Computer Security Symposium 2017
23 - 25 October 2017

ダークネットトラフィックデータの頻出パターン解析

橋本 直輝[1]　小澤 誠一[1]　班 涛[2]　中里 純二[2]　島村 隼平[3]

概要：本研究では 2016 年に注目を浴びた IoT マルウェア Mirai のソースコードが公開されるまでの期間においてダークネットトラフィックで観測された興味深い結果を報告する．観測データとして 2016 年 7 月 1 日から 2016 年 9 月 30 日まで NICT の NICTER プロジェクトで観測している/16 のセンサで収集された TCP/SYN パケットを利用した．期間中収集されたの 26,936,229 ホスト，合計 2,590,493,232 パケットに対して頻出パターンマイニングと相関ルールの解析を行った．本稿では特に，TCP ヘッダの「ウィンドウサイズ」,「宛先ポート」および IP ヘッダの「パケットの優先度」で出現した相関ルールに注目し，それらのパケットを送信したホストの一部が，公開された Mirai のソースコードの条件と合致していることを発見した．

キーワード：cybersecurity, machine learning, association rule learning, darknet traffic analysis, IoT Malware

Association Rule Analysis for Darknet Traffic Data

Naoki Hashimoto[1]　Seiichi Ozawa[1]　Tao Ban[2]　Junji Nakazato[2]　Jumpei Shimamura[3]

Abstract: In this paper we report the interesting observation of the darknet traffic before the source code of IoT malware Mirai was first opened, which gained much attention in 2016. We used TCP SYN packets collected from July 1st 2016 to September 15th 2016 with the NICT /16 darknet sensor. The frequent pattern mining and the association rule learning were performed to 2,590,493,232 packets in total which were sent from 26,936,22 unique hosts. We successfully extracted frequent patters on TCP Window Size, Destination Port and Type of Service . In addition, we show that some of the such hosts sent SYN packets satisfying the conditions known from the source code of Mirai.

Keywords: cybersecurity, machine learning, association rule learning, darknet traffic analysis, IoT Malware

1. はじめに

情報技術（IT）は人生に大きな変化をもたらし，多くの人々がインターネット上で新しい利益を得ている．近年，この IT 革命に加えて，さまざまなサービスやデバイスがネットワークに接続される IoT（Internet of Things）の大きな進歩は，私たちにさらなる革新をもたらそうとしている．しかし，その一方で IT と IoT システムの高度化に伴い，新しいシステム脆弱性を利用したサイバー攻撃が深刻化している．特に最近では，2016 年に公開された IoT マルウェア Mirai が大きな影響を与えた． Mirai は自己複製のために同様の脆弱性を持つ IoT デバイスを検出するワーム型のマルウェアである．攻撃者は，Mirai に感染した多数の IoT デバイスをボットとして操作し，これを使用して多数のパケットをターゲットホストに送信することにより分散型サービス妨害（DDoS）攻撃などを行う．

このような大規模なサイバー攻撃に迅速に対処するためには，インターネット上で発生するサイバー攻撃を広い視野で観察できる仕組みを構築する必要がある．この目的のために，ダークネットの使用が長年にわたって研究されて

1 神戸大学大学院工学研究
Graduate School of Engineering, Kobe University, JAPAN
2 国立研究開発法人情報通信研究機構
National Institute of Information and Communications Technology
3 株式会社クルウィット
clwit Inc

いる [1]． ダークネットは未使用の IP アドレス空間である．ダークネットには何も接続されておらず，サービスを提供していないために通信が発生しないと考えられるが，実際には多くのパケットが到着している．これらのパケットは主に，スキャン活動によって引き起こされる接続要求パケットもしくは DDoS 攻撃の対象となるホストからの応答パケットであるバックスキャッタである．したがって，ダークネットで観測されるパケットの多くはマルウェアもしくは悪意ある攻撃者によって生成されると考えられる．このように，ダークネットパケットの分析を通じて，インターネット上で行われているサイバー攻撃の一部を観察することができると考えられる．

本研究では，ダークネットで観測されたパケットの中からスキャンの挙動について分析を行った．特に，スキャン攻撃を特徴付ける TCP/SYN パケットに注目し，それらのパケットの TCP ヘッダに統計的なルールを見つけることを目指した．この目的のために，SYN パケットに相関ルール学習を適用し，スキャン攻撃を行うマルウェアのヘッダ情報の特徴を観測した．宛先ポートの情報に関しては，SYN パケットを分析するいくつかの先行研究が報告されている． Ban ら [1] は SYN パケットの宛先ポート番号に相関ルール学習を適用し C. Stocker や E.L. Malecot [2], [3] は，Carna ボットネットやその他のマルウェアに関連するいくつかの相関ルールを発見した．これらのルールは現在，ネットワークスキャンを実行する際にシグネチャとして使用されている．今回の実験では，TCP および IP の全てのヘッダ情報に焦点を当てた．

この論文は以下のように編成されている．第 2 節では，ダークネット解析と，FP-tree / FP-growth アルゴリズムに基づく相関ルール学習について簡単に説明する．第 3 節では，ダークネットトラフィックデータから TCP SYN パケットのウィンドウサイズに関する相関ルールを検索し，2016 年 9 月 7 日付近に観測されたダークネットトラフィックデータから発見された興味深いルールマイニング結果を示す方法を提示する．第 4 節では得られた相関ルールの結果をまとめる．第 5 節では，本稿の今後の課題について述べる．

2. ダークネットトラフィックデータに対する相関ルール学習

2.1 ダークネットトラフィックデータ

ダークネットは，インターネット上の到達可能かつ未使用の IP アドレス空間である．今日のインターネット上のほとんどの通信で使用される IPv4 では，IP アドレスは 32 ビットのデータとして表される．したがって，約 43 億の IP アドレスが存在していることになる．ただし，すべてのアドレスがホストコンピュータに割り当てられているわけではない．通常のインターネット使用では，未使用の IP アドレスへのパケット送信は発生しないが，実際にはかなりの数のパケットが到着している．この事実には主に 2 つの理由がある． 1 つはマルウェアによるスキャン活動であり，もう 1 つは DDoS 攻撃を受けたターゲットホストから送信された応答パケットのバックスキャッタである．

スキャンとは，単にポートの開閉を調べるものやホストにセキュリティの脆弱性が存在するかどうかを確認する活動のことを示している．スキャンにおいて攻撃者は，ランダムに接続要求を行ったり，広範囲の IP アドレスと宛先ポート番号を総当りに接続を試みることで，応答を参照し，宛先ホストの状態（脆弱性）のチェックを行う．このスキャンには，ネットワークスキャンとポートスキャンの 2 種類が存在している．ネットワークスキャンは主に特定のサービスを動作させているかどうか調べることが目的である．その一方で，ポートスキャンは，特定のホストのどのポートがオープンしているかを調べることが目的である．ポートスキャンの攻撃にはいくつかの種類がある．SYN スキャンは，TCP 通信で SYN パケットを送信するスキャンである．これは，SYN パケットを送信後に RST パケットを送ることでサーバー上にログを残すことなく実行されるため，ステルススキャンとして知られている．

2.2 相関ルール

本節では，データ間に潜在する興味深い相関関係を発見するために利用されている相関ルール学習 [4] について簡単に説明を行う．

2.2.1 頻出パターンマイニング

頻出パターンマイニングおよび相関ルール学習は，もともと市場バスケットデータから一緒に購入されるアイテムの頻出なグループを見つけるために提案された [5], [6]．文献 [5] の定義に従えば，相関ルール学習は以下のように定義される．

$\mathcal{D} = \{T_1, T_2, \cdots, T_N\}$ をデータベースと呼ばれる N 個のトランザクションの集合とする．$I = \{i_1, i_2, \cdots, i_M\}$ をデータベースに存在するすべての M 個のアイテム（項目）の集合とする．D のなかの各トランザクションは一意のトランザクション ID を持ち，I にあるアイテムの部分集合を割り当てられる．アイテム集合を含むデータベースにおいて支持数もしくは支持度と呼ばれる値が与えられる．支持数は条件 X を満たすトランザクションの値，支持度はその割合として $\mathrm{supp}(X)$ で定義される．

頻出パターンマイニングの際あらかじめ最低支持数もしくは最低支持度 S を閾値として設定し，全てのトランザクションの中で少なくとも S 個存在するすべてのパターン $P(P \subset I)$ を探し出すことである．

続いて**相関ルール** は，以下のような形式として定義される．

$$X \to Y, \text{ for } X, Y \subseteq I, \ X \cap Y = \emptyset. \quad (1)$$

アイテム集合 X および Y は，それぞれ前提部および結論部と呼ばれる．ルールの確信度は，条件付きの確率として以下のようになる，

$$\text{conf}(X \Rightarrow Y) = \text{supp}(X \cup Y)/\text{supp}(X). \quad (2)$$

すべての出現する可能性のあるルールの集合から強い相関ルールを抽出するために，最低確信度の閾値 C を設定し，S および C の両方を満たすものを強相関ルールと呼ぶ．

相関ルール学習 は，以下の 2 つのステップで行うことができる．

(1) 頻出パターンマイニング：各アイテム集合は最低支持数を満たす，すなわち，少なくとも S と同じ頻度で発生する．

(2) 強相関ルールの抽出：定義上，最低支持数を保証された頻出アイテム集合において作成されたルールが最低確信度の制約を満たしたものである．

2.2.2 FP-tree を用いた頻出パターンマイニング

相関ルール学習の第 1 ステップは，アイテムのすべての可能な組み合わせの冪集合 (候補集合) を検索することあるが，この集合の大きさは $\mathcal{I}$ のアイテム数 n によって指数関数的に増加する問題がある．そのため候補集合をつくることなく頻出な組み合わせを探索することが求められる．

そこで，“頻出なアイテム集合のすべての空集合ではない部分集合もまた頻出でなければならない．”という考えから，頻出なパターンマイニングのための現在最も高速で最も一般的なアルゴリズムの 1 つである，Frequent Pattern growth（FP-growth）アルゴリズムで実験を行う [4], [6]．このアルゴリズムは，与えられたデータベースのプレフィックス木表現に基づいている．プレックス木のデータ構造（いわゆる FP-tree, FP-growth）を使用することにより，トランザクションを格納するためのメモリを大幅に節約することができる． FP-growth アルゴリズムの基本的な考え方は，以下に示すような再帰的に抽出を行うプログラムとして述べることができる．

(1) 最初の過程で，頻出アイテムの集合とその支持数を取得する．トランザクションから最低支持数の条件を満たさないすべてのアイテムを削除し，すべての頻出アイテムは，頻度の高い順に見出し表に格納される．

(2) 2 番目の過程では，根のノードが ’null’ とラベル付けされた木にインスタンスを挿入して FP-tree を構築する． FP-tree の処理を高速化するために，各トランザクションのアイテムは見出し（ヘッダテーブル）と同じ順序でソートされる．アイテムを含むすべてのトランザクションがこのリストを捜査することによって参照およびカウントできるように，同じアイテムを参照するすべてのノードは，リストによって索引付けされる．リストの見出し要素は，見出し表内の対応するアイテムに関連づけられる．

(3) FP-tree の再帰的な抽出過程では，候補アイテム集合を生成し，データベース全体に対してテストすることなく，大きなアイテム集合を直接的に拡張することができる．まず頻出集合の一番頻度が小さいものから開始し，長さが 1 の条件付きアイテムベースを作成する．次に，条件付き FP-tree が作成され，元の木から投影されたカウントが条件付きのインスタンスの集合に対応し，各ノードはその子のカウントを合わせた値となる．再帰的な FP-growth は，個々の項目条件が最低支持数を満たさない場合に終了し，処理は元の FP-tree の残りの見出し項目で継続される．

(4) 再帰プロセスが完了すると，最低支持数の制約を満たすすべての大きなアイテム集合を発見することができ，相関を持つルールの作成が開始される．

2.2.3 相関ルール学習

相関ルールは，頻出なアイテム集合から以下のステップで生成することができる．

(1) 各頻出アイテム集合 l に対して，l の空でない部分集合をすべて生成する．

(2) l の空でない全ての部分集合 s について，“$s \to (l-s)$” のルールにおいて確信度が設定された最低確信度 C を満たせば強相関ルールとして出力される．

ルールは頻出なアイテム集合から生成されるため，このように作成されたすべての関連ルールは自動的に最低支持数を満たす．

3. ホストの特性分析の方法

サイバー攻撃を行うホストの挙動解析には，静的解析と動的解析があるが，本稿では，ダークネットトラフィックに相関ルール解析を適用した動的解析法を提案する．

ダークネットで観測されたスキャン活動をあらわすパケットを調査するために，TCP/SYN パケットに注目する．パケットデータの分析を行うにあたって，すべての TCP ヘッダと IP ヘッダの各要素に対して「トランザクション」の作成を行う．作成方法は，まずパケットの送信元 IP アドレスの値がトランザクションの ID となる．次に各送信元 IP アドレスごとに少なくとも 1 回は使用した値の集合を各 ID に割り当てる．以上で得られたアイテムの集合が「トランザクション」となる．この処理を行うことでスキャンを行ったホストがどのような組み合わせを用いているか観測することに注目することができる．

次に，得られたトランザクションに対して FP-growth を実行し，相関ルールの学習を行う．さらに 一定期間，相関ルール学習を続けることで時間経過に伴ってパケットデータの傾向がどのように変化するかを調査する．

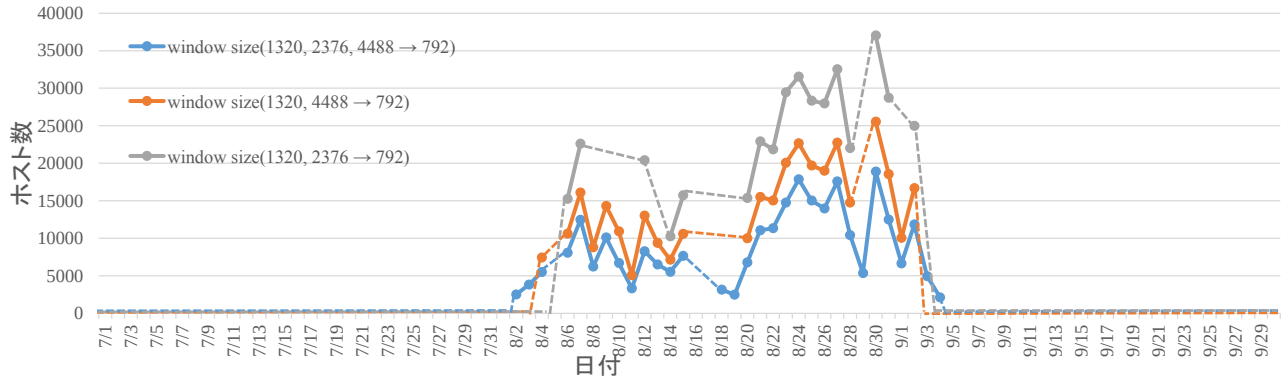

図 1　ウィンドウサイズの相関ルール

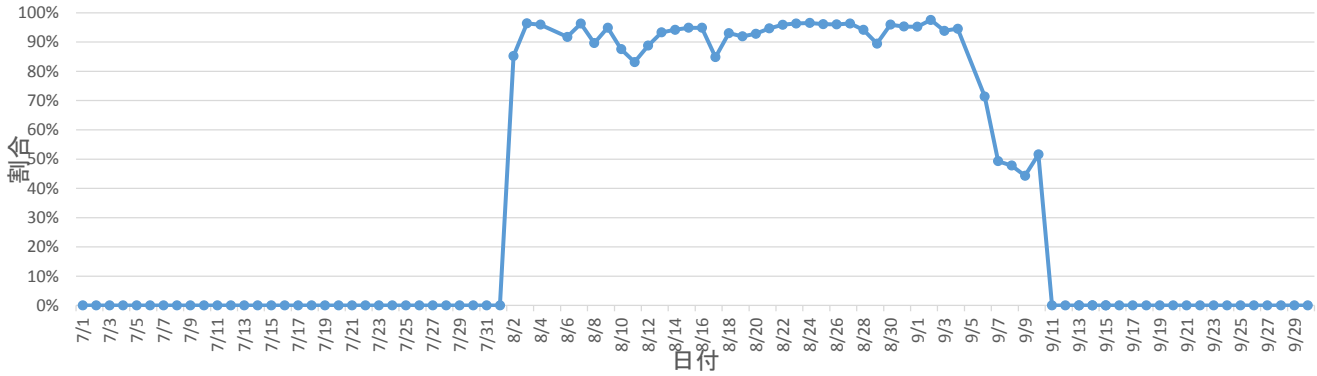

図 2　ウィンドウサイズ相関ルールの Mirai の特徴との照合

4. 実験結果

解析対象となるデータセットは，2016 年に NICT の NICTER プロジェクトで観測している/16 のセンサで収集されたパケットで構成されている．

本研究では 2016 年 9 月にソースコードが公開された Mirai に対して，その前後でダークネットにおいて観測されたパケットにどのような動向の変化があったのかを調べる目的として，2016 年 7 月 1 日から 2016 年 9 月 30 日の 3 ヶ月間おいてダークネット上で観測された TCP/SYN パケットに対して，最低支持数 1000(最低ホスト数 1000)，最低確信度 90%の条件で TCP と IP の各ヘッダ情報に対して頻出パターン解析および相関ルール学習を行った．

その結果，TCP ヘッダ情報に含まれるウィンドウサイズ (TCP Window Size)，宛先ポート (Destination Port)，存続時間 (Time to Live)，IP ヘッダ情報に含まれるパケットの優先度 (Type of Service)，の 4 項目に対して相関ルールが出現した．

この中から一定期間同じ相関ルールのみが観測されたウィンドウサイズ，宛先ポート，パケットの優先度，において 3 ヶ月間の観測実験により得られた結果について，それぞれをまとめる．

また，本研究が“Mirai”のソースコードが公開された前後の期間に焦点を当てたものである．よって，Mirai と得られた相関ルールがどのような関係を持つか調べるため追加の実験を行った．強い相関ルールと Mirai のと特徴との照合を見るために初期の Mirai の特徴として，公開されたソースコードに記されている 3 つの条件を利用する．

条件 1　シーケンス番号 = 宛先 IP アドレス,

条件 2　宛先ポート番号 = 23,

条件 3　送信元ポート番号 > 1024.

以上の条件と得られた相関ルールを組み合わせるため，まず相関ルールとして出現した値をそれぞれ少なくとも 1 回ずつ使用したホストを送信元 IP アドレスから抽出し，“ターゲットホスト”と定義する．このターゲットホストについて，送信したパケットがソースコードの 3 つの条件を全て満たす割合を求める．この割合が 90 %を超えたホストを“Mirai 感染ホスト”と定義し，ターゲットホストに対して Mirai 感染ホストがどれだけの割合で存在しているかを測定する．

4.1 ウィンドウサイズに関する相関ルール

まず，ウィンドウサイズについて実験結果を報告する．この要素については 8 月 2 日から 9 月 6 日までの約 1 ヶ月間強い相関を持つルールを複数獲得することができた．その中でも特にホストの数が非常に多い，上位 3 つのルール

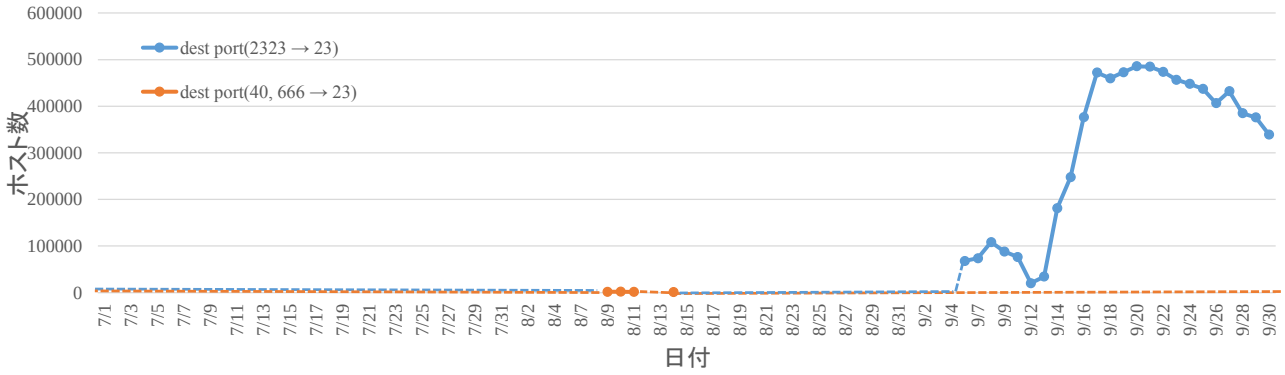

図 3　宛先ポート番号の相関ルール

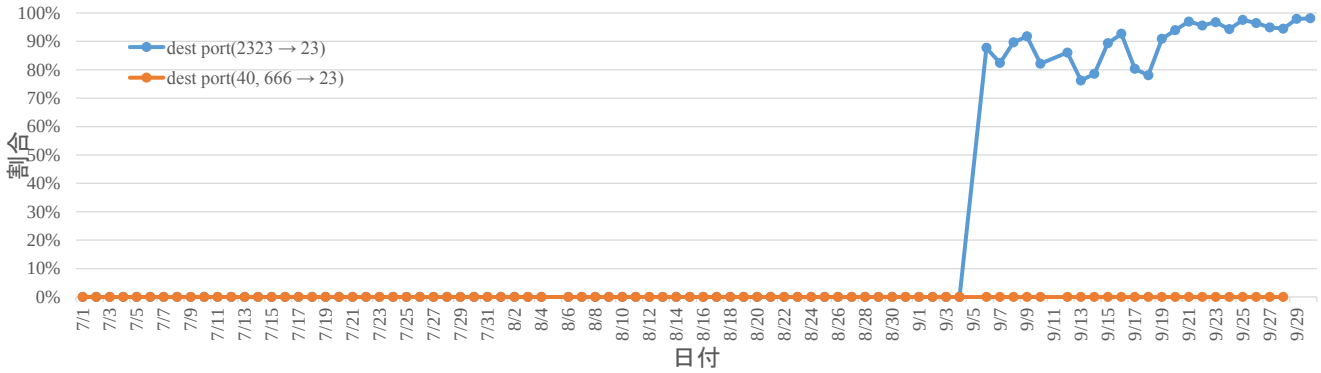

図 4　宛先ポート相関ルールの Mirai の特徴との照合

を図 1 に示す．

この結果からピーク時の 8 月 30 日には 40000 近いホストが (1320，2376 → 792) のルールに従ってスキャン活動を行っていたことがわかる．この (1320，2376 → 792) のルールは 1320，2376 のウィンドウサイズを使用したホストは 90%以上の確率で 792 のウィンドウサイズを使用している．ということを表している．

図 1 の 3 つのグラフより，ウィンドウサイズ (1320，2376，4488 → 792) 相関ルールが最も注目すべき特徴として得られた．以上の 4 つのルールについて Mirai の特徴との照合を行った．ターゲットホストは (792，1320，2376，4488) の全てのウィンドウサイズをそれぞれ一度でも使用したホストが対象となる．

照合の結果が図 2 である．この結果から，これらのウィンドウサイズの相関ルールは Mirai の初期における特徴に当てはまっていることが確認できる．

4.2　宛先ポートに関する相関ルール

宛先ポートについては既に先行研究がなされている [1] が，今回の実験では (40, 666 → 23) と (2323 → 23) の相関ルールが観測された．結果を図 3 に示す．(40, 666 → 23) のルールはピーク時約 2000 ホスト，(2323 → 23) のルールはピーク時約 500000 ホスト観測された．

これらに対しても Mirai の特徴との照合を行った．図 4 から (40, 666 → 23) のルールは Mirai との関連性は確認できなかったが (2323 → 23) のルールは Mirai の特徴に当てはまっているといえる．

4.3　パケット優先度に関する相関ルール

最後にパケットの優先度 (ToS) において得られた相関ルールを説明する．ToS については，今回の実験では (204，208，2012 → 2016) と (24，184 → 0) の相関ルールが観測された．

結果を図 5 に示す．(204，208，212 → 216)，(24，184 → 0) はそれぞれピーク時 1400 ホスト近く得られた．

このルールに対して Mirai の特徴との照合を行ったところ図 6 より，(204，208，212 → 216) は強い関連性は確認できなかったが，(24，184 → 0) のルールは 9 月の後半は 100 %に近い値となっている．宛先ポート (2323 → 23) の結果と近い形になっている．

4.4　まとめ

実験結果から，宛先ポート (2323 → 23) と，ToS(24，184 → 0) は Mirai のスキャン活動の特徴であり Mirai の動的な傾向を捕らえているといえる．9 月 17 日以降非常に多くのホストが上記二つのルールに従ってスキャン活動を行っ

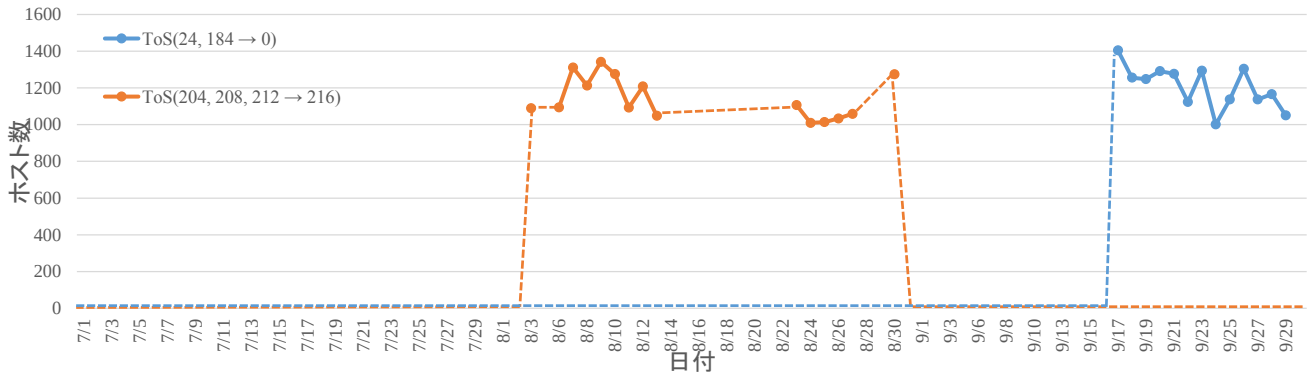

図 5　ToS の相関ルール

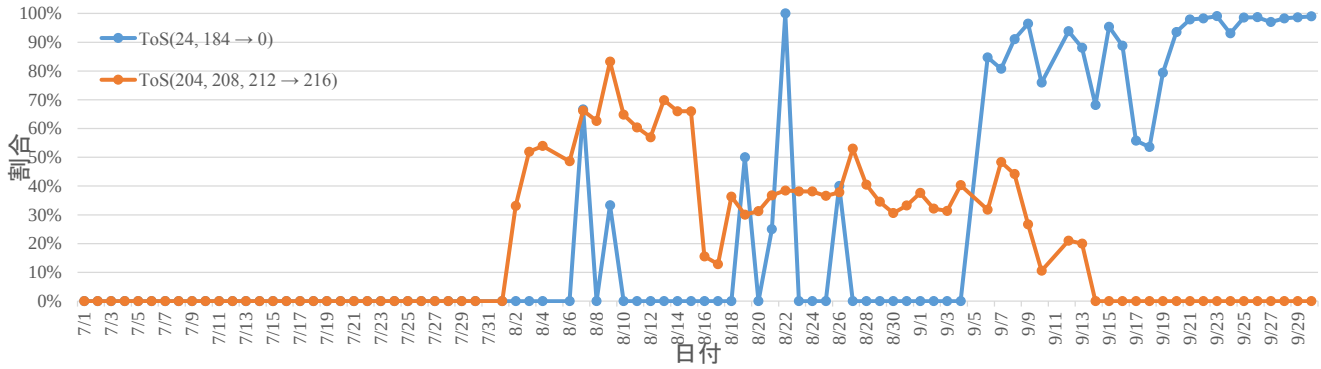

図 6　ToS 相関ルールの Mirai の特徴との照合

ており，さらに Mirai との照合の結果からほとんどのホストが高い割合で Mirai の特徴に当てはまっている．また宛先ポート (2323 → 23) に関してはソースコードにも示されており，また多くのセキュリティレポートなどでも Mirai の攻撃パターンとしてこの相関が報告されている．

一方でウィンドウサイズの相関ルールは 8 月 2 日から出現し，Mirai のソースコードが公開される 3 日前になくなっている．また，それらのホストの活動は Mirai の特徴と酷似している．されにそれらのウィンドウサイズの値は Windows OS が設定する値とは全く異なるものである．

これらの要因から今回得られた特徴は Mirai のプロトタイプの特徴であると推察できる．Windows OS ではない，いくつかの Linux ベースの機器 (IoT 機器) を利用してテストを行い，コード公開 3 日前から全く出現しなくなっている点においては，ソースコードの Mirai が完成し本格的な配布を始めたためと考えることができる．

また，公開後において相関ルールが全く出現していないことは，ソースコード内でランダムな値に設定されているためである．攻撃者が特定の機器による感染が広がっていることを隠すためにランダムな設定に変更を行ったものと推測できる．

5. おわりに

本稿では，大規模なダークネット観測網を用いた TCP スキャン攻撃の頻出パターンを解析し，相関ルールを抽出した．その結果，新しいヘッダ情報を用いて頻出なパターンを発見し，一定期間の傾向を調べることで，ウィンドウサイズ，宛先ポート，パケットの優先度におけるスキャンの特徴の変化を観測することができた．さらに，相関ルールに当てはまるホストが送信したパケットの特性から Mirai に感染した機器の送信パケットの特性と一致することを確認し，相関ルール学習の有用性を示した．

以上の結果から，今後相関ルール学習リアルタイムで行うことで，ソースコードの公開前にマルウェアの傾向を把握することができ，迅速にアラートを出すことでサイバー攻撃を未然に防ぐことができるのではないか考えられる．

謝辞　この研究は，文部科学省科学研究費補助金（B）16H02874 を受けたものである．

参考文献

[1] T. Ban, M. Eto, S. Guo, D. Inoue, K. Nakao, R. Huang, “A study on association rule mining of darknet big data,” *Proc. of International Joint Conference on Neural Networks*, pp. 1-7, 2015.

[2] C. Stocker, J. Horchert, “Mapping the internet: A

hacker's secret internet census," *Spiegel Online*, March 22, 2013.

[3] E.L. Malecot, D. Inoue, "The Carna botnet through the lens of a network telescope", In: J. Danger, et al.(eds), *Foundations and Practice of Security*, LNCS, vol 8352, Springer, pp. 426-441, 2014.

[4] J. HanJian, P.Y. Mao, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach," *Data Mining and Knowledge Discovery*, vol. 8, no. 1, pp. 53?87, 2004.

[5] R. Agrawal, T. Imielinski, A. Swami, "Mining association rules between sets of items in large databases," ACM SIGMOD Record, vol. 22, no. 2, pp. 207-216, 1993.

[6] J. Han, J. Pei, Y. Yin, "Mining frequent patterns without candidate generation," ACM SIGMOD Record, vol. 29, no. 2, pp. 1-12, 2000.

[7] C. Borgelt, "Frequent item set mining," *Data Mining Knowledge Discovery*, vol. 2, no. 6, pp. 437-456, 2012.