Vol.2011-DPS-146 No.52
Vol.2011-CSEC-52 No.52
2011/3/11

類似度の高いファイル構造に基づくマルウェア情報提供システムの提案

上井恭輔[†]　寺田真敏[††]　趙晋輝[†]

概要：本稿では，ファイル構造が類似するファイルの情報提供を目的としたマルウェア情報提供システムを提案する．提案システムはファイル内容の比較を行えるファジーハッシュのひとつである ssdeep を用いることで，入力された ssdeep の値に対して，ファイル構造が同じファイルだけでなく，構造が類似するファイル一覧を出力する．出力されたファイル一覧は，検査対象のファイルがマルウェアであるかどうかの目安としたり，新種や亜種のマルウェア解析に役立てたりすることができる．

A Proposal of Malware Information Support System using Similarity of File Structure

Kyosuke Kamii[†]　Masato Terada[†]　Jinhui Chao[†]

In this paper, we propose Malware Information Support System for preventing leakage of privacy information in the file for analyzed and supplying other file information whose structure is similar to the file's structure for analysis. This system use ssdeep for comparison of file content, prevent file information using similarity of file structure with ssdeep value user sent. A user can make use of the file information to know whether the file is malware. Also, security researcher can make good use of that for analysis of unknown malware.

1. はじめに

インターネットが社会のインフラとして普及する一方，マルウェアによる被害が増加している．数年前まではマルウェアの作者達の目的は愉快犯のような自己顕示であったが，最近は個人情報を悪用する金銭的なものへと変わってきている．これと共に，作成されるマルウェアの数も飛躍的に増えている状況にある[1]．手元にあるファイルがマルウェアに感染しているかどうか判定する方法の 1 つとして，シグニチャファイルを用いたパターンマッチングによるウイルス検知がある．これにより，ウイルス名を明らかにすることはできるが，ウイルス対策ベンダによってはウイルス名が異なっていること，作成されるマルウェアも多くなってきていること，該当ファイル自身の交換そのものも運用上難しくなってきていることなど，該当ファイルを識別でき，さらに，ファイル自身の属性情報を付加した形態での外部組織との情報交換も必要になってきている．

そこで本稿では，ファジーハッシュのひとつで，ファイル内容の比較を行える ssdeep を用いることで，該当ファイル自身を交換することなく，ファイル構造が類似するファイルの情報提供を目的としたマルウェア情報提供システムを提案する．これにより，ファイル構造が同じファイルだけでなく，構造が類似するファイル一覧を提示することができる．また，出力されたファイル一覧は，検査対象のファイルがマルウェアであるかどうかの目安としたり，新種や亜種のマルウェア解析に役立てたりすることができる．

2. マルウェアによる被害状況と関連対策の動向

2.1 マルウェアによる被害状況

数年前までは，知識と技術を競うようにして巧妙なウイルスを作り，世間を騒がす事件を起こすことがマルウェアの作者達の主な目的であった．2000 年に流行したメールで感染を広げてパソコン内のデータを破壊する LoveLetter や，2003 年頃に流行したネットワークに接続しているだけで感染する Blaster もこのパターンである．

しかし，ここ数年でマルウェアの作者達の目的は，愉快犯のように世間を騒がすことから，マルウェアを用いて個人情報を盗み出そうとする金銭的なものへと変わってきている．目的の変化に連れて，マルウェアの種類も感染したマシンのファイル破壊等を行うウイルスから，現在では攻撃者が感染したマシンを操ることができるボット

[†] 中央大学大学院
Graduate School of Engineering，Chuo University
[††] 日立製作所
Hitachi Ltd.

Vol.2011-DPS-146 No.52
Vol.2011-CSEC-52 No.52
2011/3/11

が主流となっている．また，マルウェアの一般的な感染経路として電子メール，Web サイト，ファイルによる感染が挙げられるが，正規ソフトを装ったマルウェアも発見されている．さらに，コンピュータだけでなく，携帯電話やスマートフォン等で動作するボットの存在も確認されている．

2.2 関連対策の動向

作成されるマルウェアの数も飛躍的に増えている状況にあるという状況において，シグニチャファイルに依存せずに，マルウェアの挙動を基に検知を行うマルウェア動的解析システムの研究が進められてきている[2][3]．また，非マルウェアのホワイトリストデータベース[4]などの，独自のマルウェア対策機能も研究されている．

上記のような新たな検知方法を研究するアプローチとは別に，多くのマルウェア検体を集め，それら検体をウイルス対策ベンダに送付する仕組みなども進められている．2006 年から稼働しているサイバークリーンセンター（Cyber Clean Center）[5]では，インターネットにおける脅威となっているボットをハニーポットで収集し，その特徴を解析することにより，ユーザのコンピュータからボットウイルスを駆除するためのボットウイルス駆除ツール「CCC クリーナ」を作成しユーザに配布している．さらに，収集した検体は，ウイルス対策ベンダに対して提供し，各社の対策ソフトのシグニチャファイルに反映させる活動を取り組んでいる．スペインのセキュリティベンダ Hispasec Sistemas[6]が運営するマルウェアのオンラインスキャンサイトである Virus Total[7]は，2004 年 6 月からサービスが開始され，検査対象のファイルをサイトにアップロードすることで，上限の 20M バイト以下であれば 2011 年 2 月時点で 43 種のウイルス対策ソフトにより検査できる．また，ファイルから生成したハッシュ値を送信することで，過去に解析されたファイルのウイルス対策ソフトによるスキャン結果，ファイル構造情報等のファイル情報を取得することができる．ここで，アップロードされた検体については，必要に応じてウイルス対策ベンダに対して提供されている．

3. ファイル内容，ファイル構造，マルウェア挙動における関連性

本章では，ファイル内容，ファイル構造，マルウェア挙動にそれぞれ対応する ssdeep，セクション構造，ウイルス名称の関連性についての調査結果について述べる．

3.1 調査に用いるデータ

本調査には，サイバークリーンセンターが収集したボット観測データ CCC DATAset 2008/2009/2010 の攻撃元データを用いる．本データは半年から 1 年間，約 100 台のハニーポットで取得したログデータであり，各ログレコードは検体取得日時，送信元／宛先 IP，送信元／宛先ポート，プロトコル，検体のハッシュ値（SHA1），ウイルス名称等で構成されている．さらに，これら攻撃元データの中から，ハッシュ値がオンラインスキャンサイト Virus Total に登録されている 15,300 個を選別し用いた．

3.2 調査項目

関連性の調査にあたっては，Virus Total から取得したファイル情報に含まれる ssdeep，セクション，ウイルス名称の 3 項目を用いる．

（1） ssdeep
（a） 概要

ssdeep[8]はファジーハッシュのひとつで CTPH（Context Triggered Piecewise Hashes）を算出するツールである．ファイルの先頭から順に文字列を生成していくため，生成した文字列を用いてファイル内容の比較を行うことができる．例として，内容が似通う 2 つのプログラムを図 1 に示す．2 つの違いは感嘆符の有無のみであり，ファイル内容が類似するものである．これらのファイルから生成した SHA1，MD5，ssdeep の値を表 1 に示す．

プログラム①

プログラム②

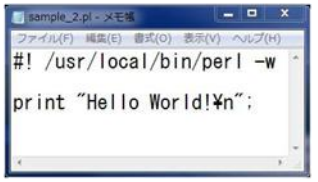

図 1 ファイル内容が似通う 2 つのプログラム

表 1 図 1 のファイルにおける SHA1，MD5，ssdeep の値

項目	プログラム①	プログラム②
SHA1	9108de73c96fa11782f2 e3583d2bfce8608c6f03	34f69f67bbebbddb4962 33e5c1a845be8634983b
MD5	7f5d6406b6db5ea8 feab9b214eb25470	fb143dccb845f5a9 1093bdaa4b0e631e
Ssdeep	3:TFKLncAT7NclPsEB+n :J0ncATBc1s8+	3:TFKLncAT7NclPsE8VHKn :J0ncATBc1srK

Vol.2011-DPS-146 No.52
Vol.2011-CSEC-52 No.52
2011/3/11

2 つのファイルにおける内容はほとんど等しいが，表 1 に示す通り，SHA1 や MD5 といった従来のハッシュ関数では全く異なる値となるが，これに対して，ssdeep の値は似通ったものとなる．このように，ファイル内容の比較に，ssdeep を用いて生成した値を用いることができる．なお，ssdeep の値は「blocksize」「hash1」「hash2」の 3 項目で構成されている（表 1）．ハッシュ値は 2 種類あるが，hash1 の文字列長の上限が 64 文字であることに対し，hash2 は半分の 32 文字であるため，hash1 の方がより詳しくファイル内容を反映している．

（b） 類似度の算出

ssdeep の値における類似度の算出には，レーベンシュタイン距離（Levenshtein Distance，LD）を用いる．LD は文字列の異なり具合を表す数値であり，一方の文字列から他方の文字列に変換するために必要な最小回数である．なお，ssdeep の類似度を算出するに当たり，LD のみでは距離の大きさがそのまま類似度に影響するが，ssdeep の値は固定長ではないため，本研究では hash1，hash2 から算出した LD を以下の式に当てはめて類似度を算出する．

$$\text{LDを用いたssdeepの類似度} = 100 - \left(\left(\frac{\text{LD}}{\text{長い方の文字列長}} \right) \times 100 \right)$$

（2） セクション

（a） 概要

セクションは EXE ファイル内の PE ヘッダの後にある，コンピュータに理解できる言語で記述されたネイティブコード，デバッグに必要なデバッグ情報等の実行に必要な各データである．そのため，セクションの類似度が高い程，ファイル構造が似通っていると言える．Virus Total から取得できるセクションは「name」「viradd」「virsiz」等の 6 項目で構成されている．取得したセクションの各項目と例を表 2 に示す．

表 2 各セクションの項目と内容

項目	内容	例
name	セクション名	.text
viradd	メモリ上におけるセクションの位置	0x1000
virsiz	メモリ上におけるセクションのサイズ	0x244B
rawdsiz	ファイル上におけるセクションのサイズ	0x2600
ntropy	エントロピー	6.27
md5	対応する内容のハッシュ値（MD5）	9d41bd85xxx...

（b） 類似度の算出

比較は項目の name（セクション名）が等しいもの同士で行う．セクションの類似度を算出するに当たり，一致数だけでは比較回数に関係なく類似度に影響するが，ファイルに含まれるセクションの個数は一定ではなく，含まれていないものもあるため，比較で得た 6 項目全部が一致した個数が比較数に占める割合を類似度に用いる．

なお，本稿における「ファイル構造が類似する」とは，セクションの類似度が 50% 以上のものとする．

（3） ウイルス名称

（a） 概要

Virus Total から取得したウイルス名称に関する情報は「ウイルス対策ソフト名」「ウイルス名称」の 2 項目で構成されており，比較はウイルス対策ソフト名が等しいもの同士で行う．また，ウイルスベンダはマルウェアの挙動からウイルス名称を命名しているため，ウイルス名称の類似度が高い程，マルウェア挙動が似ていると言える．

（b） 類似度の算出

本調査で用いるデータ 1 個につき平均 38 種のウイルス対策ソフトが解析しているが，解析日時によっては検体を解析したウイルス対策ソフト数は異なるため，ウイルス名称の類似度は，科名部分が一致した個数が比較数に占める割合を用いる．

3.3 ファイル内容の類似度別におけるファイル構造，マルウェア挙動の関連性調査

本節では，3.2 節で述べた ssdeep，セクション，ウイルス名称を用いて実施した，ファイル内容，ファイル構造，マルウェア挙動の関連性の調査結果について示す．

（1） 調査方法

ssdeep の類似度別にグルーピングしたデータにおいて，セクションやウイルス名称の類似度を調査する．ssdeep の類似度別におけるセクションの類似度を表 3 に，ウイルス名称の類似度を表 4 に示す．なお，各類似度は各グループにおける総比較数分，解析した結果における平均値である．また，比較 1 回におけるセクションの平均比較数は約 3 回，ウイルス名称の平均比較数は約 42 回である．

Vol.2011-DPS-146 No.52
Vol.2011-CSEC-52 No.52
2011/3/11

表 3　ssdeep の範囲別類似度におけるセクションの類似度

hash 別 類似度の範囲		総 Group	総 Hash	総 比較数	6 項目の 一致割合
hash1	60～70	1,993	49,296	516,118	20.73
	70～80	2,491	637,146	119,899,676	23.61
	80～90	3,166	936,472	283,370,088	39.29
	90～100	2,104	102,964	10,006,231	60.28
hash2	60～70	3,573	767,921	104,700,975	22.32
	70～80	3,712	617,076	77,226,427	41.55
	80～90	3,137	308,993	43,470,832	51.66
	90～100	1,582	26,560	1,447,057	57.73

表 4　ssdeep の範囲別類似度におけるウイルス名称の類似度

hash 別 類似度の範囲		総 Group	総 Hash	総 比較数	科名の 一致割合
hash1	60～70	1,993	49,296	516,118	42.76
	70～80	2,491	637,146	119,899,676	16.85
	80～90	3,166	936,472	283,370,088	39.43
	90～100	2,104	102,964	10,006,231	70.89
hash2	60～70	3,573	767,921	104,700,975	14.11
	70～80	3,712	617,076	77,226,427	21.98
	80～90	3,137	308,993	43,470,832	62.22
	90～100	1,582	26,560	1,447,057	68.75

（2） 調査結果

表 3，表 4 より，ssdeep の類似度が高いものはセクションやウイルス名称の類似度も高い結果が得られた．特に，ssdeep の類似度が 90%以上であれば，セクションの約 60%が 6 項目全て一致しており，ウイルス名称は約 70%が少なくとも科名部分が一致しているという結果が得られた．

以上の結果より，ssdeep の値を入力値とすることにより，解析対象のファイルとファイル構造やマルウェア挙動が類似しているマルウェアの情報を提供できると考えた提案システムについて，4 章で述べる．

4. マルウェア情報提供システム

本章ではマルウェア情報提供システムを提案するにあたり，提案システムの概要と提案システムで用いるデータベースの構造について述べる．

4.1 提案システムの概要

提案システムでは，解析対象であるファイル自身を交換することなく，ファイル構造が類似するマルウェアの情報を提供することを目的とする．3 章の調査において，ssdeep の類似度が高い場合には ssdeep，セクション，ウイルス名称の関連性が高いため，提案システムでは入力された ssdeep の値に対して，ssdeep の hash1，hash2 の両類似度が一定値以上のマルウェアの情報を提供する．なお，提案システムでは CCC DATAset 2008/2009/2010 攻撃元データにおける検体のハッシュ値を用いて Virus Total から取得した「ウイルス対策ソフトのスキャン結果」「ssdeep」「セクション」等のファイル情報を，従来のオンラインスキャンサイトと同様に Web ページに表示する．

提供したマルウェア情報により，ユーザは解析対象のファイルとファイル構造が同一，または類似しているマルウェアの情報を把握でき，対象のファイルがマルウェアに感染しているかどうかの 1 つの目安とできる．また，セキュリティ研究者は新種や亜種のマルウェアの解析に役立てることができる．提案システムでは，15,300 個のファイル情報の中から該当するものを提供する．

4.2 データベース構造

本節では，提案システムにおけるデータベース構造について述べる．提案システムで用いるデータベースの構成を図 2 に示す．データベースには提供するデータとして，3 章の調査でも用いた CCC DATAset 2008/2009/2010 攻撃元データにおけるマルウェア情報 15,300 個を，ssdeep の blocksize 毎に作成したテーブル内に格納する．各テーブルに保存されたマルウェア情報は，ssdeep の hash1，hash2 の両類似度が 90%以上のもの同士でグループを構成し，グループ毎にツリーを構成する．そして，ユーザから受信した ssdeep の値に対しては，各グループの親ノードとしたマルウェア情報のものを比較し，hash1，hash2 の両類似度が一定値以上である場合，その親ノードが属するグループのレベル 1 からレベル 3（親ノードにおける孫ノード）までのマルウェア情報を Web ページに表示する．

このように blocksize 毎にマルウェア情報を分別し，さらに親ノードのみを比較対象にして，その子ノードとの比較を省略することで処理の軽量化を図っている．なお，図 2 のようにグループ毎に格納されているマルウェア情報の数は異なるため，提供するマルウェア情報の数は受信した ssdeep の値により異なる．

Vol.2011-DPS-146 No.52
Vol.2011-CSEC-52 No.52
2011/3/11

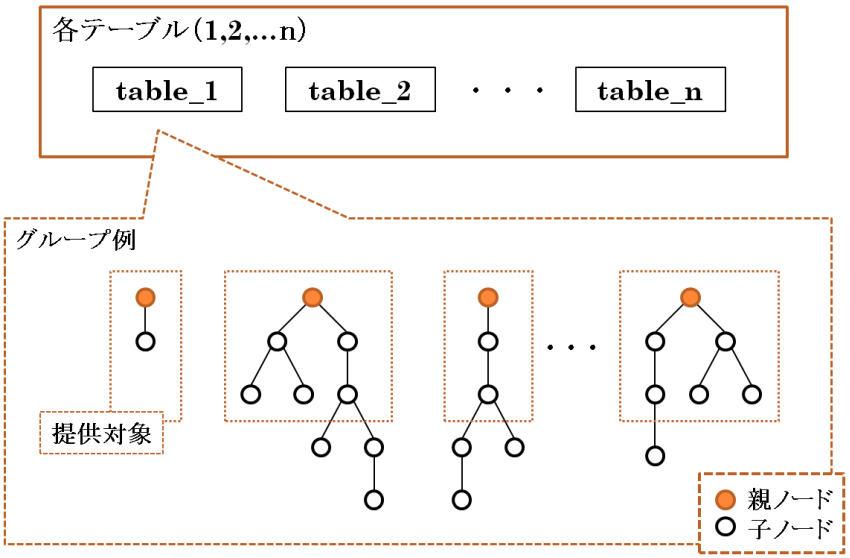

図 2　データベースの構成

5. 評価

マルウェアは同じ科名のものでも，ファイル構造が類似するものと異なるものがある．そこで本章では，実装した提案システムのプロトタイプシステムを用いて，前述した 2 つのタイプのマルウェアに対して提案システムの有用性を評価する．なお，提案システムの動作環境は以下の通りである．

- VMware Player 3.1.2
- OS：Fedora Core 10
- データベース：MySQL 5.0.83

5.1 評価方法と結果

共通の科名を持つマルウェアにおいて，ファイル構造が類似するものと異なるものの 2 タイプのものを対象に，提案システムの有用性を評価する．具体的には，まず，共通の科名を持つマルウェアから生成した ssdeep を入力とした場合に提案システムから提供されるマルウェア情報のファイル構造における類似度を算出する．次に，ファイル構造が類似するものは前者のマルウェアとして，異なるものは後者のものとして評価を行う．

今回の評価には，10 種のウイルス対策ソフトにおける計 153 個の科名を含んだウイルス名称を持つマルウェア計 1,133 個の ssdeep を入力として用いた場合に得られるマルウェア情報 8,845 個を用いた．評価に用いたウイルス対策ソフト別における科名数，入力に用いたマルウェア数，提供されたマルウェア情報数を表 5 に示す．なお，今回の評価では入力の ssdeep の値に対して，類似度が 80%以上のファイルを提供情報として用いた．

表 5　評価に用いたウイルス対策ソフト数別の科名数と入力数

ウイルス対策ソフト	科名数	入力用のマルウェア数	提供されたマルウェア情報数
AVG	31	270	2,618
ClamAV	2	3	4
Comodo	11	134	1,879
DrWeb	4	15	132
Emisisoft	10	109	401
Ewido	3	40	21
F-prot	11	62	206
NOD32	32	207	630
NOD32Beta	42	226	2,153
NOD32v2	7	67	801
合計	153	1,133	8,845

入力用として用いた 1,133 個の各 ssdeep の値に対して，提供されたマルウェア情報同士におけるセクションの類似度は全てが互いに類似していた．したがって，今回の評価で用いたデータ，すなわち，CCC DATAset 2008/2009/2010 の攻撃元データからハッシュ値が Virus Total に登録されているデータとして選別したものについては，表 5 に記したウイルス対策ソフトにおいて，共通の科名を持つものはファイル構造が類似するもののみであったと思われる．一方，共通の科名を持ち，ファイル構造が異なるマルウェアに対する評価を行うことができなかったため，今後の課題とする．

6. おわりに

本稿では，解析対象であるファイル自身を交換することなく，ファイル構造が類似するファイルの情報提供を目的としたマルウェア情報提供システムを提案した．まず，

Vol.2011-DPS-146 No.52
Vol.2011-CSEC-52 No.52
2011/3/11

ssdeep，セクション，ウイルス名称における関連性の調査を行い，ファイル内容，ファイル構造，マルウェア挙動（ウイルス名称）における関連性の調査結果を示した．調査結果から，ファイル内容が類似していれば，ファイル構造やマルウェア挙動（ウイルス名称）も類似していることを示し，提案システムが提供するマルウェア情報の実現可能性を示した．また，同じ科名を持つマルウェアでも，ファイル構造が類似するものと異なるものがあることから,提案システムのプロトタイプシステムを用いて，2 タイプのマルウェアに対する提案システムの評価を試みた．具体的には，共通の科名を持つマルウェアから生成した ssdeep の値を入力とした場合において，提案システムが提供したマルウェア情報におけるファイル構造の類似度を調査し，評価を行った結果，本稿で評価に用いたマルウェアにおいて，共通の科名を持つものは全てファイル構造も類似していると思われる．

今後の課題としては，共通の科名を持ちファイル構造が異なるマルウェアに対する評価，マルウェア情報の表示操作性の向上，LD 以外の ssdeep に対する類似度算出方法の適用などを検討し，提案システムを改善する予定である．

参考文献

1) Panda Security: 2010 Annual Security Report, http://press.pandasecurity.com/news/2010-annual-security-report/
2) C. Willems, et al.,“Toward Automated Dynamic Malware Analysis Using CWSandbox,” IEEE Security and Privacy Magazine, Vol.5, Issue 2, 2007.
3) D. Inoue, et al.,“Automated Malware Analysis System and its Sandbox for Revealing Malware's Internal and External Activities,” IEICE Trans. Information and Systems, Vol.E92-D, No.5, 2009
4) National Software Reference Library, http://nsrl.nist.gov/, accessed at 14/02/2011.
5) サイバークリーンセンター, https://www.ccc.go.jp/
6) Hispasec - Seguridad Informatica，http://www.hispasec.com/
7) Virus Total，http://www.virustotal.com/jp/
8) Fuzzy Hashing and ssdeep，http://ssdeep.sourceforge.net/