CCC DATAset における連携するマルウェアの変遷

大類 将之 † 菊池 浩明 † 寺田 真敏 ‡ Nur Rohman Rosyid††

† 東海大学大学院 工学研究科 情報理工学専攻
259-1292 神奈川県平塚市北金目 4-1-1
yama, kikn@cs.dm.u-tokai.ac.jp

‡ 日立製作所 Hitachi Incident Response Team (HIRT)
212-8567 神奈川県川崎市幸区鹿島田 890 日立システムプラザ新川崎
masato.terada.rd@hitachi.com

†† モンクット王工科大学ラカバン校
Chalongkrung Road, Ladkrabang Bangkok 10520, THAILAND
nrohmanr@gmail.com

あらまし 本研究では，CCC DATAset 2008 から 2010 の攻撃通信データ，攻撃元データを用いて，3 年間に渡るマルウェアの振る舞いに着目し，データマイニング手法である Apriori と PrefixSpan を適用することで，その変遷や特徴を報告する．

Evolution of Botnet Coordinated Patterns from CCC DATAset

Masayuki Ohrui† Hiroaki Kikuchi† MasatoTerada‡ Nur Rohman Rosyid††

†Course of Information Science and Engineering,
Graduate School of Engineering, Tokai University
4-1-1 Kitakaname, Hiratsuka, Kanagawa 259-1292, JAPAN
‡Hitachi Incident Response Team (HIRT), Hitachi, Ltd.
890 Kashimada, Kawasaki, Kanagawa, 212-8567, JAPAN
††King Mongkut's Institute of Technology Ladkraban
Chalongkrung Road, Ladkrabang Bangkok 10520, THAILAND

Abstract This paper aims to apply data mining technique Apriori and PrefixSpan, detecting feature and evolution coordinated attacks from using the captured packets data and the downloadinglogs of the CCC DATAset 2008-2010 by focusing on the behavior of malware over the past three years.

1 はじめに

近年のマルウェアは，数多くの亜種が存在し，複数のダウンロードサーバに分散して感染するなど，複雑化，高度化が進んでいる．特に分散されたサーバによるマルウェアの連携感染は検出をより困難としている．

我々は MWS2009 にて，攻撃元データに対して，価値ある相関ルールを抽出するデータマイニング手法であるアソシエーション分析 (以下，Apriori) を適用し，関連性の強いマルウェアの組み合わせ，すなわち，連携感染を自動抽出する手法を提案した [1]．また，系列パターンマイニングである PrefixSpan を用いることで，Apriori の欠点であった時系列を考慮したルー

表 1: データマイニング方法の比較

	Apriori	PrefixSpan
提案者	Agrawal，他 [4]	Pei，他 [5]
抽出対象	相関ルール (A，B → C)	シーケンスパターン (A，B，*，C)
精度	支持度，確信度	確信度
特徴	アイテムの集合 (順序なし)	シーケンス (順序あり)

ルが抽出可能であることを示した [2]．Apriori と PrefixSpan の違いを表 1 に示す．

しかし近年，新たに Gumblar をはじめとする Web 感染型マルウェアが台頭し，被害が増加している．逆に，攻撃元データに含まれるマルウェアの感染数は減少傾向にある．これは攻撃の主流が Web 感染型マルウェアに移行して来ていることを示唆している．では，実際にマルウェアの連携感染は減少しているのだろうか．

そこで本研究では，研究用データセット CCC DATAset 2008 から 2010 の攻撃通信データ，攻撃元データ [3] を用いて，3 年間に渡るマルウェアの振る舞いに着目し，データマイニング手法である PrefixSpan を適用することで，連携感染の変遷を調査した．その結果明らかになった特徴を報告する．

2 要素技術

2.1 Apriori[4] アルゴリズム

Apriori アルゴリズムは，Agrawal らが提案した代表的な相関ルール抽出アルゴリズムである．支持度と確信度という閾値に最小値を与え，その値を元に数多く抽出される相関ルールの中から，効果的に価値の低いルールを枝刈りする．これにより，価値の低いルールを除き，効率よく価値あるルールを発見できる．

2.2 PrefixSpan[5] アルゴリズム

PrefixSpan アルゴリズムは，Pei らが提案した Prefix Projection という射影を用いた系列パターンマイニングのアルゴリズムである．系列データから，射影対象の系列より後ろに存在するアイテムのみを抽出し，深さ優先で射影を繰り返すことで，頻出する系列パターンを効果的に発見する．

3 連携感染

3.1 定義

複数のダウンロードサーバによる連携で，個別のマルウェアを組み合わせて感染させる攻撃を連携感染と定義する．CCC DATAset 2009 攻撃通信データから得た例を以下に示す．

1. `PE_VIRUT.AV` に感染 [124.86.165.*]
2. ss.ka***.com (DNS) を名前解決し，hub.56***.com (IRC) に接続 [67.43.226.*]
3. always***.com (DNS) [67.215.1.*]，zone t***.info (DNS) [72.10.166.*] を名前解決
4. 各サーバから `TROJ_BUZUS.AGB` (/vot.exe) [67.215.1.*]，`WORM_SWTYMLAI.CD` (/vss.exe) [72.10.166.*] を HTTP GET によりダウンロードする
5. ポートスキャンを行う

このように連携感染は，起点となるマルウェアをダウンロード後，IRC サーバに接続し，他のマルウェアを HTTP GET によりダウンロードする．2008 年，2010 年もサーバの数や感染するマルウェアの数など若干の違いはあるが，この一連の流れで感染を行っている．

3.2 実験データ

実験データとして，CCC DATAset 2008，2009，2010 の攻撃通信データと攻撃元データを使用する．攻撃通信データは定期的にクリーンな状態にリセットされるため，Windows XP が送信

するNTPパケットを利用して，各データを一定間隔で分割した．これをタイムスロット（以下，スロット）と呼ぶ．スロットを1つのトランザクションとし，その間にダウンロードされたマルウェアの種類をそのトランザクションに生じるアイテムとして，頻出アイテムを抽出する．同様に，攻撃元データもスロットに分割する．

3.3 調査結果

3.3.1 マルウェアの活動傾向

攻撃元データから3年間観測されているマルウェアが存在するか調査した．結果を表2に示す．3年間を通して上位のマルウェアは，PE_VIRUT.AVである．PE_VIRUT.AVは，3.1節で述べた連携感染の起点となるマルウェアである．感染数は減少傾向だが，マルウェアファミリ名もPEが占めており，その勢力は衰えていない．

次に，連携感染に用いられるIRC，DNSを攻撃通信データから抽出した．それぞれ表3，表4に示す．抽出には，PINGの送信先を使い，数は各スロットのユニーク数としている．IRCでは，3年間共通してhub.*****.comが使用されていた．これは，いずれもPEを起点とした連携感染で用いられるドメインである．同様に，DNSでも一部ドメインが3年間使用されているのを確認できた (表中太字)．以上の結果から，連携感染はなくなっておらず，最低でも1つ以上のボットネットで使用されていると考えられる．

3.3.2 連携感染の活動傾向

本節では，攻撃元データから，Apriori，PrefixSpanを用いて連携感染を抽出し，連携感染の活動傾向を調査する．なお，2008年の攻撃元データはハニーポットの記載がないため，使用していない．

まず，連携感染数の変化を図1に示す．全ハニーポット730日分のデータに対し，Aprioriを用いて相関ルールの抽出を行い，月平均を求めた．Aprioriを使用した理由は，PrefixSpanに比べ，抽出されるルール数が正確であったためである．系列の長さ3以上のパターンに絞っ

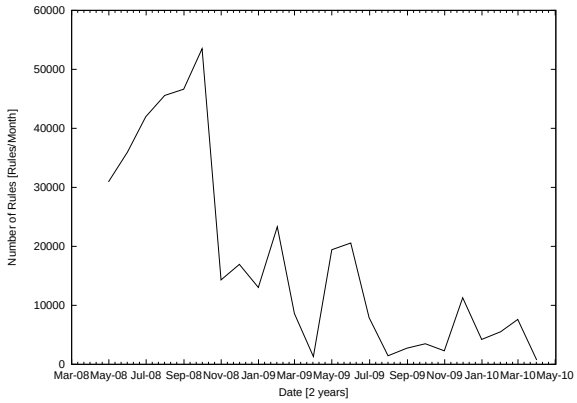

図 1: 2009，2010年の連携感染数の推移 (全体)

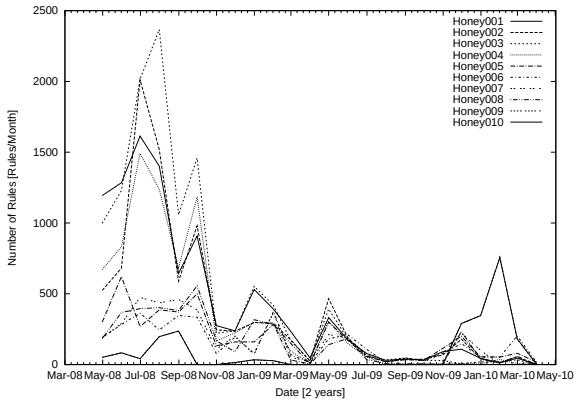

図 2: 2009，2010年の連携感染数の推移 (Honey001 ～ 010)

ており，2種類以下のマルウェア間のルール及びUNKNOWNを含むルールは除外している．図1から，マルウェアの減少に伴い，連携感染も減少傾向であることが分かる．

Honey001 ～ 010をまとめたものを図2に示す．ハニーポットによって傾向の違いこそあるものの，全体の傾向としてはやはり減少傾向であった．攻撃通信データを解析した結果でも，2009年は3種類の連携感染が確認できたことに対し，2010年は1種類のみである．

次に，連携感染は何種類のマルウェアで構成された感染を行うのか，すなわち，平均連携マルウェア数を調査する．抽出にはPrefixSpanを用いて連携マルウェア数は3以上とし，全ハニーポットの連携感染パターンを抽出して平均を求めた．PrefixSpanは時系列を考慮できるため，感染順を考慮して連携感染パターンを抽出する

表 2: 3 年間共通して観測されたマルウェア

マルウェア名	2008 年		2009 年		2010 年	
	順位	Uniq.	順位	Uniq.	順位	Uniq.
PE_BOBAX.AK	8	47654	3	94324	32	8018
PE_VIRUT.AV	9	46741	2	222207	1	194557
WORM_ALLAPLE.IK	10	45033	12	30319	19	12564
PE_VIRUT.XV	20	26518	28	16625	31	8424
PE_VIRUT.XZ	46	14315	51	8885	33	7181
PE_VIRUT.PAU	63	10749	47	9347	21	11815
BKDR_VANBOT.HG	93	6050	43	11206	24	10404

表 3: 3 年間共通して用いられた IRC サーバ

順位	2008 年		2009 年		2010 年	
	IRC ドメイン	数	IRC ドメイン	数	IRC ドメイン	数
1	hub.40***.com	81	hub.14***.com	35	pwned30.i***.net	31
2	i	38	-	-	pwned28.i***.net	30
3	hub.56***.com	36	-	-	hub.63***.com	23
4	hub.44***.com	31	-	-	hub.48***.com	20
5	aaa.59***.com	3	-	-	hub.27***.com	14
6	irc.foo***.com	2	-	-	no***.org	13
7	bl*.com	2	-	-	s*.com	8
8	FE7B03EC	1	-	-	ja**.org	5
9	F3B4433F	1	-	-	irc.fo***.fo	1

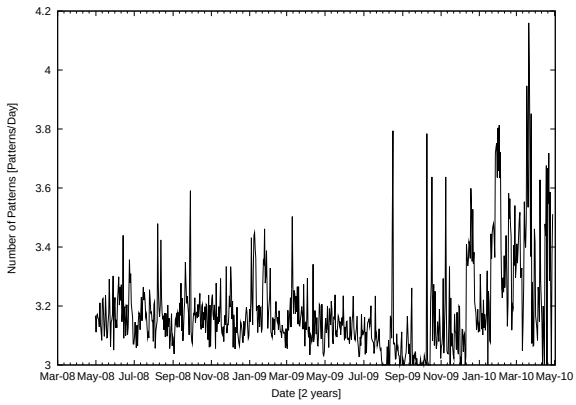

図 3: 連携感染の平均連携マルウェア数

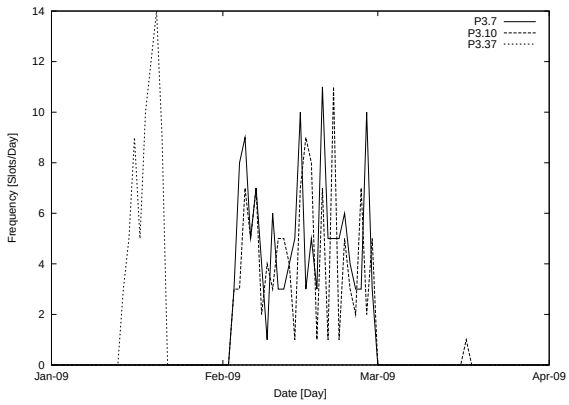

図 4: 連携感染の活動期間 [2]

ことができるためである．こうして求められた平均連携マルウェア数の変化を図 3 に示す．連携感染総数の減少とは逆に，構成マルウェア数は増加している．これは連携感染はより複雑化し，巧妙化してきていることを示している．

また，攻撃通信データを解析したところ，HTTP GET でダウンロードされるマルウェア数は，2008，2009 年で 2 種類であったのに対し，2010 年では 5 種類となっており，ここからも連携パターンの複雑化が裏づけられる．

最後に，連携感染の活動期間を調査する．重複しない 3 種類のマルウェアによる連携パターンの観測期間を図 4 に示す．1 日に観測されたスロット数を元にしている．このように連携感染の活動期間は 2 週間〜1 ヶ月と非常に短い．例えば，PE_VIRUT.AV は各年で連携するマルウェアが変わっている．

表 4: 2010 年の DNS ドメインとその比較

順位	DNS ドメイン	数	2008 年	2009 年
1	botz.noreta***.com	133		
2	proxim.ntkrn***.info	62		
3	checkip.dyn***.org	60		
4	www.whatism***.org	52		
5	tx.mostafaaljaaf***.net	35		
6	tx.nadersam***.org	32		
7	www.whatsmyipaddr***.com	31		
8	www.getm***.org	28		
9	**ss.ka***.com**	19	31 位	1 位
10	**ss.nadnad***.info**	16	81 位	5 位
11	**ss.MEMEH***.INFO**	15	90 位	
12	videogale***.com	12		
13	blah.swapixtr***.com	10		
26	xx.nadna***.info	2		

表 5: 2010 年の攻撃通信データと攻撃元データの比較 (一部)

No.	マルウェア名	Prot.	攻撃通信	攻撃元	誤差
1	WORM_DOWNAD.AD	TCP	118	79	−39
2	WORM_PALEVO.SMD	TCP	106	36	−70
3	WORM_PALEVO.BL	TCP	49	12	−37
4	PE_VIRUT.AV	TCP	42	26	−16
5	TROJ_BUZUS.MC	TCP	25	11	−14
6	BKDR_RBOT.SMA	UDP	43	13	−30
7	BKDR_MYBOT.AH	UDP	13	4	−9
8	WORM_SDBOT.CEM	UDP	6	5	−1
9	WORM_MYTOB.IR	UDP	1	0	−1
合計	-	-	512	261	−251

3.4 考察

本節では，なぜダウンロード数，連携感染数が減少したのかを考察する．

CCC DATAset 2010 の攻撃元データ数が不自然に少なく，検出誤りが生じていると考えた．そこで，攻撃通信データを解析し，各スロットを TCP と UDP の通信に分割し，それぞれ tcpflow，TFTPgrab を用いて分析することでマルウェア検体を得て，攻撃元データとの比較を行った．表 5 に比較結果を示す[1]．攻撃通信データに含まれるマルウェアが攻撃元データに含まれていない．

特に TCP 通信では WORM_PALEVO.SMD，UDP 通信では BKDR_RBOT.SMA が数多く得られたが，攻撃元データとの差はそれぞれ 70 個と 30 個と大きい．

以上より，2010 年の攻撃元データの 261 件には，攻撃通信データの 512 件に対して，1.961 倍の未検出があったことがわかる．2009 年では，攻撃元データが 200，攻撃通信データが 221 で 1.105 倍の未検出に留まっていた．この結果を考慮して補正したダウンロード数の推移を図 5 に示す．実際の未検出数がわかる 2009 年，2010 年の攻撃通信データを使用して，その値を元に補正を行っている．ダウンロード数は全ハニーポットにおける総数である．なぜならば，2008 年は，攻撃元データにハニーポット ID が含まれず，区別できないためである．なお，2009 年 5 月末で 2 台のハニーポットが停止している．

全体の傾向として，HTTP GET によってダウンロードされたマルウェア，あるいは，UDP

[1] 3 年間の攻撃元データ及び Virus Total に該当するハッシュ値がないマルウェア検体は調査対象から外した．なお，2008 年の攻撃通信データから分析して得られたマルウェア数は 673 個で圧倒的に多いが，攻撃元データと比較できないため省略する．

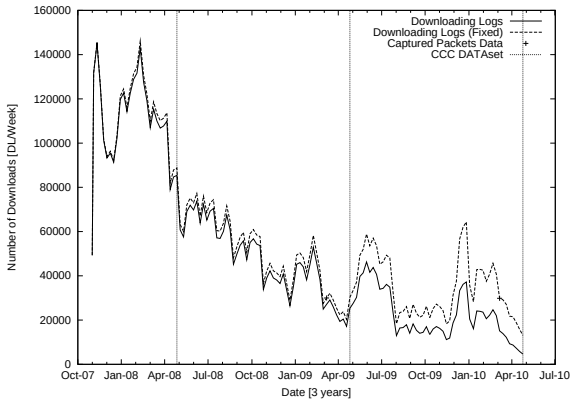

図 5: マルウェアのダウンロード数の推移

通信によるマルウェアが未検出であることが多かった．同一のマルウェアを検出できていることもあったため，マルウェア検体自体が原因，もしくは，収集しているハニーポット固有の問題と考えられる．

4 おわりに

過去 3 年間の攻撃通信データ，攻撃元データを用いて，連携感染の変遷及び特徴を報告した．連携感染数が減少する一方で，感染するマルウェアの連携パターン数は増加していた．この増加は，2010 年の攻撃通信データ，攻撃元データの解析結果から裏づけられた．

謝辞

本研究を遂行するにあたり，ご助言及びご協力を下さった日立製作所の藤原将志氏に深く感謝する．

参考文献

[1] 大類，他，“分散ハニーポット観測からのダウンロードサーバ間の相関ルール抽出”，マルウェア対策研究人材育成ワークショップ 2009（MWS2009）, 2009.

[2] N. R. Rosyid，et al.，“Frequent Sequential Attack Patterns of Malware in Botnets”，第 48 回 コンピュータセキュリティ研究会 (CSEC48)，2010.

[3] 畑田，他，“マルウェア対策のための研究用データセット ～ MWS 2010 Datasets ～”，マルウェア対策研究人材育成ワークショップ 2010（MWS2010），2010.

[4] R. Agrawal，T.Imielinski，A. Swami，“Mining Association Rules between Sets of Items in Large Databases”，Proc. of ACM SIGMOD-93，pp. 207-216，1993.

[5] J. Pei，et al.，“PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth”，Proceedings of the 17th international Conference on Data Engineering，pp. 215-224，2001.