早稲田大学大学院 基幹理工学研究科

博士論文審査報告書

論　文　題　目

Discovering the Hidden Cyber Attacks: Machine Learning Based Approaches

伏在するサイバー攻撃の発見: 機械学習によるアプローチ

申　請　者

Bo　SUN

孫　博

情報理工・情報通信専攻 ネットワークシステム研究

2018 年 2 月

サイバー攻撃を行う攻撃者には、攻撃の成功率を高める動機がある。一方で防御側の立場からすれば、攻撃の成功率を低めることに目標がある。これらの相反する要求の結果、攻撃者は防御側にとって検知がきわめて困難となるように攻撃手法を変化させてきた。すなわち攻撃者は、通常の通信と攻撃通信との判別が困難となるように、攻撃に関連するデータや痕跡を巧妙に偽装する。このような巧妙な偽装が行われた場合、サービスの利用者が攻撃に対処することはもとより、攻撃の存在に気が付くことすらできなくなる。検知を回避することを意図して巧妙に作成された攻撃を、本論文では「伏在するサイバー攻撃」と呼ぶ。伏在するサイバー攻撃は様々なサービスに遍在し、利用者やサービス提供者に甚大な被害をもたらすリスクがある。伏在するサイバー攻撃を早期に検出し、攻撃を未然に防ぐことは、重要なセキュリティ課題のひとつである。

本論文は、利用者が多いインターネット上のサービスである（１）ウェブサービスおよび（２）モバイルアプリ配布サービスに着目し、これらのサービスにおける「伏在するサイバー攻撃」を効率的に検出する技術を提案し、その技術の有効性評価を行った結果を報告している。これらのサービスはきわめて多数の利用者から構成されるため、セキュリティ上の脅威が高く、実用上の意義は高い。これらのサービスが扱う情報量は膨大であるため、伏在するサイバー攻撃を人手で検知することは不可能である。したがって、自動化された技術が必要である。自動化を実現するための手段として、機械学習技術を活用している点に本研究の特徴がある。

第１章では、本論文の背景となる情報、研究の目的、および全体的な貢献を述べている。

第２章では、ウェブサービスに伏在するサイバー攻撃を検出する技術を提案している。ウェブサイトの中には、アクセスするだけで利用者端末にマルウェアをインストールする悪性サイトが存在する。それらの悪性サイトのアドレスを予め知ることができれば、利用者がそのようなサイトにアクセスすることを未然に防ぐことができる。しかしながらウェブサイトのアドレスは無数に存在し、かつアドレスは動的に生成されることが多い。したがってウェブサイトのアドレスが悪性であるか、良性であるかをアドレスそのものから得られる情報、例えばドメインやファイルパスなどによって判定することは困難である。さらに攻撃者はパターンファイルによる検出を逃れるために、ウェブサイトごとに異なる特徴を有するようなしかけを用意している。すなわち悪性サイトはウェブ空間内で伏在している。

本論文はこのような問題に対処するために、（１）無数のアドレス空間から悪性サイトである確率が高いアドレスを探索し、（２）任意に指定することが可能な悪性ウェブサイト群と類似した特徴を持つアドレスを検索する技術を提案している。アドレスの探索においてはドメイン名とＩＰアドレスの関係をもとに悪性サイトが潜んでいる可能性が高い空間に絞ってアドレスを収集

する。アドレスの検索では様々な悪性サイトに固有な特徴を集中的に学習する。検索の手順は、はじめに同様な特徴を持つ悪性サイト、例えば同一のエクスプロイトキットから生成された可能性が高いサイトの集合を用意し、次に検査対象のアドレスが前記集合に属する確率を評価する。あるアドレスに対して高い確率を得たら、そのアドレスを悪性サイトとして判定する。このように検索方法としてオンデマンド・クラスタリングのアプローチをとることにより、様々な組み合わせの悪性サイト群に類似した悪性サイトを検索することができる。すなわち単純に十把一絡げで悪性・良性の判定を試みるのではなく、数ある悪性サイトがもつ固有な性質のそれぞれに着目し、個々の特徴を有する確率が高い悪性サイトを抽出することに特徴がある。実データを用いた性能評価の結果、同一のエクスプロイトキットを用いて作成された悪性ウェブサイトやフィッシングサイトを従来方法と比較して高精度に検索できること、およびアドレス空間の探索と検索に要する時間は従来方法とくらべて大幅に削減できることを実証している。

第３章では、モバイルアプリ配布サービスに伏在するサイバー攻撃を検出する技術を提案している。スマートフォンやタブレット等のモバイル端末では、マーケットと呼ばれるプラットフォームを通じてアプリを配布するモデルが広く採用されている。このようなモバイルアプリ配布プラットフォームでは、開発者がマーケットにアップロードした個々のアプリに対し、利用者によるレーティングや、レビューコメント等の評判情報を公開することが多い。このようなサイトにおいて、攻撃者は悪性アプリ（マルウェア）をアップロードし､かつそのアプリに対して偽の評判情報を投稿することができる。攻撃者がレビューコメントに工夫をすれば、偽の評判情報と通常の評判情報を区別することは容易ではない。偽の情報に騙された利用者は、自身の端末にマルウェアをインストールするリスクがある。

本論文ではこのような伏在する悪性の偽情報を検出するために、自然言語処理と機械学習を組み合わせたアプローチをとる。はじめにマルウェアに対してのみレビューを投稿したユーザを悪性ユーザとして抽出する。次にそれらの悪性ユーザがつけたレーティングおよび投稿したレビューコメントの特徴を学習し、それらの評判情報と近い特徴を持つコメントを偽情報候補として抽出する。最後に、そのような偽情報と考えられるコメントが多数投稿されたアプリを解析し、悪性アプリであるかを判定する。レーティング情報からは平均的なレーティング値からの外れ度に基づくスコアを特徴とする。レビューコメントが持つ特徴として、文字数や N-gram などの量的な特徴、および意味解析、感情分析による質的な特徴を取り入れている。機械学習として複数の教師あり機械学習アルゴリズムを用い、最も性能が高かったものを採用している。手動でラベル付をした約 1,700 件の評判情報データを用いて提案手法の性能を評価した結果、真陽性率が 90%、偽陽性率が 5.8%と実用的な精度を達成できることを示している。つぎに約 100 万のアプリに対して約

1,400 万人のユニークなユーザが投稿した、約 5,700 万件の評判情報データに対して、前記で訓練した識別モデルを適用し、攻撃の特徴や規模を調査している。調査の結果、約 289,000 の潜在的な攻撃アカウントが検出された。これらの攻撃アカウントが評判情報を投稿したアプリの約 15%は悪性アプリであった。潜在的な攻撃アカウントその中でも特に悪性度が高いと考えられる 1,000 のアカウントを解析した結果、136 のアカウントが悪性アプリにのみ評判情報を投稿していること、また、これらの攻撃者がコメントを投稿したアプリの中には、アンチウィルスソフトが検出できかった悪性アプリが多数含まれていた。すなわち、提案手法によって効率的に伏在する偽評判情報、およびそれに紐づく悪性アプリを検出することができる。

第４章では、本論文の制限事項や課題、および研究の発展性について論じている。

第５章では、伏在するサイバー攻撃に対して本研究が提示した２つの方法を俯瞰し、それらのまとめと本論文の結論を述べている。

以上を要するに、本論文は防御者による検出から逃れる巧妙な「伏在するサイバー攻撃」の問題に焦点を当て、これらの攻撃を検出するための具体的な技術の提案と、その有効性評価を行っている。本論文で検討対象としたウェブおよびモバイルアプリプラットフォームは、いずれも非常に多くの利用者を擁するサービスであり、セキュリティ脅威によるインパクトが高い。本研究の成果は、セキュリティオペレータが有効に活用できるデータ解析支援ツールを提供するものであり、実用的にも価値がある。本論文で扱った２つの課題に共通するアプローチとして、機械学習技術の適用が挙げられる。それぞれの問題に応じた特徴の構成と、システムとしての実装方法に学術研究としての特徴と新規性がある。本論文が提案したアイディアや考え方は、他のサービスにおける伏在するサイバー攻撃への対策にも有効であり、この研究領域のさらなる発展が期待できる。よって、本論文は博士（工学）早稲田大学の学位論文として価値あるものと認める。

2018 年 2 月

審査委員

主査　早稲田大学准教授　　博士（情報科学）（早稲田大学）森　達哉

早稲田大学教授　　工学博士（東京大学）　　後藤　滋樹

横浜国立大学准教授　博士（工学）（横浜国立大学）　吉岡　克成