様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月　 ５ 日現在

機関番号：１４５０１
研究種目：基盤研究(C)
研究期間：2012～2014
課題番号：２４５００１７３

研究課題名（和文）追加学習とアクティブ学習を導入した学習型マルウェア検出・分類システムの開発

研究課題名（英文）Development of Malware Detection/Classification System Introducing Incremental Learning and Active Learning

研究代表者

小澤　誠一（Ozawa, Seiichi）

神戸大学・工学（系）研究科（研究院）・教授

研究者番号：７０２１４１２９

交付決定額（研究期間全体）：（直接経費）　　3,900,000円

研究成果の概要（和文）：本研究では，マルウェア感染を誘導する悪性スパム攻撃からネットユーザを守り，マルウェアなどによる悪意ある活動を広域的に観測するため，機械学習を導入した３つの学習型システムを開発した．一つは悪性度の高いスパムメールを自動収集し，クローラ型ウェブ解析システムによる自動ラベリングとオンライン学習可能な悪性スパム検知システムである．二つ目は，未使用IP群であるダークネットへのパケットを収集し，そのトラフィック特徴をクラスタリングすることで，サブネットのマルウェア感染状況を広域監視するシステムである．最後は，ダークネットトラフィックの特徴からDDoS攻撃のバックスキャッタを判定する広域監視システムである．

研究成果の概要（英文）：In order to protect network uses from malicious spam mail attacks that can lead to malware infections and to conduct a large-scale monitoring of malicious activities by malwares, we developed three types of learning systems introducing machine learning techniques. First, we developed a malicious spam mail detection system with the following three sophisticated functions: an automatic mechanism to collect suspected malicious spam mails, an automatic labelling (malicious or benign) function for collected spam mails by a crawler-type of web security analyzer, and online learning function for automatically collected training data. Second, we developed a large-scale monitoring system which can observe transitions of subnet infection states by allocating the most similar typical patters obtained by performing the hierarchical clustering for darknet traffic features. Finally, we developed a large-scale monitoring system which can detect DDoS backscatter from observed darknet traffic features.

研究分野：知能情報学

キーワード：サイバーセキュリティ　機械学習　オンライン学習　悪性スパムメール検知　ダークネット解析　DDoS　バックスキャッタ判定　マルウェア感染モニタリング　テキスト解析

１．研究開始当初の背景

近年，マルウェア感染による被害がその深刻さを増しており，企業の知的財産や企業秘密情報を狙ったもの，個人の財産を狙うものなどが現れ，年々凶悪化している．2011 年の統計では，約 1.37 秒ごとに新型または亜種のマルウェアが発生しているとされ，前年に比べて 14.5%増加した．急速に増殖する，このような脅威に対して，もはやブラックリストベースのアンチウィルスソフトだけでは，感染による深刻な被害が生じる前に検出・処理することが困難になりつつある．計算機とネットワークが，我々の社会経済のインフラとして欠かせない存在である以上，サイバー世界でのセキュリティを確保することは最優先課題の一つであり，早急に根本的な対策を要する．

この問題に対し，最近，機械学習を取り入れた学習型マルウェア検出システムが注目を浴びている．Xu らは，計算機上で稼働する OS やソフトウェアのメッセージ（コンソールログ）から，マルウェア感染などで生じる不正プロセスを検出する方法を提案している．この方法では，コンソールログから抽出されたキーワード情報を特徴ベクトルとし，まず正常プロセスの特徴ベクトルに主成分分析（PCA）を適用して特徴空間（PCA の場合，固有空間に相当する）を学習する．そして，この特徴空間の補空間成分を求め，その大きさによって不正プロセスの検出を行う．また，Masud らは，実行形式から抽出されたバイナリ n-gram と実行形式を逆アセンブルした命令語や DLL コールを特徴ベクトルとして不正プロセスの検出を行う学習モデルを提案している．このモデルはアンサンブル学習器によって構成され，オンラインで追加学習を行えるだけでなく，識別対象の特性（クラス決定境界など）が時間的に変化する，いわゆるコンセプトドリフトにも対応している．

上記の先行研究では，以下のようなマルウェアの特性を考慮した学習方式が提案されている．

(1)マルウェアに特異的な挙動が，正常挙動の特徴空間とは異なる領域に現れる．
(2)新種や亜種の出現でマルウェアの挙動パターンは多様化し，検出精度を上げるには，識別器をオンラインで追加学習し，コンセプトドリフトにも対応する必要がある．

しかし，サイバーセキュリティ分野で実用性の高い学習システムを構築するには，上記 2 点だけでなく，次のような仕組みを導入した学習方式を取り入れる必要がある．

(3)マルウェアなどによる悪意ある活動は膨大な通信パケットの中のごく一部に含まれており，大多数の通信パケットはマルウェアに関係のない通常通信のものである．よって，通常通信と悪意ある通信をふるい分ける仕組みが必要である．
(4)通信パケットが悪意ある活動の結果生じたものか，そうでないかの情報は通常与えられない．よって，教師なし学習を適用するか，膨大な量の通信トラフィックパターンに対して，効率よく教師ラベルを与える仕組みを導入する必要がある．
(5)膨大な量の通信パケットデータが常時発生している状況で追加学習を行うには，学習に有効なデータのみを選択して学習する仕組みを導入する必要がある．
(6)未知の通信トラフィックパターンを識別器に判定させることは，信頼性の観点から適切でない．よって，識別器には「判定できない」と出力させて，監視者に判断をゆだねる仕組みを導入する必要がある．

(3)は学習データの自動収集機能，(4)は自動ラベリング機能（監視者がシステムに割り込む形でラベルを与えるものも含む），(5)はアクティブ学習機能，(6)は判定結果の外れ値検出機能と呼ばれる．よって，本研究で開発すべき学習システムは，以上 4 つの機能を目的別に複数個組み込んだ自律学習型システムと言え，このようなアプローチは，著者の知る範囲では従来のサイバーセキュリティの研究にはない．

２．研究の目的

本研究では，悪意をもって送り込まれるマルウェアの感染は完全に防げないことを前提に，その感染を通して行われる悪意ある活動を検出・分類する学習型マルウェア検出・分類システムの開発を目指す．悪意ある活動は多種多様かつ時間とともに変化し，前節の(3)～(6)で述べたような特性も持ち合わせている．よって，開発すべきシステムは，どのような悪意ある活動を検知・分類するのかに応じて適切な学習スキームを導入する必要がある．

本研究では，以下の３つの課題に取り組み，学習型システムの開発を行う．

①悪性スパムメールの悪性度判定

スパムメールの目的は単純な広告からマルウェア感染サイトへの誘導など多様化している．特に，URL が含まれている場合，その危険性を判断することは非常に困難である．悪性度は，URL に接続することで判定可能であるが，マルウェア感染の危険性や接続に要する時間的コストが高く個人で行うことは現実的でない．また，スパムメールの文面などは，攻撃者のキャンペーンにより不定期に変化し，ブラックリスト方式などで対応することは難しい．そこで，本研究ではスパムメールの悪性度を 2 段階に定義し，スパムメールを受信するたびに悪性度を判定する逐次学習型システムを提案する．スパムメールの悪性度は URL を安全に巡回するクローリングシステムを利用することで判定

し，学習用ラベルとして扱う．その後，ラベリング結果をもとにオンライン学習し，辞書や識別器の更新を逐次行うことで高い分類制度を常に保つことが目的である．つまり，前述の(2)～(5)の仕組みを取り入れた新しい学習システムを開発する．

②ダークネットトラフィックに基づくサブネットの広域的脆弱性判定

近年，コンピュータ及びネットワークの発達により多くの人がその恩恵を享受するようになったが，不正なプログラム（マルウェア）による信用不安も広がっている．マルウェアは他のコンピュータの脆弱性探索などの感染行為を行うとき実在しない IP アドレス（ダークネット）を宛先としたパケットを発生させることがある．また，マルウェアに感染した多くのコンピュータが短時間に特定のサーバに向け集中的にパケットを送信し，機能を停止させる攻撃を DDos 攻撃というが，この攻撃は虚偽の送信元を記載したパケットを標的に送信するため，これを受信したサーバからの返答（バックスキャッタ）が，ダークネットに届くことがある．よって，ダークネットに届くパケットはマルウェアの活動に連動して発生している可能性が高く，その解析によりマルウェアによる感染の程度や拡大の様子を知ることが可能と考えられる．しかし，世界中のコンピュータの個々の感染の把握は困難である．IDS，IPS，FIREWALL などマルウェア対策ソフトウェアは組織ごとに適用される場合が多く，グループ単位で特定のマルウェアに感染することが多いことより，組織（サブネット）ごとに解析することで，インシデント観測を効率よく行える．危険なサブネットをあらかじめ把握できると，そのサブネットと通信をしている組織への警告，危険なサブネットのユーザへの注意喚起ができ，感染拡大阻止に効果があると考えられる．本研究では，サブネットの脆弱性を判定することを目標とし，サブネットの通信トラフィックの様子から特徴ベクトルを生成し，それを用いてサブネットのクラスタリングを行う．さらに，クラスタリングされたサブネットの通信トラフィックを解析，可視化を行い，クラスタリングの有効性を検証することを目的とする．つまり，前述の(1) (3)(4)(6)の仕組みを取り入れた新しい学習システムを開発する．

③ダークネットトラフィックに基づく DDoS 攻撃の広域的バックスキャッタ検知・分類

様々なサービスがネットワークに依存している現在，これらのサービスの停止は一時的なものでさえ多額の損害が発生しうる．その結果，攻撃者は金銭的損害を与えることが目的で，ボットに感染した多数のコンピュータを操り，攻撃対象にパケットを大量に送りつける DDoS 攻撃が行われるようになった．攻撃者は自身の特定を困難とするために送信元の IP アドレスを詐称することが多い．特に DDoS 攻撃では，攻撃者が多数のコンピュータを操り，複数のホストが送信元を詐称してパケットを送りつける．このとき攻撃を受けたサーバは，正常な通信と DDoS 攻撃による通信を容易には区別できず，また攻撃者の身元を特定することも困難である．そのため，詐称した IP アドレスを用いた DDoS 攻撃をいち早く発見し，サービス停止に追い込まれる前に対応することが課題となっている．攻撃を受けたサーバは，通常通信と攻撃の通信が区別できないため，すべての通信に対して返答を行う．そのため，詐称された送信元に対しても返答を行い，コンピュータ等が存在しない未使用 IP アドレスのネットワーク（ダークネット）に対しても返信パケットが送信される．この返信パケットは DDoS 攻撃の跳ね返りパケットであり，バックスキャッタと呼ばれる．ダークネットからバックスキャッタを観測することで DDoS 攻撃の特徴や傾向などを分析し，早期に攻撃を発見し，DDoS 攻撃によってサーバがサービス不能になる前に対策を行うことが可能となる．ダークネットを利用する利点としては，通常のサーバやネットワークの観測とは違い，広範囲にパケットを得ることが容易に可能なため，DDoS 攻撃の特徴が得られやすいこと，ダークネットに通常の通信によるパケットが来ることはほとんどないため，異常な通信の検出が容易という点が挙げられる．

本研究では，ダークネットで観測された短時間のパケットデータから DDoS 攻撃によるバックスキャッタを判別する手法を提案する．この際，ラベル付けが機械的に可能な 80 番ポートからの TCP パケットと 53 番ポートからの UDP パケットを訓練データとして用い，ラベル付けが機械的に可能でないパケットに対するバックスキャッタの判別を行うことを目的とする．つまり，前述の (2)～(6)の仕組みを取り入れた新しい学習システムを開発する．

３．研究の方法

①悪性スパムメールの悪性度判定

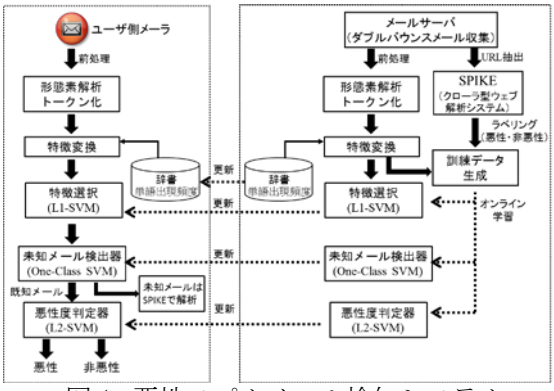

図１ 悪性スパムメール検知システム

提案システムを図１に示す．システムは大きく２つに分かれ，メーラソフトのプラグインとしてユーザに提供する悪性度判定部（図１の左側）とそこで使われている３つの識別器モデルを学習する学習部（図１の右側）で構成される．３つの識別器は，特徴選択を行う L1-SVM，未知メールを判定する One-Class SVM，そして悪性度判定を行う L2-SVM である．これら３つの識別器モデルは，学習部で生成された訓練データを用いてオンライン学習され，随時更新される．

学習に利用する訓練データ（スパムメール）の収集を行い，そのスパムメールの中に含まれる URL の悪性度を SPIKE と呼ばれるクローラ型ウェブ解析システムで調べ，それに基づいて自動ラベリングが行われる．なお，識別器に与える入力特徴は，まず形態素解析とトークン化を行い，単語出現頻度(Term Frequency, TF)に基づいて Bag-of-Words アプローチで特徴変換される．

②ダークネットトラフィックに基づくサブネットの広域的脆弱性判定

広域的かつリアルタイムのマルウェアの感染状況の把握のため，感染単位をネットワークの部分空間であるサブネットとし，マルウェアの感染推移を追跡し，ネットワークの異常検出を行うためのシステムを開発する．サブネットの脆弱性は，鈴木らが提案した TAP 分析とマルウェアシグネチャを特徴量として判定する．TAP 分析とは，宛先 IP アドレスやポート数などの情報から送信元ホストの振舞いタイプを高速かつ効率的に分類する手法である．マルウェアシグネチャとは，TAP 分析の結果にポート番号などのマルウェア特有の特徴を組み合わせた判定基準である．１日ごとのこの TAP 分析結果の判定の分布を特徴ベクトルとし，階層クラスタリングを行い，属するクラスタの時間変化をモニタリング（可視化）することで感染推移を監視する．

③ダークネットトラフィックに基づく DDoS 攻撃の広域的バックスキャッタ検知

図２に開発した DDoS バックスキャッタ検知システムの処理フローを示す．攻撃を受けたサーバは，通常通信と攻撃の通信が区別できないため，すべての通信に対して返答を行う．そのため，詐称された送信元に対しても返答を行い，コンピュータ等が存在しない未使用 IP アドレスのネットワーク（ダークネット）に対しても返信パケットが送信される．この返信パケットは DDoS 攻撃の跳ね返りパケットであり，バックスキャッタと呼ばれる．ダークネットからバックスキャッタを観測することで DDoS 攻撃の特徴や傾向などを分析し，早期に攻撃を発見し，DDoS 攻撃によってサーバがサービス不能になる前に対策を行うことが可能となる．

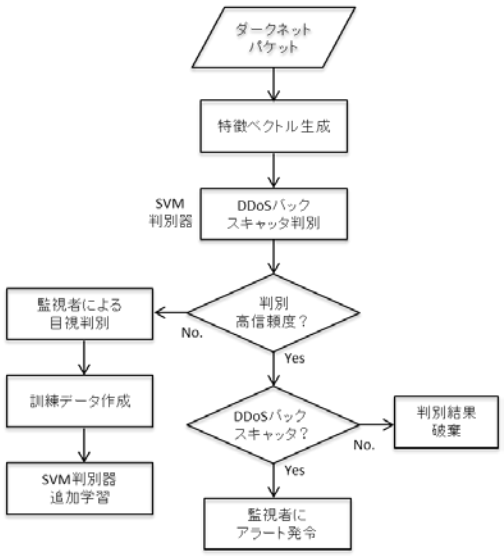

図２ 追加学習を導入した DDoS バックスキャッタ判別システム

本研究では，ダークネットで観測された短時間のパケットデータから DDoS 攻撃によるバックスキャッタを判別する追加学習型システムを提案する．訓練データと評価用データには，バックスキャッタか否かのラベル付けが機械的に可能な 80 番ポートからの TCP パケットと 53 番ポートからの UDP パケットを用いる．判別に用いる特徴は送信元ポートの番号自体の情報は用いず，送信元/送信先ポートや送信先 IP の統計情報，ペイロードに関するものなど，他の通信形態でも見られるものを使用する．送信ホストごとにダークネットパケットを特徴ベクトルに変換し，SVM 識別器を使って DDoS バックスキャッタ判定を行う．この際，判定の信頼度が高い場合は識別器の出力を採用し，バックスキャッタと判定されたら監視者にアラートを出す．一方，判定の信頼度が低い場合，ダークネットトラフィックの特徴をタイルグラフで可視化し，監視者による目視判別が行われる．このとき，バックスキャッタと判定されれば，その特徴にクラスラベルを付加して SVM 識別器の追加学習を行う．

４．研究成果

①悪性スパムメールの悪性度判定

性能評価には，情報通信研究機構（NICT）で 2013 年 11 月 1 日～2014 年 9 月 23 日の 326 日間収集された 61,477 通のスパムメールを用いた．最初の 60 日間のスパムメールで初期学習を行った後，過去 30 日間のスパムメールを用いて，一日ごとにオンライン学習を行う．オンライン学習を行うのは，図１に示した特徴選択を行う L1-SVM，未知メール検出(Outlier Detection)を行う One-Class SVM，悪性度判定を行う L2-SVM の３つの識別器モデルである．未知メール検出機能を導入するシステムと導入しないシステムの

適合率（precision），再現率（recall），F 値を調べたところ，表 1 に示すような結果が得られた．また，オンライン学習による悪性度検知の精度（accuracy）を図 3 に示す．

以上より，未知メールを検出し，それを SPIKE に与えて悪性度解析を行ったうえで追加学習する効果が示された．

表 1　未知メール検出機能の有無に対する性能比較[%]

	適合率	再現率	F 値
あり	91.0	96.2	93.5
なし	90.1	94.3	92.2

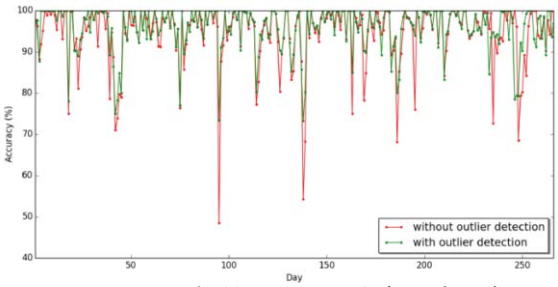

図 3　オンライ学習による正確度の時間変化

②ダークネットトラフィックに基づくサブネットの広域的脆弱性判定

評価実験には，NICT の/16 ダークネットセンサで 2014 年 2 月 1 日～2 月 28 日の 28 日間に観測された 303,733,994 パケットのデータを使用した．IP 空間を/16 サブネットに分割し，個々のサブネットから送出された 30 秒間のダークネットパケットを TAP 分析を通して特徴ベクトルに変換し，階層型クラスタリングで 15 個の典型的な感染状況を表す特徴ベクトル（プロトタイプ）を得た．このプロトタイプを色表示し，50 のサブネットに対し，2 月 1 日～28 日までの感染状況を可視化したものを図 4 に示す．

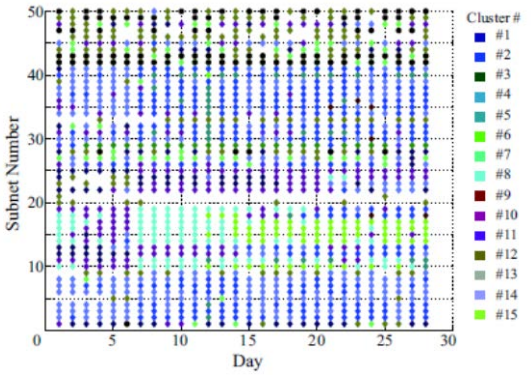

図 4　サブネットの感染状況の時間推移

これからわかるように，サブネット#10 と #14～#17 が類似した時間推移パターンを持っていることがわかる．この推移では，クラスタ番号が#10 → #8 → #15 と移り変わっており，他のサブネットに対して特異的な推移パターンをもっている．これらのサブネットで何が行っているかを調べたところ，5000 番ポートへのパケットが際立って多いことがわかった．これは，2 月 28 日にパンデミックが報告された Synology NAS (port 5000/TCP)を攻撃するマルウェアの活動であることがわかった．

興味深いことに，開発した脆弱性判定システムでは，2 月 7 日頃から特異的な遷移パターンを観測していることである．これは，パンデミックが起こる 20 日前に新種のマルウェアの活動を検知したことを示している．このことから，開発したシステムが未知のマルウェアによる感染拡大の予兆を捉えられる可能性が示唆された．

③ダークネットトラフィックに基づく DDoS 攻撃の広域的バックスキャッタ検知

NICT のダークネットセンサで観測された 2013 年 1 月から 2014 年 1 月までのパケットデータを用いた．このうち，2013 年 1 月から 2013 年 12 月までにおいて，80 番ポートと 53 番ポートから送信されたパケットのデータを初期データとし，2014 年 1 月における 80 番ポートと 53 番ポート以外のパケットデータを評価データとした．また，本性能実験では SVM の判別信頼度による判別後の処理の区別は行わず，すべての評価データに対して，目視による評価と追加学習を行う．まず，評価データのうち，1 月 1 日から 10 日までのデータから作成した評価データを学習させた判別器に与え，判別を行う．そして，これらのホストの活動を確認し，手動でラベルを付与し，評価を行う．つぎに，評価を行った評価データを新たに訓練データに加え，判別器の再学習を行う．再学習後，1 月 11 日から 20 日までのデータから作成した評価データに対して判別を行い，同様の手順で評価・再学習を行う．最後に，1 月 21 日から 31 日までのデータで作成した評価データに対して判別・評価を行う．各データにおけるラベリングの内訳を表 2 ～ 3 に示す．

表 2　初期データとして用いたパケット数

	Backscatter	Non Backscatter
80/TCP	6,869	411
53/UDP	2,525	168

表 3　訓練データとして用いたパケット数

	Backscatter	Non Backscatter
1～10 日	1,433	1,575
11～21 日	1,037	1,847
22～31 日	871	1,553

表 4　追加学習による性能改善［%］

	真陽率	陰陽率	正確度
1～10 日	79.7	10.0	42.6
11～21 日	85.0	92.3	89.2
22～31 日	81.6	91.4	87.7

評価実験の結果を表４ に示す. 1〜10 日のダークネットパケットを評価データとして与えたとき, 陰陽率が 10%, 正確度(accuracy)が 42.6%と大きく性能が劣化している．これは，1〜10 日の評価データに，初期データにはなかった DDoS バックスキャッタパターンが含まれており，これを非バックスキャッタと判定したためである．ここで誤判定されたパターンは，監視者がタイルグラフを見て，バックスキャッタか否かの判定をやり直し，その結果をクラスラベルとして与えて追加学習する. その結果, 表４からわかるように，11〜20 日における真陽率，真陰率，正確度ともに向上し，その後の 21〜31 日でも，性能を高く維持できている．これは，監視者の知識に基づいてラベリングを行って追加学習した効果であり，教師ラベルが陽に与えられないパケットデータのうち，ラベリングが必要なものに限定して行うことの有効性を示唆している.

本研究では，図２に示した識別器の信頼度に基づいた監視者への目視判別プロセスを導入していないが，これを導入することで，今後，大規模なダークネットパケットデータを扱ったリアルタイム・オンライン学習が可能になると考えられる．これについては，今後の課題としたい.

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１９件）

① Hironori Nishikaze, Seiichi Ozawa, Jun Kitazono, Tao Ban, Junji Nakazato, Jumpei Shimamura, Large-Scale Monitoring for Cyber Attacks by Using Cluster Information on Darknet Traffic Features, Proc. of 1st INNS Conference on Big Data, 査読有, 2015, 8 pages (in press)

②Ali Siti Hajar Aminah, Seiichi Ozawa, Tao Ban, Junji Nakazato, Jumpei Shimamura, An Online Malicious Spam Email Detection System Using Resource Allocating Network with Locality Sensitive Hashing, Journal of Intelligent Learning Systems and Applications, 査読有, Vol.7, No.2, 2015, pp. 42-57.
DOI: 10.4236/jilsa.2015.72005

③Annie A. Joseph, Takaomi Tokumoto, Seiichi Ozawa, Online Feature Extraction based on Accelerated Kernel Principal Component Analysis for Data Stream, Evolving Systems, 査読有, 2015, 13 pages.
DOI: 10.1007/s12530-015-9131-7

④ Yuli Dai, Shunsuke Tada, Tao Ban, Junji Nakazato, Jumpei Shimamura, Seiichi Ozawa, Detecting Malicious Spam Mails: An Online Machine Learning Approach, Neural Information Processing, Lecture Notes in Computer Science 8836, 査読有, 2014, pp 365-372
DOI: 10.1007/978-3-319-12643-2_45

⑤Nobuaki Furutani, Tao Ban, Junji Nakazato, Jumpei Shimamura, Jun Kitazono, Seiichi Ozawa, Detection of DDoS Backscatter Based on Traffic Features of Darknet TCP Packets, Proc. 9th Asia Joint Conference on Information Security, 査読有, 2014, pp. 3-5.
DOI: 10.1109/AsiaJCIS.2014.23

〔学会発表〕（計２０件）

①古谷暢章，班 涛，中里純二，島村隼平，北園淳，小澤誠一，ダークネットトラフィック観測による DDoS バックスキャッタ判定，電子情報通信学会情報通信システムセキュリティ研究会，2014.11.28，東北学院大学（宮城県）

②多田隼輔，中里純二，班 涛，小澤誠一，スパムメールに対するオンライン悪性度判定システムの開発，暗号と情報セキュリティシンポジウム，2014.1.24，城山観光ホテル（鹿児島県）

③西風宗典，班涛，小澤誠一，ダークネットトラフィックデータの解析によるサブネットの脆弱性判定に関する研究，コンピュータセキュリティシンポジウム 2013，2013.10.23，サンポートホール（香川県）

６．研究組織

(1)研究代表者
小澤 誠一（OZAWA, Seiichi）
神戸大学・大学院工学研究科・教授
研究者番号：70214129

(2)研究分担者
安藤 類央（ANDO, Ruo）
情報通信研究機構・ネットワークセキュリティ研究所・主任研究員
研究者番号： 30446596

(3)連携研究者
北園 淳（KITAZONO, Jun）
神戸大学・大学院工学研究科・助教
研究者番号：00733677

班 涛（BAN, Tao）
情報通信研究機構・ネットワークセキュリティ研究所・主任研究員
研究者番号：80462878

中里 純二（NAKAZATO, Junji）
情報通信研究機構・ネットワークセキュリティ研究所・研究員
研究者番号：60435782

(4)研究協力者
島村隼平（SHIMAMURA, Jumpei）