2012年度 SWAN Working Group 活動報告

宮本大輔 (daisu-mi@nc.u-tokyo.ac.jp)
Gregory Blanc (gregory.blanc@telecom-sudparis.eu)　門林雄基 (youki-k@is.naist.jp)

2013年1月29日

目 次

1 はじめに

SWAN (Security for Web 2.0 Application) WG では，最近の悪意あるウェブサイトの動向を観測し対策を検討している．ウェブを介した攻撃にはその攻撃空間が広いという特徴があり，本研究グループはその広い特性に対応した研究を行なっている．これまでの活動としては，脆弱性を持つウェブ2.0のアプリケーションをWIDEメンバーに提供する試みや，悪意あるウェブサイトによく見られる難読化されたJavaScriptの構造に着目した解析，PCだけではなくAndroidなどで動くマルウェアの解析技術のハンズオンなどが挙げられる．

今年度は，ウェブを悪用した攻撃であるフィッシングサイトに着目した研究を行った．マルウェアではなく人間を騙すことによって攻撃が発生する点でこれまでの研究とは異なるが，SWAN WGでは人間の行動だけではなく，フィッシングサイトをホスティングしているサーバ，関連して発生するネットワークインフラへの影響など，WGの趣旨通り，広い特性を考慮した研究開発を行なっている．

以降SWAN WGの本年度の主な活動について報告する．

2 フィッシングサイト検知技術の個人に合わせた調整手法の研究

本研究課題では，ユーザの過去の判断 (Past Trust Decision, PTD) の記録を活用したフィッシングサイト検知についての考察を行う．我々の先行研究では，ユーザがウェブサイトを見て「正規サイトである」「フィッシングサイトである」という判断を行った結果を，既存のヒューリスティクスを用いたフィッシングサイト検知手法に取り入れることを提案している．しかし先行研究では平均的な精度こそ向上していたものの，ユーザによっては彼/彼女らのPTDを用いない場合に精度が向上する場合が確認された．そこで本論文では，PTDを活用すべきユーザ，そうでないユーザについて考察を行う．ユーザを分類する手法として，ユーザの判断能力の構成要素をコンテンツの文言のみで判断を行っていないこと，URLやSSLを確認し怪しさに気付いていること，当該サイトを過去に利用した経験を判断に活用できていることなどあると仮定し，これら要素に基づいたクラスタ分析を行う．また，クラス

表 1: PTD データベースの例

URL	Actual Condition	The user's decision	Heuristics #1	···	Heuristics #N
Site 1	phishing	phishing	phishing	···	legitimate
Site 2	phishing	phishing	phishing	···	legitimate
Site 3	phishing	phishing	phishing	···	legitimate
···	···	···	···	···	···
Site M	legitimate	legitimate	legitimate	···	phishing

タ毎に先行研究の提案によるフィッシングサイト検知を行い，活用すべき PTD をもつユーザの傾向について考察を行う．

2.1 HumanBoost

先行研究である，HumanBoost [2] 方式について概要を説明する．HumanBoost 方式は，エンドユーザがウェブサイトを信頼できる，信頼できないといった判断を行った結果を，フィッシングサイトの検知に活用するという提案である．エンドユーザはウェブサイトに個人情報を入力する時，つねに何らかの意思決定を行っていると考えられる．言い換えれば，エンドユーザはウェブサイトに対し正規サイトあるいはフィッシングサイトであるという出力を行う装置であるとも考えられる．

仮に，表 1 のように，各エンドユーザに PTD のデータベースが存在していると考える．データベースのスキーマはウェブサイトの URL 及びそのサイトの実際の状況，さらにユーザの意思決定の結果と既存のヒューリスティクスの結果によって構成される．PTD のデータベースを $N+1$ 個の説明変数と 1 個の目的変数を持つバイナリ行列とみなすと，フィッシングサイトの検知は機械学習における分類問題の一種であると捉えられる．

そこで，エンドユーザにウェブサイトを閲覧させ PTD を作成し，PTD を用いた場合，用いない場合の比較検討を行った．機械学習手法には AdaBoost を利用した．理由の 1 つは，我々の先行研究 [3] において AdaBoost の性能が高かったためである．他の理由としては AdaBoost は，あるヒューリスティクスが正確に解答できなかったウェブサイトについて，正しく解答できた他のヒューリスティクスに高い重みを割り当てるという理論的背景があるためである．これにより，エンドユーザが間違えやすいフィッシングサイトを正しく判別できるヒューリスティクスに高い重みが割り当てられ，各エンドユーザの能力に応じた検知が行えるのではないかと期待した．なお，複数のエンドユーザが PTD を共有することは，プライシーの問題もあり本論文では考慮しない．

第 1 回の被験者実験は，2007 年 11 月に奈良先端科学技術大学院大学のインターネット工学講座に所属していた 10 名を対象として実験を行った．被験者らは全員 22 歳から 29 歳の男性で，3 人は過去 5 年以内に修士課程を卒業しており，残りは修士課程の学生であった．この実験はこれまでに行われてきた著名な被験者実験の手法を踏襲して行われ，被験者らは Windows XP 上で動作する Internet Explorer (IE) 6 を操作し，正規サイト 6 件，フィッシングサイト 14 件を閲覧した．被験者らの判断の平均の誤り率は 19.0%，既存のヒューリスティクスを AdaBoost で組み合わせた検知の誤り率が 20.0% であったのに対し，HumanBoost 方式の場合は誤り率が 13.4% と改善が見られたことを観測した．なお，特殊な設定として IE 6 を多国語ドメイン名を表示できるようにした事を除いては，OS やブラウザはインストールされたままの標準的な設定を用いた．

また，第 2 回の被験者実験として，2010 年 3 月に北陸先端科学技術大学院大学の篠田研究室に所属していた 11 名を対象として追試を行った．被験者らは全員 23 歳から 30 歳までの男性で，2 人が過去 5 年以内に修士課程を卒業しており，残りは修士課程の学生であった．この実験では被験者らにはブラウザを操作させず，IE 6 のスクリーンショットを紙に印刷したプリントを用いて判断させた．一部の正規サイトは第 1 回目の実験の時からデザインが変更されていた為，フィッシングサイトもそれらに合わせて調整を行った．被験者らの判別の平均誤り率は 40.5%，ヒューリスティクスによる組み合わせは 10.5% であったのに対し，HumanBoost 方式では 9.7% であった．

これらの被験者実験の共通の問題点としては，被験者の偏りが挙げられる．どちらの実験においても，被験者は少数であり，全員男性であり，情報工学分野の修士課程の学生または卒業生であった．また，被験者によっては PTD を利用せず，既存のヒューリスティクスのみによって判別を行う場合に誤り率が少なくなるケースも確認された．

2.2 大規模な被験者実験の概要

第 3 回目の被験者実験として，2010 年 7 月にインターネット調査企業に依頼して 309 人分の解答を採取した．この 309 人のうち，男性は 131 人で，女性は 178 人であった．職業は技術職の会社員が 48 人，事務職の会社員が 58 人，主婦が 61 人，学生が 18 人であった．

設問は，年齢や職業などといったユーザの基本的な属性に加え，ウェブサイトの利用経験の調査，エンドユーザの意思決定の根拠の調査，最後に 2.1 節で述べた HumanBoost 方式の実験の追試を目的とした調査を目的として設定した．なお，この実験に用いたウェブサイト群は論文 [1] を参照して頂きたい．

これまでの被験者実験では，被験者は利用経験の全くないウェブサイトについてもフィッシングサイトか否かの判断を下さねばならなかった．しかし，エンドユーザにとってウェブサイトに事前知識がある場合，ない場合において同様の判断が行えるとは考えがたい．そこで，実験で用いるウェブサイト群について利用経験の有無の調査を行った．

また，過去に行われた著名な被験者実験ではエンドユーザはウェブサイトの内容を意思決定の拠り所としていることが知られている．しかし，フィッシングサイトのページの内容は本物そっくりであり，ページの内容に頼った判断ではフィッシングサイトを正規サイトであると誤って判断する率が高くなると考えられる．そこで，被験者らにいくつかのウェブサイトのスクリーンショットを閲覧させ，それぞれ正規サイトかフィッシングサイトと思うかを判断させた．また，被験者らにはその判断の根拠が「ページの内容」「ウェブサイトの URL」「ブラウザの表示するセキュリティ情報」「その他」（その他の場合はその事由）の何であったかについて，1 つ以上の選択肢を解答させた．

2.3 被験者実験の解析

2.3.1 被験者のクラスタリング

実験結果から，このような選択基準がウェブサイトの判断に好影響をもたらすのか，あるいは悪影響をもたらすのかを調査した．利用経験がないと答えられたウェブサイトの平均誤り率は 48.6% であったのに対し，利用経験があるサイトでは 42.7% であることが観測された．これから，利用経験の有無は意思決定の結果に何らかの好影響を及ぼしていると推測し得る．

表 2: 判断基準と平均誤り率

ページの内容	URL	セキュリティ	誤り率
v			61.9 %
	v		25.5 %
		v	36.8 %
v	v		51.7 %
v		v	60.0 %
	v	v	17.9 %
v	v	v	49.8 %

表 3: 被験者グループのクラスタリング

クラスタ ID	被験者数	要素 1	要素 2	要素 3	要素 4	要素 5
1	48	0.827	0.766	0.787	0.543	0.787
2	30	0.576	0.580	0.479	0.835	0.240
3	61	0.591	0.405	0.674	0.047	0.965
4	17	0.753	0.953	0.953	0.039	0.992
5	71	0.411	0.160	0.168	0.058	0.950
6	11	0.561	0.000	0.014	0.000	1.000
7	34	0.124	0.006	0.008	0.000	1.000
8	37	0.441	0.290	0.177	0.421	0.822

次に，各項目と平均誤り率についての調査を行った．その結果を表 2 に示す．v は該当する選択肢が選択された事を示す．例えば，ページの内容によってのみ判断している被験者の平均誤り率は 62.1% であった．表 2 からは，ページの内容のみによって判断している場合は誤り率が高く，ウェブサイトの URL とブラウザの表示するセキュリティ情報をみて判断している場合に誤り率が少ない．従って，ページの内容に頼って判断することは意思決定の結果に何らかの悪影響を及ぼしており，ウェブサイトの URL とセキュリティ情報をみて判断することは好影響を及ぼしていると考えられる．なお，選択肢として「その他」を定義したが，選択される頻度が少なかったため本論文では分析の対象外とする．

そこで，被験者らがフィッシングサイトを判断するための能力の構成要素は，以下の 5 項目であると仮定する．また，論文 [1] に示す方式で，これらの各構成要素について，$(0 \cdots 1)$ に正規化された範囲内における数値化を試みた．

要素 1 過去に利用経験のあるウェブサイトについては，この経験を活かした判断を行うこと．

要素 2 ページの内容に基づいた判断を行っていないこと．

要素 3 ウェブサイトの URL に基づいた検知を行っていること．

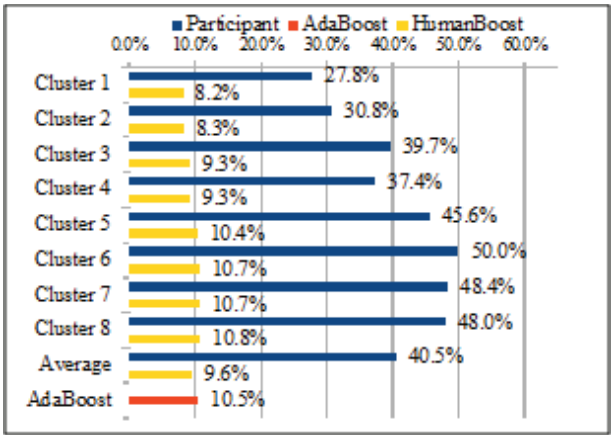

図 1: 各クラスタの誤り率

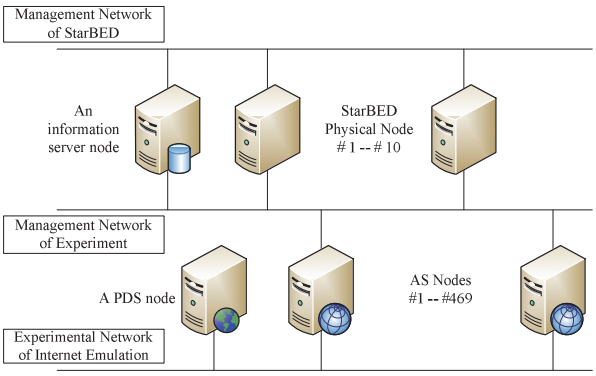

図 2: 実験トポロジー

要素 4 SSL (EV SSL) を利用しているサイトでは，ブラウザの表示するセキュリティ情報に基づいた判断を行っていること．

要素 5 SSL (EV SSL) を利用していないサイトでは，ブラウザの表示するセキュリティ情報に基づいた判断を行っていないこと．

次にこれらの数値化された 5 個の構成要素に基づいたエンドユーザのクラスタリングを行う．本研究では，クラスタリング手法の選定として，代表的なクラスタリングのアルゴリズムである EM 法を用いた．なお，クラスタ数はベイズ情報量規準を用いて探索した結果から 8 クラスタとした．

クラスタリング結果を表 3 に示す．クラスタ ID は便宜上付与した名前であり，それぞれ 8 個のクラスタを意味している．被験者数は各クラスに分類された被験者数である．要素 1〜5 の数字は，数値化された被験者の能力の構成要素について，クラスタ毎に平均を計算した値である．例えばクラスタ 1 のユーザは，利用経験のあるサイトを正しく判定でき，コンテンツの内容より URL やブラウザのセキュリティ情報に重きをおいて判断していることがわかる．

各クラスタごとの平均誤り率を図 1 に示す．青色のグラフが被験者クラスタ毎の判別の誤り率の平均，赤色のグラフがヒューリスティクスによる誤り率の平均，黄色のグラフが HumanBoost 方式を用いた場合の，各被験者クラスタにおける誤り率の平均である．既存のヒューリスティクスによる誤り率は 10.5%，各被験者の誤り率は 40.5%，HumanBoost による誤り率は 9.6% であった．クラスタ 1〜5 に属するユーザの誤り率は，既存のヒューリスティクスによる誤り率より低いという観測結果から，これらのユーザは PTD を活用すべきユーザであると考えられる．反対に，クラスタ 6〜8 に属するユーザの誤り率は，既存のヒューリスティクスより高くなってしまっている．これは，これらのユーザの PTD は活用せず，ヒューリスティクスに任せた検知を行うべきだと推測される．

最も平均誤り率が低かったクラスタは 1 であった．この結果から，HumanBoost 方式で活用すべき PTD を持つユーザは,「事前知識を検知に役立てることができる」かつ「ページの内容に頼った判断を行っていない」「ウェブサイトの URL に基づいた検知を行える」事ができ，また「ブラウザの表示するセキュリティ情報を注目できる」といった能力を持つ被験者であると考察できる．

3 フィッシングサイトの模倣インターネット上での再現

本研究ではインターネットにおけるトポロジをテストベッド上に再現する技術である模倣インターネットに着目する．模倣インターネットは，インターネット規模のアプリケーションやプロトコルを検証する目的で開発されたものである．このような技術は，実際にインターネットにおいて有効性を試すことが理想ではあるが，インターネットは様々な組織によって独立して運用されており，検証のための操作や観測を行うことは極めて困難である．このため，テストベッド空間に

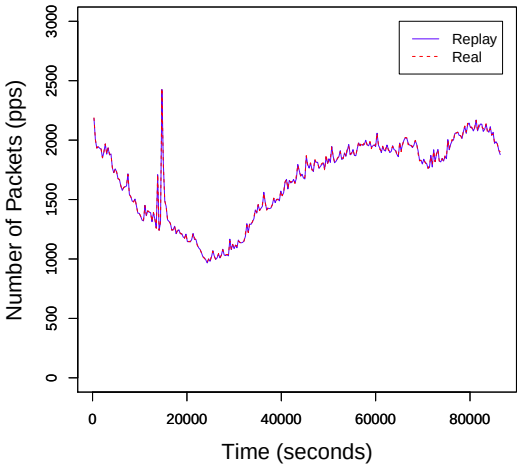

図 3: 再現された単位時間あたりのパケット数と，データセットの値との比較

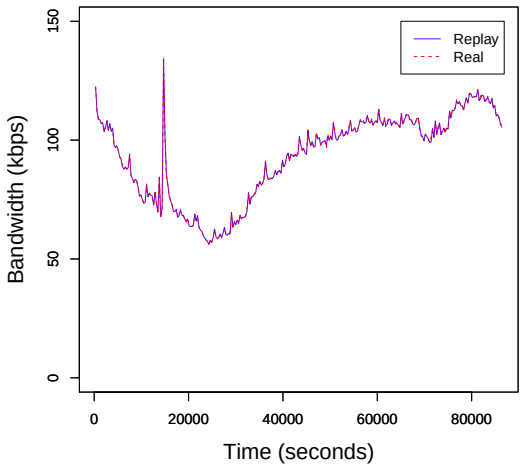

図 4: 再現された単位時間あたりのトラフィック流量と，データセットの値との比較

インターネットを再現するというアプローチによって，アプリケーションやプロトコルの検証が進められている．この技術を活用し，本研究では，取得したフィッシングサイトを模倣インターネット上に再現するシステムを実現する．これにより，対策システムをより現実的な環境で試験ができ，有効性，普遍性の検証を目指している．

2011 度は模倣インターネットを用いたテストベッドの作成及びフィッシングサイトの再現を行った．2012 年度はフィッシング攻撃以外のサイバー攻撃を同じテストベッド上に発生させ，その二つを協調して検知するような対策技術の方式研究を行った．なお，成果の一部は WIDE 5 月研究会においてポスター展示として発表した．

3.1 DoS 攻撃対策システムとの連携

ヒューリスティクスに基づいたフィッシングサイトの検知システムは，フィッシングサイトらしさを算出し，この値としきい値を比較する．例えば，フィッシングサイトらしさを 0〜1 の範囲で算出するアルゴリズムを考える．1 はフィッシングサイト，0 は正規サイトを示し，フィッシングサイトらしさの値が 0.5 以上か否かでフィッシングサイトであるかを判断する．

ここで，算出されたフィッシングサイトらしさの値が，しきい値に近い 0.4 のような値であった場合を考える．当然，しきい値を下回るため正規サイトと判定するのであるが，仮にこのサイトがボットネット上にホスティングされていた場合はどうであろう．正規のサイトであればボットネットではなく著名なクラウド・データセンターを利用していることが多いと考えられうる．このような情報を用いれば，フィッシングサイトらしさの判断をより高度化できる可能性があるのではなかろうか．0.4 というフィッシングサイトらしさの値を 0.5 以上に押し上げるだけの要因となりえるのではないか．

そこでボットネットを介して行われる代表的なサイバー攻撃であるサービス運用妨害 (Denial of Service, DoS) と，フィッシングサイトの検知技術の協調した解析を考える．DoS 攻撃では発信源の IP アドレスが偽装されることが多く，発信源の探索には IP トレースバック技術が必要となる．トレースバック技術は，本研究と同じく模倣インターネットの環境で実験されており，フィッシングサイトと DoS 攻撃を同時に再現することは可能であると考える．

3.2 DoS 攻撃対策システムの評価環境

そこで，フィッシングサイトと DoS を同時に再現し，評価をするテスト基盤を研究する．まず，図 2 に

示す構成を用い，日本国内の 469 AS を再現する模倣インターネットを作成する．模倣インターネットの利用，フィッシングサイトの再現に関しては Deep Space One / Nerdbox Freaks Working Group とも連携して行なっており，2011 年度の報告書を参照していただきたい．

次に，模倣インターネットにおけるトラフィックの再現を考える．テストベッド上で DoS 攻撃を再現し，検知する技術を評価する際には，DoS 攻撃だけでなく，正常状態の通信も再現できる必要がある．DoS 攻撃はパケットを大量に生成するだけで簡単に再現できるが，正常状態の通信の再現は難しい．

そこで，正常状態の通信と思われるトラフィックデータセットを入手し，これを AS 単位に分割した上で，模倣インターネットの各 AS から協調してトラフィックを送出するプログラムを実装した．なお，今回使用したデータセットは全て DNS トラフィックであり，UDP データグラムによって構成されているという特徴があった．また，再現されたトラフィックに起因するであろう ICMP Port Unreachable メッセージを抑制するよう，模倣インターネットの各ノードを設定した．

トラフィックの再現結果を図 3， 4 に示す．横軸はトラフィックが観測された時間 (秒)，縦軸はそれぞれパケットの数 (個数) と使用した帯域 (Kbps) である．赤線は実インターネット上で観測されたトラフィック，青線は模倣インターネットで再現したトラフィックを示す．線がほぼ重なっているため，紫色で表示されている．念のため時系列回帰を行ったところ，どちらの場合も時差が 0 の場合に最も高い相関係数が観測された．これらの結果から，本データセットに関しては正常状態のトラフィックを模倣インターネット上で再現できていると考えられる．

本研究で開発したソフトウェアは，トラフィックを各 AS 単位に分離するもの，分離されたトラフィックを送信するものの 2 種類から構成される予定であった．ところが，送信を担うプログラムについて，送信時間の制御がうまくいかない不具合があった．例えば，n 秒と m 秒 (m >n) において観測されたトラフィックを再現するとき，n 秒にトラフィックを送信，$(m-n)$ 秒間ウェイトしてから，m 秒に観測されたトラフィックを送信，というような処理となる．このウェイトが想定通りに働かないという不具合であった．実験では著名なトラフィック再生ツールである `tcpreplay` を用いることによって解決したが，このツールでは,「トラフィックを 2 倍速で再現する（待ち時間を半分にする）」など，テストベッドの特性を用いた実験を行うことが難しい．実験の要求項目を整理した上で，このウェイト時間問題の解決を試みることは今後の課題である．

4 おわりに

本年度の SWAN Working Group の活動は，悪性ウェブサイトの１つであるフィッシングサイトに焦点を当て，また WIDE の幅広い人脈を活かし，テストベッド等の研究開発やトラフィック観測・分析との連携を行った．来年度も幅広い種類の悪性ウェブサイトについて，多面的な解析を行うことが課題である．

参考文献

[1] Daisuke Miyamoto, Hiroaki Hazeyama, Youki Kadobayashi, Takeshi Takahashi, *"Behind HumanBoost: Analysis of Users' Trust Decision Patterns for Identifying Fraudulent Websites,"*, *Journal of Intelligent Learning Systems and Applications*, (Accepted).

[2] Daisuke Miyamoto, Hiroaki Hazeyama, Youki Kadobayashi, *"HumanBoost: Utilization of Users' Past Trust Decision for Identifying Fraudulent Websites,"*, *Journal of Intelligent Learning Systems and Applications*, Vol. 2, No. 4, pp.190-199, December 2010.

[3] Daisuke Miyamoto, Hiroaki Hazeyama, Youki Kadobayashi, *"An Evaluation of Machine Learning-based Methods for Detection of Phishing Sites,"*, *Australian Journal of Intelligent Information Processing Systems*, Vol. 10, No. 2, pp.54-63, November 2008.