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

効率的な解析を目的とした
自動マルウェア分類に関する研究

Automatic malware classification for efficient analysis

申　請　者

岩村	誠
Makoto	IWAMURA

情報理工学専攻　情報構造研究

2012 年 2 月

近年、機密情報等の漏えいやサービス妨害攻撃等のセキュリティ侵害の背後で基盤ツールとして暗躍するマルウェア（Malware）が社会問題化している。マルウェアの種類数は増加の一途を辿り、マルウェアが及ぼす脅威を解明することが困難になっている。こうした事情を鑑み本研究では、多数のマルウェアを効率的に解析する仕組みを構築することを目的とする。これにより、マルウェアが備える脅威の全容解明を可能にし、マルウェア駆除ツールの作成やネットワークでの攻撃遮断といった対処を促進することを目指す。

本研究は、次の大きく二つの取り組みから構成される。一つ目は、マルウェアの解析に要する作業を自動化する取り組みであり、二つ目は、マルウェアが備える脅威の全体像を効率よく把握する取り組みである。

一つ目の取り組みでは、種々の方法でコード隠ぺいがなされているマルウェアを解析するためのアンパックならびに逆アセンブル技術の構築、ならびに機能把握の要となるインポートアドレステーブル（以下、IAT）に関して、その格納場所を特定する手法を提案した。実験では、5 種類のマルウェアに関してインポートアドレステーブルの格納場所を予測し、提案手法と従来技術の予測精度を比較した。その結果、提案手法の MCC（Mathews Correlation Coefficient）は 98.4%～100%を示し、従来技術と比較し安定して優れた予測精度であることを明らかにした。

二つ目の取り組みでは、プログラムコードに基づく自動マルウェア分類システムを新たに提案・構築した。これにより、同じプログラムコードの断片を共有するマルウェアを発見するとともに、優先して解析すべきマルウェアを選定することが可能になる。インターネットで収集されたマルウェアを分類した実験では、代表的な 5 つのクラスタから 1 検体ずつを選択して解析するだけで、全マルウェアの約 77.5%のプログラムコードを把握できることを明らかにした。

本論文は 7 章からなる。以下に各章の概要を述べ評価を加える。

第 1 章では本研究の目的並びに概要を述べている。

第 2 章はマルウェア解析のためのアンパック手法を提案している。多くのマルウェアは、ランタイムパッカーと呼ばれる一種の難読化ツールにより、そのプログラムコードが隠蔽（以下、パック）されている。このため、プログラムコードに基づいてマルウェアを分類するには、隠蔽されたマルウェアのプログラムコード部分を抽出（以下、アンパック）する必要がある。本章では、従来のアンパック手法における二つの課題を指摘し、それらを解決する手法を提案・評価した。従来技術では、アンパックされたプログラムコードのエントリポイント（以下、OEP: Original Entry Point）を特定することに留まってお

り、アンパックされたプログラムコードの始点・終点を決定できなかった。これに対し本研究では相対分岐命令の分岐元と分岐先が、複数の実行モジュール間を跨がないことに着目し、実行モジュールの境界を推定する手法を提案した。実験では、 対象となるプログラムコード領域の前後に、他のプログラムコード領域が接している場合であっても、OEP を含むプログラムコード領域だけを識別できることを示した。二つ目の課題は、マルウェアが多重にパックされている場合に、各層のアンパックが完了するたびに、オリジナルコードの候補が抽出されてしまう点にある。これに対し本研究では、得られたオリジナルコードの候補に関して、隠れマルコフモデルに基づく確率モデルにより、コンパイラ出力コードの尤もらしさを算出し、オリジナルコードを特定できる新たなアンパック手法を提案した。実験では、従来技術の方式により抽出された約 230 のオリジナルコードの候補に関して、真のオリジナルコードを正確に特定可能なことを示した。

第３章では、こうして得られたバイト列を逆アセンブルする技術を開発した。一般的に、デバッグシンボル情報等の入手が困難なマルウェアに関して、正確な逆アセンブル結果を得ることは難しい。多くのマルウェアは、通常のソフトウェアと同様、迅速なバグ改修や機能追加のために、よく知られたコンパイラが用いられる。そこでここでは、隠れマルコフモデルに基づく確率的逆アセンブル手法を提案した。本手法は、よく利用されるコンパイラが出力する実行ファイルの傾向（機械語命令・データにおける各バイト値の出現確率等）を学習することで、正確な逆アセンブル結果を得ることを可能にする。これにより従来技術の MCC では逆アセンブルの精度が 90～91%程度となる一方で、提案手法では安定して 99%以上の結果が得られた。

第４章では、マルウェアの逆アセンブル結果をもとに、マルウェア間の類似度を算出する手法を提案した。従来研究では、ベーシックブロックやコールツリー等、プログラム構造を手動で再構築する必要があり、これがマルウェア分類の全自動化の妨げとなっていた。また N-gram/N-perm による手法では、一種の統計情報により類似度を定義しているため、実際に変化のあった場所を抽出することは難しいといった問題もあった。こうした問題に対し、本研究では機械語命令単位の LCS（Longest Common Subsequence）を抽出し、その LCS の長さに基づき類似度を決定する手法を提案した。本手法が必要とするのは逆アセンブル結果のみであり、容易にマルウェア分類作業を自動化することができる。さらには、提案手法により算出された類似度は機械語命令単位の LCS であるため、解析に要する作業量（読むべき機械語命令数）を正確に見積もることも可能になる。ただ、マルウェアの中には、機械語命令数が 100,000 を超えるものも存在し、単純に機械語命令列同士の LCS を抽出するには多くの計

算時間を要する。このため提案手法では、機械語命令を独自の縮約命令で表現することで、LCS 抽出アルゴリズムのビットベクトル化を可能にした。これにより、SSE2 命令を用いた実装では、単純な LCS 抽出アルゴリズムと比較し 100 倍程度の高速化を達成した。

第５章では、前述のアンパック・逆アセンブル・類似度算出に関する提案手法を組み合わせることで、自動マルウェア分類システムを構築した。実際のインターネットで収集されたマルウェアに対する実験では、代表的な５つのクラスタから１ 検体ずつを選択し解析するだけで、全マルウェアの約 77.5% のプログラムコードを把握できることを明らかにした。さらにソースコードが存在するマルウェアを用いた実験では、コンパイラや最適化オプションが同じであれば、ソースコードの類似度と同じ相関関係を維持できていることが分かった。

第６章では、マルウェアの機能を把握するために要となる IAT エントリ格納場所の特定方法を提案した。網羅的に分岐命令の候補を抽出する従来技術は、実行モジュールの再配置による錯乱手法に対して弱い。また逆アセンブル手法に基づく従来技術は、逆アセンブル結果の不正確さが IAT エントリ格納場所の特定にも悪影響を与えていた。そこで本研究では、実行モジュール内の各バイト値が機械語命令である確率と、IAT エントリを根とするコールツリーを用いることで、IAT エントリ格納場所を精度よく抽出する手法を提案した。実験では、提案手法が各種従来技術よりも高い精度で IAT エントリ格納場所を特定できることを示した。

第７章はまとめと今後の課題について言及している。

以上を要するに、本論文ではマルウェア解析の自動化を図ることにより迅速かつ高速な解読・分類技術を実現し、これにより、安全・安心な IT 社会の構築に大きな貢献をした。よって本論文は、セキュリティ分野及びソフトウェア工学分野の進展に多大なる貢献をしたものとみなすことができ、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。

2012 年 2 月
審査員

主査	早稲田大学教授	Ph.D.（イリノイ大学）	村岡洋一
	早稲田大学教授	工学博士（慶應義塾大学）	中島達夫
	早稲田大学教授	工学博士（早稲田大学）	
		Ph.D.（ｽﾀﾝﾌｫｰﾄﾞ大学）	松山泰男
	早稲田大学教授		筧捷彦