ることを表している. つまり，異常なイベントのオリジネータは何らかの対処がなされることでイベントの生成が阻害される結果となった.

さらに，ラベル付データの取得期間と学習の精度を知るために以下の 3 つの手法を調査し

た．① 収集したラベル付データセットのみを学習に使用，② 収集したラベル付データ

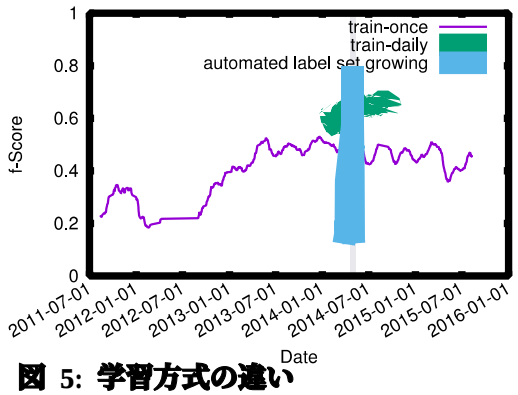

図 5: 学習方式の違い

セット内の IP アドレスのうち，ラベルが存在しない期間に現れる IP アドレスから特徴量抽出し学習に使用したもの，③学習データとして，ラベル付データに加えて分類器が出力したラベルをラベル付データとして使用した場合．図 5 の結果に示すように，一度の学習データを使用する場合の性能劣化は大きく，IP アドレスを複数の曜日データに適用し，学習データとすることで，より高い学習効果を得ることができた．さらに，学習した結果から生成された新しいラベルを使用する場合には，ほぼその新しいラベルは性能向上に寄与しないことが明らかとなった．分類性能の向上のためには，適当な期間のラベルデータ収集，とりわけ異常のラベルデータが必要であることが示された．

表 2: 分類性能

	multi class	2 class	2 class + smote
DT	0.47	0.66	0.68
RF	0.67	0.68	0.76
SVM	0.37	0.55	0.54

さらにバックスキャッターによるイベント分類をスキャンに特化するために，マルチクラスの分類ではなく，2 クラス分類による評価を行った．潜在的な問題として，スキャンのラベル付データはその存在数が少なく，その他のラベル付データの存在数が大きいことから，機械学習手法を単純に適用する場合には，期待する精度が得られないというデータのインバランス問題が生じる．そのため，確率的なアップサンプリング手法(SMOTE: Synthetic Minority Over-sampling Technique)を用いることでこの問題に取り組んだ．これら 3 種類の特徴量データに対して，3 種類の機械学習アルゴリズム(DT; Decision Tree, RF; Random Forest, SVM; Support Vector Machine)を適用した場合の分類性能(F 値)を表 2 に示す．分類性能としては，RF が最も性能が高い結果となったが，マルチクラス，2 クラスの性能差はほぼ確認できなかった．それに対して，2 クラス+アップサンプリング手法ではインバランス問題が改善され，より高い分類性能が得られる結果となった．これは，ラベル付データの収集の観点では大変興味深い結果である．すなわち，ある一定数のラベル付スキャンイベントのデータをダークネット等の他のデータ源より一定間隔で収集することができれば，かなりの精度でスキャンの分類が可能となることを意味している．

今後の課題としては，より発見が困難な IPv6 ネットワークでのネットワークスキャンが検出可能かという点である．IPv6 ネットワークはアドレス空間が 128 ビットあり，IPv4 ネットワークと比較するとネットワークスキャンの効果が薄いとされている．しかし，Zmap 等の高速スキャンツールが IPv6 でも使用可能となり，IPv6 ネットワークでのスキャンは今後より増えると予想される．アドレスのカバー率という意味ではダークネットには限界があることから，DNS バックスキャッター技術のさらなる応用が期待できると考える．

５．主な発表論文等
（研究代表者は下線）

〔雑誌論文〕（計 2 件）

(1) Jun Liu, Kensuke Fukuda “An Evaluation of Darknet Traffic Taxonomy”, Journal of Information Processing, pp.148-157, vol.26, DOI: 10.2197/ipsjjip.26.148, 2018（査読有り）

(2) Kensuke Fukuda, John Heidemann, Abudul Qadeer, “Detecting Malicious Activity with DNS Backscatter Over Time”, pp.3203-3218, vol.25, DOI:11.1109/TNET.2017.2724506, 2017（査読有り）

〔学会発表〕（計 0 件）

〔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〕
ホームページ等

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)研究協力者
〔主たる渡航先の主たる海外共同研究者〕
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

〔その他の研究協力者〕
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
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