早稲田大学大学院 基幹理工学研究科

博 士 論 文 概 要

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

2011 年 12 月

近年，機密情報の漏えいやサービス妨害攻撃等のセキュリティ侵害が，その背後で基盤ツールとして暗躍するマルウェア（Malicious Softwareの混成語）とともに社会問題化している．加えて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はおろか，流行のマルウェアを把握することも困難になっている．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，次の二つの取り組みから構成される．一つ目は，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．また二つ目は，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る．

一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．

また，二つ目の取り組みでは，マルウェアの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．

以下では，各取り組みにおける背景および提案手法ついて本論文の章立てに沿って概説する．まず，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の構築において，その要素技術となるアンパック（第2章）・逆アセンブル（第3章）・プログラムコードの類似度算出（第4章）について説明する．

昨今の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，まず隠蔽されたマルウェアのプログラムコードを抽出（以下，アンパック）する必要がある．本論文の第2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．従来技術における一つ目の課題は，従来技術の目的が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・終点を決定できないことにある．この課題に対し本研究では，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．これにより，OEPを含む一つの実行モジュール，つまりアンパックされたプログラムコードの領域を決定することを可能にした．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．従来のアンパック手法における二つ目の課題は，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．

こうして得られるマルウェアのアンパック結果はバイト列として表現される．マルウェアの解析作業では，このバイト列を逆アセンブルし，その機能を明らかにしていく．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（機械語命令とデータのラベル付）結果を得ることは難しい．一方で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．そこで第3章では,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．一つ目の実験では，Visual C++でコンパイルされた8つのアプリケーションに関して，提案手法と従来技術との比較を行った．その結果，コンパイラオプションによっては従来技術のMCCが90～91%程度となる一方で，提案手法では安定して99%以上の精度を示していることを明らかにした．また，高度に難読化された10種類の実行バイナリに関しても，従来技術のMCCが約35.2%～48.7%であったのに対し，提案手法は約91.1%と非常に高い精度を達成した．

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．また，本システムが同一と判断したマルウェアに関して，アンチウィルスソフトでは異なる複数の検出名が確認される状況もあり，マルウェアに対する命名の難しさが明らかになった．他にも，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大小関係を維持できていることが分かった．一方，同じソースコードのマルウェアであっても，コンパイラや最適化オプションが異なる状況では，種の違いよりも大きくその類似度が低下することも確認された．ただ，マルウェア開発によく使われるコンパイラやそのオプションの種類数には限りがある．このため，マルウェア作者がコンパイラやそのオプションを変化させながらマルウェアを作成したとしても，依然として，本提案システムは解析コストの削減に効果を発揮すると考えられる．

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．

個人情報の売買やスパムメール配信代行といったブラックマーケットの成熟に伴い，マルウェアにも迅速なバグ改修や機能追加が求められるようになってきた．維持管理が容易な開発環境が利用され，また多くのプログラムコードが再利用されているのも，マルウェア開発に対するコスト意識が高まった結果であろう．こうしたマルウェア開発の現状に則し，マルウェアの分類・解析作業の自動化を実現した本研究は，日々増加するマルウェアに対しても，その脅威の全容解明に大きく貢献するであろう．

早稲田大学　博士（工学）　学位申請　研究業績書

岩村　誠　　印

（2011 年 11 月　現在）

種 類 別	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む）
論文　○	題目　Towards Efficient Analysis for Malware in the Wild 掲載紙名　Proceedings of IEEE International Conference on Communications 2011 発表年月　2011 年 6 月 著者　Makoto Iwamura, Mitsutaka Itoh, Yoichi Muraoka
○	題目　機械語命令列の類似性に基づく自動マルウェア分類システム（推薦論文） 掲載紙名　情報処理学会論文誌 Vol.51, No.9, pp.1622--1632 発表年月　2010 年 9 月 著者　岩村誠,伊藤光恭,村岡洋一
総説	題目　Anti-Malware Technologies 掲載紙名　NTT Technical Review, Vol.8, No.7 発表年月　2010 年 7 月 著者　Mitsutaka Itoh, Takeo Hariu,Naoto Tanimoto, Makoto Iwamura, 他 5 名
	題目　マルウェア対策技術 掲載紙名　NTT 技術ジャーナル,2010 年 3 月号 発表年月　2010 年 3 月 著者　伊藤光恭,針生剛男,谷本直人,岩村誠, 他 5 名
○	題目　研究用データセット：マルウェア検体編　機械語命令列の類似性に基づく自動マルウェア分類システム 掲載紙名　情報処理学会会誌 Vol.51, No.3, pp.292--295 発表年月　2010 年 3 月 著者　岩村誠,伊藤光恭
講演　○	題目　IAT エントリ格納場所の特定方法 発表箇所　マルウェア対策研究人材育成ワークショップ 2011 発表年月　2011 年 10 月 著者　岩村誠,川古谷裕平,針生剛男
○	題目　マルウェアのエントリポイント検出後におけるコード領域識別手法（インターネットアーキテクチャ研究会学生研究奨励賞受賞） 発表箇所　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 発表年月　2010 年 6 月 著者　岩村誠,伊藤光恭,村岡洋一
	題目　次世代ネットワークを変容させるネットワークセキュリティ技術：機械語命令列の類似性に基づく自動マルウェア分類システム（招待講演） 発表箇所　情報処理学会創立 50 周年記念（第 72 回）全国大会 発表年月　2010 年 3 月 著者　岩村誠

早稲田大学　博士（工学）　学位申請　研究業績書

種 類 別	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む）
講演　○	題目　機械語命令列の類似性に基づく自動マルウェア分類システム 発表箇所　マルウェア対策研究人材育成ワークショップ 2009 発表年月　2009 年 10 月 著者　岩村誠,伊藤光恭,村岡洋一
○	題目　コンパイラ出力コードモデルの尤度に基づくアンパッキング手法 発表箇所　マルウェア対策研究人材育成ワークショップ 2008 発表年月　2008 年 10 月 著者　岩村誠,伊藤光恭,村岡洋一
○	題目　隠れマルコフモデルに基づく新規逆アセンブル手法 発表箇所　電子情報通信学会総合大会 発表年月　2008 年 3 月 著者　岩村誠,伊藤光恭,村岡洋一
著書	題目　アナライジング・マルウェア――フリーツールを使った感染事案対処 出版元　オライリー・ジャパン 出版年月　2010 年 12 月 著者　新井悠,岩村誠,川古谷裕平,青木一史,星澤裕二
その他 （論文）	題目　Design and Implementation of High Interaction Client Honeypot for Drive-by-download Attacks 掲載紙名　IEICE Transactions on Communication, Vol.E93-B No.5 pp.1131--1139 発表年月　2010 年 5 月 著者　Mitsuaki Akiyama, Kazufumi Aoki, Yuhei Kawakoya, Makoto Iwamura, Mitsuataka Itoh
	題目　Memory behavior-based automatic malware unpacking in stealth debugging environment 掲載紙名　Proceedings of 5th IEEE International Conference on Malicious and Unwanted Software 発表年月　2010 年 10 月 著者　Yuhei Kawakoya, Makoto Iwamura, Mitsutaka Itoh
	題目　能動的攻撃と受動的攻撃に関する調査及び考察 掲載紙名　情報処理学会論文誌 Vol.50, No.9, pp. 2147- 2162 発表年月　2009 年 9 月 著者　青木一史,川古谷裕平,秋山満昭,岩村誠,針生剛男,伊藤光恭
（講演）	題目　実行命令トレースに基づく動的パッカー特定手法 発表箇所　マルウェア対策研究人材育成ワークショップ 2011 発表年月　2011 年 10 月 著者　川古谷裕平,岩村誠,針生剛男

早稲田大学　博士（工学）　学位申請　研究業績書

種 類 別	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む）
その他 （講演）	題目　Controlling Malware HTTP Communications in Dynamic Analysis System using Search Engine 発表箇所　The 3rd IEEE International Workshop on Cyberspace Safety and Security 発表年月　2011 年 9 月 著者　Kazufumi Aoki, Takeshi Yagi, Makoto Iwamura, Mitsutaka Itoh 題目　Dense Ship:サーバ型ハニーポット用仮想マシンモニタ 発表箇所　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 発表年月　2011 年 6 月 著者　川古谷裕平,岩村誠,伊藤光恭 題目　メモリ拡張によるアドレスに依存しないブレークポイント技術の提案 発表箇所　マルウェア対策研究人材育成ワークショップ 2010 発表年月　2010 年 10 月 著者　中山心太,青木一史,川古谷裕平,岩村誠,伊藤光恭 題目　動的解析における検体動作時間に関する検討 発表箇所　マルウェア対策研究人材育成ワークショップ 2010 発表年月　2010 年 10 月 著者　青木一史,川古谷裕平,岩村誠,伊藤光恭 題目　検索エンジンによるマルウェア接続先評価手法の提案 発表箇所　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 発表年月　2010 年 6 月 著者　青木一史,秋山満昭,岩村誠,伊藤光恭 題目　Gumblar の長期観測による分析 発表箇所　電子情報通信学会,インターネットアーキテクチャ研究会 発表年月　2010 年 6 月 著者　秋山満昭,佐藤一道,岩村誠,伊藤光恭 題目　OEP 自動検出によるマルウェアアンパック手法 発表箇所　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 発表年月　2010 年 6 月 著者　川古谷裕平,岩村誠,伊藤光恭 その他 14 件
（特許）	登録済　8 件（第 4091528 号,第 4253215 号,第 4358648 号,第 4551316 号,第 4643201 号,第 4709160 号,第 4739962 号,第 4755658 号） 出願中　4 件(特開 2009-193161,特開 2010-092179,特開 2011-086147,特開 2011-154727)