Vol.2021-DPS-186 No.49
Vol.2021-CSEC-92 No.49
2021/3/16

ダークネットにおける大規模調査パケットを考慮したポート番号埋め込みベクトルによるスキャンパケット解析

石川 真太郎[1]　中藤 大暉[1]　班 涛[2]　小澤 誠一[1]

概要：近年，IoT デバイスの脆弱性を利用したサイバー攻撃による被害が深刻になっており，対策が求められている．本研究では，機械学習を用いてダークネットで観測されたマルウェア感染デバイスによる攻撃と，その変化を追跡する手法を提案する．最初に，ダークネット観測において，ノイズとなっている大規模調査パケットを除外する．次に，FastText による特徴抽出を行い，スキャンパケットの宛先ポート番号から，ターゲットとなっているネットワークサービス間の相関関係を捉える．最後に，UMAP と DBSCAN を用いてホストの可視化と，同じ攻撃パターンを持つホストのクラスタリングを行い，マルウェアの傾向把握や新たなマルウェア亜種の出現の検知を行う．実験では，同手法で大規模調査パケットを削除した場合としない場合を比較し，大規模調査パケット削除の有効性を示し，また，既知の大規模調査を行う組織情報などを利用し大規模調査パケットを正しく分離できていることを示した．その上で 1 日ごとの解析情報を追跡することにより，宛先ポート番号だけに着目した解析では判断できない，ホストの時間的な活動情報に基づいたマルウェア判定を行うことができることを示した．

Scan Packet Analysis by Port-number Embedding Vector Considering Large-scale Survey Packets in Darknet

Abstract: In this research, we propose a method for tracking the attacks by malware-infected devices observed in the darknet and their changes using machine learning. First, we exclude large-scale survey packets that are noisy in darknet observations. Then, feature extraction using FastText is executed, and the correlation among targeted network services is captured from the destination port numbers of scan packets. Finally, UMAP and DBSCAN are used to cluster hosts with the same attack pattern as host visualization, to grasp malware trends and detect the emergence of new malware variants. In the experiment, we study the effectiveness of the proposed method where large-scale scanners are identified and ignored their traffic. By tracking the cluster transitions, we verify that the time transient of malware activity can be captured by tracking the portset clusters.

1. はじめに

近年，IoT の脆弱性を悪用するサイバー攻撃が増加している．マルウェアに侵入された IoT デバイスはボットネットを形成し，重要なインフラストラクチャにサイバー攻撃を実行することが問題となっており対策が急務となっている．有名な例として，IP カメラやルーターなどのデバイスを標的とする IoT マルウェアである Mirai [1] が 2016 年に発見された．Mirai は，TCP/SYN パケットをランダムな IP アドレスに送信して，実行中のサービスを探索するネットワークスキャンを実行し，デバイスに存在する脆弱性を悪用して侵入し，大規模なボットネットを形成する．ボットネット形成後，C&C サーバーの指示に従って，大量のパケットをターゲットサーバーに送信することにより，DDoS 攻撃が実行されるというマルウェアである．2016 年 9 月に GitHub で Mirai のソースコードが公開されたことで，ボットネットを悪用する攻撃が急増し，他の脆弱性を狙うように変更が加えられた多数の亜種が出現している．

Mirai の亜種やその他の新しいマルウェアによる被害を軽減するために，攻撃対象となっている脆弱性などの特徴的な情報を早期に得ることが重要である．そのために，サイバー攻撃を広い視野で観察できる，未使用の IP アドレス空間であるダークネットの使用が長年にわたって研究されている．ダークネットでは，マルウェアに感染したデバ

[1] 神戸大学
[2] 情報通信研究機構

イスと密接に関連しているパケットを収集できるが，近年，そのようなパケット以外にダークネットトラフィックの大規模な調査を目的としたパケットが多く観測されている．このようなパケットは，マルウェアの解析においてノイズとなるので，正確なマルウェアの傾向把握を行う上で取り除く必要がある．ダークネットセンサーでは，大規模調査パケットを除いたとしても，大量のパケットが観測される．膨大なネットワークトラフィックから，絶えず変化する攻撃の傾向を特定し新たなマルウェアを検出するには，機械学習が重要な役割を果たす．

機械学習を用いたダークネットトラフィック解析の一環として，石川ら [2], [3] は，テキストマイニング手法である FastText [4] を使用して，宛先ポート番号において共起する部分文字列の関係をモデル化するポート番号埋め込みベクトルを提案した．石川らの提案手法においても，調査目的のパケットがノイズとなり，正確なポート番号のベクトル表現が得られていないという問題がある．

本研究では，大規模調査パケットを除去することで，ノイズの影響を受けずにマルウェアに感染したホストからのポートスキャンの傾向を把握する手法を提案する．ポートスキャンに着目するため，ダークネットで観測された TCP/SYN パケットを収集して分析を行う．最初に，大規模調査パケットを投げるホストに見られる特徴から除去ルールを設け，そのルールにマッチするホストを除去する．続いて，IoT 関連のポートスキャンにおいて，（23/TCP, 2323/TCP）や（80/TCP, 8080/TCP）のような宛先ポート番号のペアが，関連するマルウェアの亜種の間で共通して見られる点に着目する [3]．この特徴は，ポート番号に紐づけられたサービスの関係を維持して，類似サービスに対して類似したポート番号を割り当てたいという人間の心理が働いた結果と考えられる．この特徴を利用することで，マルウェアの亜種間の類似性を宛先ポート番号間で共通するサブワードから推測でき，マルウェアの亜種間の関係をモデル化して評価することが可能となる．サブワードを考慮するために，宛先ポート番号に FastText を用いて，ポート番号のベクトル表現を作成する．また，ホストの可視化とクラスタリングを行うことで，大規模調査パケットによる影響を受けることなく，マルウェアの傾向把握や新たなマルウェア亜種の出現の検知を行う．

2. 提案手法

本節では，大規模調査パケットによる影響を受けることなく，マルウェアのスキャン活動を迅速かつ正確に把握する方法を提案する．ダークネット解析において，大規模調査パケットによるノイズの影響が大きくなっている．正確にマルウェアの傾向把握を行うために，適切に大規模調査パケットを除去するルールを設け，マルウェアによるトラフィックにのみ焦点を当てた解析を手法を提案する．大規模調査パケットを除去した後，ポート番号において頻繁に共起する部分文字列の関係に基づいた，ポート番号埋め込みベクトル表現を学習し，このベクトルにより，マルウェア亜種間の類似性に基づいたクラスタリングを試みる．

表 1 ダークネットセンサーで観測されたパケット年間統計

年	2015	2016	2017	2018	2019
#パケット ($\times 10^9$)	54.5	128.1	150.4	212.1	327.9
#IP アドレス ($\times 10^3$)	280	300	300	300	300

2.1 ポートセットの作成

マルウェアに感染したホストを類似したスキャン活動でグループ化するために，ダークネットで収集された TCP/SYN パケットから送信元 IP アドレスごとに宛先ポート番号を抽出し，これをポートセットと呼ぶ [3]．それぞれのポート番号にネットワークサービスが割り当てられているため，ダークネットで観察されるパケットの宛先ポート番号は，マルウェアの標的となるネットワークサービスを表す．従って，ポートセットはマルウェアに感染したホストによるスキャン活動を表現するデータとみなすことができる．作成したポートセットの類似性に基づいて，マルウェアに感染したホストのクラスタリングを試みる．

2.2 大規模調査パケットの除去

近年，ダークネットにおいて観測される調査目的のパケットが増加している．表 1 は，NICT が運用しているダークネットセンサーで観測されるパケット数を示したものである．2017 年から 2019 年にかけてのパケット数の増加は，主に海外組織からの調査目的とみられるスキャンの増加が主な原因である [5]．2019 年においては，約 1,750 億パケットが調査目的のスキャンとして判定されている．調査パケットと判定された IP アドレスは Shodan [6] や Open Port Statistics [7] などのセキュリティ関連組織のホストであるとわかっている．また，1 日あたり数千万から数億のスキャンパケットを送信する運用組織が不明なホストも観測されており，調査目的のスキャンが増加していることがわかっている．

このような大規模調査ホスト群から送られるパケットは，マルウェア活動の解析でノイズとなるため，適切に検知して除去する必要がある．そこで，本提案手法では以下の 2 つのルールを設け，大規模調査ホストの特定を行った．

1. Shodan の公開 IP アドレスから送信するホスト
2. ポートセットに 30 個以上のポートを含むホスト

マルウェアに感染したホストは，標的として定められた特定のポート番号にスキャンパケットを送信するが，大規模調査を目的とするホストは，多数のポート番号に対してスキャンパケットを送信する．そのため，マルウェアに感染したホストによるポートセットと，大規模調査を目的とし

たホストによるポートセットは，ポートセット内のポート数に違いが出ると考えられ，その閾値を経験的に 30 としてルールに反映した．このルールに従い，調査目的の IP アドレスから抽出されたポートセットを除外する．こうすることで，マルウェアによって生じるトラフィックにのみ焦点を当てた解析を行える．

2.3 ポート番号埋め込みベクトル

（23/TCP，2323/TCP）と（80/TCP，8080/TCP）は，関連するマルウェアの亜種間で共通して見られるポート番号である．これは，ポート番号に紐づけられたサービスの関係を維持して，類似サービスに対して類似したポート番号を割り当てたいという人間の心理が働いた結果と考えられる．この性質を利用することで，類似したマルウェアの亜種を同じグループに分類することが可能となる．ポートセットにテキストマイニング手法の一つである FastText を適用することで，宛先ポート番号のサブワードを考慮した埋め込みベクトルを取得できる．

図 1 に，提案手法であるポート番号埋め込みベクトルを取得する流れを示す．本手法では，ポートセットはドキュメント分析における文と見なされ，このポートセットに FastText を適用することで，各ポート番号が n-gram のサブワードに分割され，次に，n-gram サブワードと元の宛先ポート番号のベクトルが取得される．各宛先ポート番号とそれを構成する全てのサブワードの平均をとることで，ポート番号埋め込みベクトルを取得できる．

マルウェアに感染したホストによるポートセットと，大規模調査を目的としたホストによるポートセットは，ポートセット内に出現するポート番号やポート番号の数に違いが出る．そのため，マルウェアによるポートセットと大規模調査によるポートセット内に同じポート番号が出現したとしても，そのポート番号と他のポート番号の共起性が変わってくる．FastText は，ポートセット内のポート番号の共起性からベクトル表現を取得するため，このような違いにより適切なポート番号埋め込みベクトルが取得できなくなってしまう．適切なポート番号埋め込みベクトルを学習するために，大規模調査パケットによるポートセットを除いた学習データを用いる．最終的なポートセットのベクトルは，ポートセット内に出現する宛先ポート番号のポート番号埋め込みベクトルを平均化することで得られる．これにより得られるベクトル表現をポートセットベクトルと呼ぶ [3].

2.4 スキャン活動の可視化

スキャン活動のクラスタリングと新たなスキャン活動を検出をするために，ポートセットベクトルを 2 次元の埋め込み空間で可視化する．ポートセットベクトルの次元を削減する方法として，非線形次元削減法である UMAP [8] を使用する．UMAP では，最初に高次元空間のデータをグラフとして表し，次に，対応する低次元空間のグラフ構造が，高次元空間のグラフ構造にできるだけ類似するように最適化される．これにより，UMAP は高速で高性能の次元削減を実現している．

本手法では，マルウェアに感染したホストの分布を観測するために UMAP を使用し，ポート番号埋め込みベクトルがマルウェアによるスキャン活動を適切に反映できているかどうかを確認する．

2.5 スキャン活動のクラスタリング

ポートセットベクトルをクラスタリングすることにより，同じスキャンパターンを持つマルウェアに感染したホストのグループを自動的に識別する．クラスタリング手法は，密度ベースでクラスタリングを行う DBSCAN [9] を用いる．DBSCAN は 2 つのステップで構成される．まず，全てのデータ点を，指定された半径内に指定された数を超える隣接点を持つコア点，コア点から指定された半径内に存在するコア点同士を全て結んだネットワークに属する到達可能点，どちらにも当てはまらない外れ値に分ける．次に，コア点と到達可能点の集合をクラスタとすることでクラスタリングを行う．

DBSCAN は事前にクラスタ数を決定する必要がないため，日々特徴が変化しクラスタ数を定義することが困難であるダークネット分析に有効なクラスタリング手法である．ポートセットベクトルを DBSCAN に入力することで，類似したスキャンパターンを持つマルウェア感染ホストを自動的にクラスタリングすることができる．

2.6 詳細解析

本提案手法は，可視化やクラスタリングで得られた結果から，より有用な情報を得るために以下のような詳細解析を行う．

本提案手法は IoT 関連のポートスキャンにおいて見られる特徴に着目しているので，有名な IoT マルウェアである Mirai のスキャン活動に基づいた詳細解析を行う．クラスタリング結果と Mirai の特徴を持つポートセットの照合を行うために，公開されたソースコードに記されている以下の 2 つの条件を利用する．

1. シーケンス番号=宛先 IP アドレス
2. 送信元ポート番号 >1024
3. 宛先ポート番号=23

特定のホストから送信されたパケットの 90%以上がこの条件を満たすとき，このホストを Mirai 感染ホストとし，このホストから抽出されたポートセットを Mirai 感染ポートセットとする．この Mirai 感染ポートセットのクラスタリング結果から，Mirai 関連のクラスターの特定や，その周囲のポートセットから亜種の傾向把握，新たに出現した亜

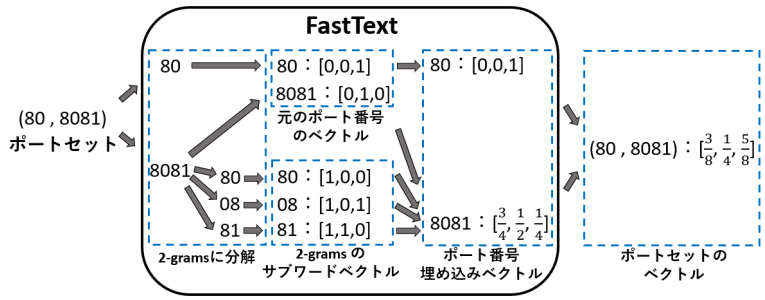


図 1 2-gram の FastText によるポート番号埋め込みベクトルの取得.

表 2 除外された大規模調査ホストの日次変化（観測期間：2020 年 2 月 10 日〜15 日）

	2/10	2/11	2/12	2/13	2/14	2/15
除外パケット数	507,472	537,045	567,991	555,938	545,280	510,805
除外ホスト数	872	976	992	951	1004	981
前日除外ホストとの重複数	-	785	889	864	872	880
重複率 (%)	-	80.4	89.6	90.9	86.9	89.7

種の検知を行う.

3. 評価実験

本節では，提案手法である大規模調査パケットによる影響を除去した，ポート番号埋め込みベクトルの有効性を評価する.

以下の実験では, 情報通信研究機構 (NICT) が運用する/16 ダークネットセンサーで観測された 2020 年 2 月 1 日から 2020 年 2 月 29 日までの 9,875,671,868 パケットを使用した.

3.1 大規模調査パケットの除去

大規模調査パケットを除去することによる，解析対象データへの影響を調査した．調査対象となるパケットデータには，2020 年 2 月 10 日から 2020 年 2 月 15 日に観測された TCP/SYN パケットを用いた.

1 日毎のパケットデータから，大規模調査を行うホストから送信されるパケットと，特定のホストから送信されるパケットの宛先ポート番号のユニーク数が 30 以上となる場合，そのホストから送信されるパケットを取り除いた.

まず，2 月 10 日から 2 月 15 日のパケットデータから，宛先ポート番号のユニーク数が 30 以上であるホストを取り除く．取り除かれたパケット内で，前日と重複しているホストの数とその割合を調べた結果を表 2 に示す．ただしデータの重複を削除しているため 1 ホストにつき 1 つのパケットデータが割り当てられている.

次に大規模調査を行うホストの情報を利用して，宛先ポート番号のユニーク数が 30 以上のパケットを除いたとき，そのホストがどの程度取り除けているのかを 2 月 10 日と 11 日のデータを用いて表 3 に示す.

次に大規模調査パケットを考慮した場合とそうでない場合を比較した．具体的には，2020 年 2 月 11 日のパケットデータを解析対象とし，学習モデルとしては 2 月 4 日から 10 日のパケットデータを使用し，情報通信研究機構が所有する 2 月 11 日における Mirai 感染ホストの宛先ポート番号の情報と照会することで可視化の結果を比較した．その結果を図 2 に示す。

また，図 2 に示した注目点付近のホスト群のうち，Mirai 感染ホストと感染不明ホストをランダムにサンプリングし，それらのポートセットを図 3 に示す．これからわかるように，Mirai シグネチャーにはマッチしなかったため感染が不明とされたホストのポートセットは，Mirai 感染が確認されているホストのポートセットに極めて類似している．さらに詳細な解析が必要と思われるが，感染不明ホスト群は Mirai 亜種に感染している可能性があると言える.

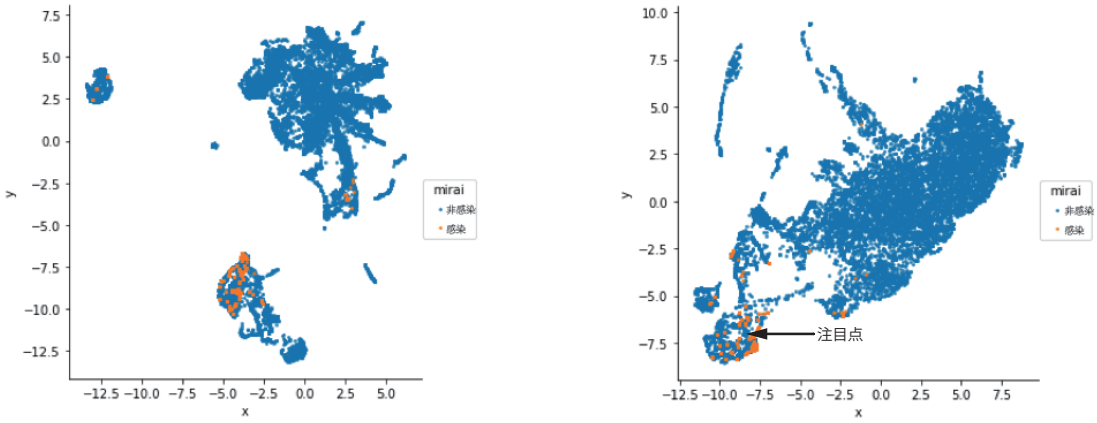
3.2 スキャン活動の追跡

スキャンパケットからマルウェアの傾向や変化を捉えるため，提案手法を用いてポートセットを可視化し，その変化を追跡した．またクラスタリング結果の変化についても追跡を行った．解析対象データとして，2020 年 2 月 10 日から 2020 年 2 月 15 日の TCP/SYN パケットを用いた．ここで FastText の学習データは，2020 年 2 月 4 日から 2 月 10 日のパケットデータを使用し，可視化を行うデータには 2020 年 2 月 11 日から 1 週間のパケットデータを使用した．その結果を図 4 に示す.

クラスター 1 について，クラスタリング内のポートセット数の追跡を行った．図 2 と図 4 において，2 月 11 日の Mirai 感染ポートセットの可視化結果とを比較すると，図 4 右上のクラスターに Mirai 感染ホストが位置していると考えられる．ここでは，シグネチャーがない，マルウェア感染が疑われるホスト群の解析を行いたいため，11 日に図 4 の右下に出現した緑色で示したクラスターの追跡を行った．このクラスター内のポートセットはマルウェアに

表 3　既知の大規模調査ホストに対する検知精度（観測期間：2020 年 2 月 10 日〜15 日）

データセット名	2/10	2/11	2/12	2/13	2/14	2/15
既知の大規模調査ホスト	23	22	21	21	21	20
検知ホスト数	20	20	20	20	20	19
再現率	0.87	0.91	0.91	0.91	0.91	0.95



(a) 除去前　(b) 除去後

図 2　大規模調査ホストの除去による Mirai 感染ホストクラスターの変化．オレンジ色の点が Mirai 感染ホスト．

Mirai感染

	port
0	23 80 60001
1	23 80 60001
2	23 80 60001
3	23 80 60001
4	23 80 60001
5	23 80 60001
6	23 80 60001
7	23 80 60001
8	23 80 60001
9	23 80 60001
10	23 80 60001
11	23 80 60001
12	23 80 60001
13	23 80 60001
14	23 80 60001
15	23 80 60001
16	23 80 60001
17	23 80 60001
18	23 80 60001
19	23 80 60001
20	23 80 60001
21	23 80 60001
22	23 80 60001
23	23 80 60001
24	23 80 60001
25	23 80 60001
26	23 80 60001
27	23 80 60001
28	23 80 60001
29	23 80 60001

不明

	port
0	80 2323 60001
1	23 5555 60001
2	80 2323 60001
3	80 2323 60001
4	23 5555 60001
5	2323 60001
6	80 2323 60001
7	80 81 2323
8	80 2323 60001
9	80 2323 60001
10	23 80 81 2323 8080 60001
11	23 8080 60001
12	80 2323 60001
13	23 80 8081 52869
14	80 2323 60001
15	80 2323 60001
16	2323 60001
17	80 2323 60001
18	80 2323 60001
19	2323 60001
20	80 2323 60001
21	80 2323 60001
22	80 2323 60001
23	2323 60001
24	80 2323 60001
25	80 2323 60001
26	80 2323 60001
27	80 2323 60001
28	80 2323 60001
29	80 81 2323

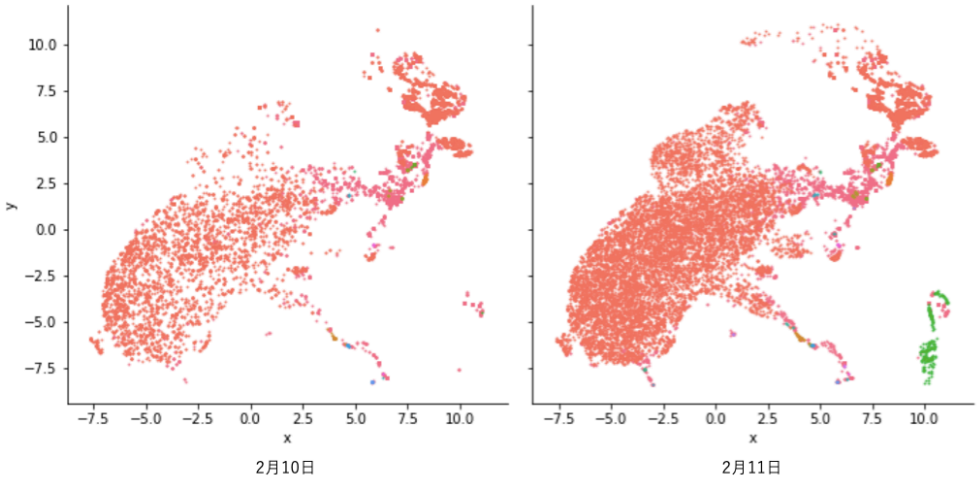
図 3　図 2 の注目点付近に含まれるホスト群のポートセット．

よるスキャンが多く見られる宛先ポート番号の 23/TCP，80/TCP，445/TCP などを含むポートセットが大多数を占めていた．その結果を図 5 に示す．ただし追跡対象のクラスターを“クラスター 1”，その他のクラスターを“他クラスター”と表記する．また（ ）内の数値はポートセット数を表している．

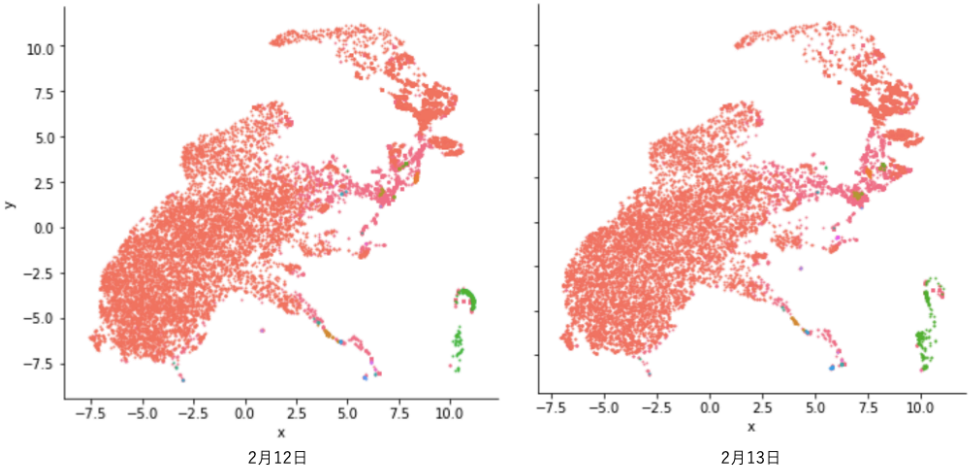
3.3　考察

まず、大規模調査パケットを取り除いた結果であるが，大規模調査のもう 1 つの特徴として継続的な活動が挙げられ，その条件を満たしさらに宛先ポート番号のユニーク数が 30 以上であるという条件を満たせば，大規模調査パケットである可能性が高いと考えられる．結果は表 2 のとおり取り除いたホストの内少なくとも 80 %以上のホストが継続的な活動をしていることがわかる．よってこの手法により大規模調査パケットをデータセットから削減できている．また表 3 の結果も大規模調査パケットの 9 割程度を削減できていることから，この手法により大規模調査パケットの多くを削減できていると考えられる．

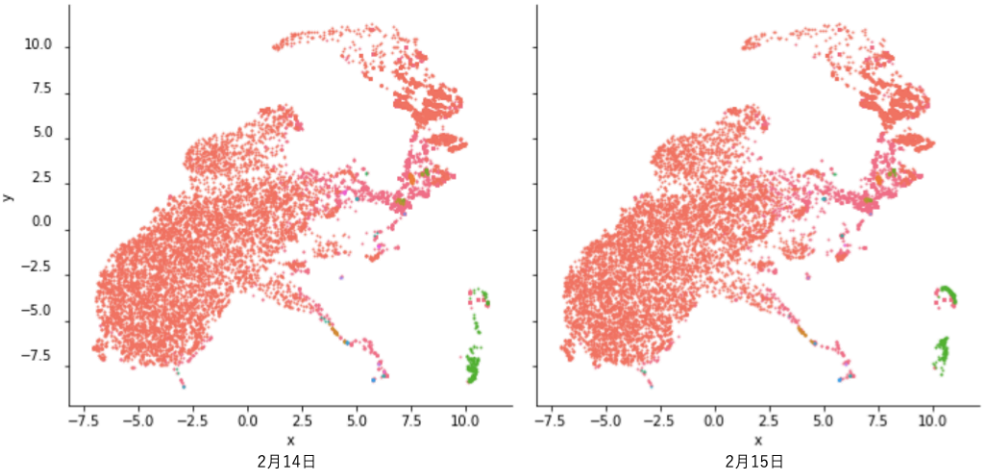
次に大規模調査パケット削除前後の比較の結果，図 2 が示すように，先行研究の手法では次元削減後の可視化ではあるが，主な MIrai 感染ホストのクラスターから離れた位置に配置されていた Mirai 感染ホストが，大規模調査パケット削減後では，比較的近い位置に配置されている．このことから次元削減前のベクトル空間においても距離が近くなっていると考えられ，Mirai 感染ホストに対するベク



(a) 10 日〜11 日



(b) 12 日〜13 日



(c) 14 日〜15 日

図 4 可視化結果

トル付与の精度が向上していると考えられる．しかし全体的なクラスターのまとまりが損なわれているように見受けられるため，さらに改善の余地はあるように思われる．

最後にクラスターの追跡を行ったが，10 日から 11 日に移ると現れた対象のクラスターは，その後も継続的に類似宛先ポートのパケットを送信し続けていることが表 5 よりわかる．表 5 で他クラスターに分類されたパケットを調べた結果，対象クラスターの近傍に位置する桃色で表されて

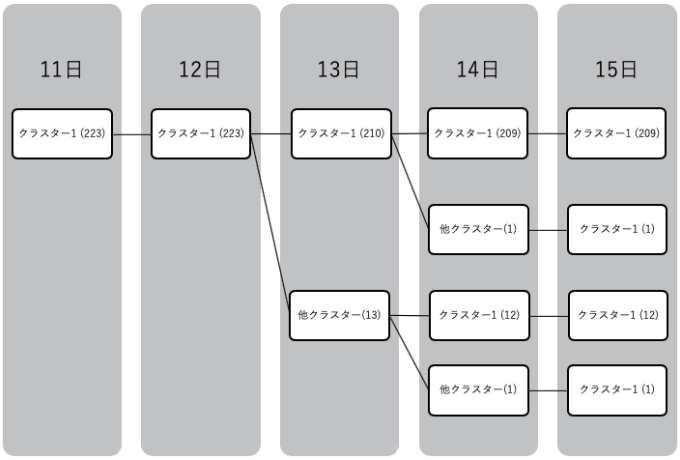


図 5 クラスター内ホストの追跡

いるパケットであることがわかった．このことから，マルウェア感染ホストと疑われたが，実際には，マルウェアがスキャンするポート番号に焦点を当ててパケットを送信している組織による調査パケットと推測される．このように日次でのクラスター解析の推移を観察することで，マルウェアの活動と組織による調査活動を区別できることがわかった．

4. 結論

本研究では，ダークネットで観測されるパケットの宛先ポート番号に着目してマルウェアの活動を解析する手法を提案した．具体的には，高精度な解析を行うため，大規模調査ホストを特定して，その影響を排除し，マルウェアによる感染が疑われるホスト群のクラスターを追跡して解析する手法を提案した．

近年マルウェアの増加，多様化に対し多くの解析が必要になる反面，解析目的でダークネットに本来存在しない多くのパケットが送信されることで，宛先ポート番号による特徴抽出とクラスタリングにおいて，大規模調査パケットの情報がマルウェア感染ホスト特徴に影響してしまうという問題が生じていた．大規模調査パケットを考慮した解析は，大規模調査パケットの宛先ポート番号のユニーク数に注目しユニーク数が 30 以上のものを除外したが結果は実験からもわかるとおり，効果的なデータの絞り込みができ上記の問題を軽減できた．

また，ダークネットセンサにより収集された TCP/SYN パケットから宛先ポート番号を抽出して解析を行なったが，宛先ポート番号からマルウェア感染ホストの特徴ベクトルを作成し，クラスタリングした後，クラスターの追跡を行うことでマルウェア感染ホストと大規模調査パケットの活動の違いを観測できたことで，判別に有効な１つの特徴を捉えられたと考える．

また時間軸での追跡は大規模調査パケットとマルウェア感染ホストの判別だけでなく，マルウェアの解析にも効果的であり，解析結果からマルウェア感染ホストのスキャン活動を把握することが可能となる．未知のマルウェアなどに対応するためには特徴をとらえる必要があるが，その点マルウェア感染ホストによるスキャンパケットと大規模調査パケットの時間的な活動間隔やパケット数には大きな違いがあると考えられるので時間的特徴を解析することは非常に有効な手段である．

今後の課題としては，マルウェア感染ホストの時間的なスキャン間隔を継続的に調査し，その特徴を把握できるシステムの構築がある．

謝辞

本研究は，科研費基盤研究（B）（課題番号 16H02874）の助成を受けたものである．

参考文献

[1] Kolias, C., Kambourakis, G., Erhan., Stavrou, A., Voas, J.: DDoS in the IoT: Mirai and Other Botnets. *IEEE Computer*. 50, 80 - 84 (2017)
[2] Ishikawa, S., Ozawa, S., Ban, T.: Port-Piece Embedding for Darknet Traffic Features and Clustering of Scan Attacks. *Neural Information Processing. (ICONIP 2020) . Lecture Notes in Computer Science*. 12533, 593-603 (2020)
[3] 石川 真太郎，小澤 誠一，班 涛: ポート番号埋め込みベクトルを用いたダークネットスキャンパケット解析. コンピュータセキュリティシンポジウム 2020 論文集, 1010-1016 (2020)
[4] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*. 5, 135-146 (2017)
[5] National Institute of Information and Communications Technology. NICTER Observation report 2019, Retrieved June 20, 2020, from https://www.nict.go.jp/cyber/report/NICTER_report_2019.pdf
[6] The search engine for the Internet of Things. https://www.shodan.io/.
[7] Open Port Statistics. http://openportstats.com/.
[8] McInnes, L., Healy, J., Melville, J.: UMAP: Uniform

Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802. 03426* (2018)

[9] Ester, M., Kriegel, P.H., Sander, J., Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining,*. 226-231 (1996)

[10] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781* (2013)

[11] Ramos, J.: Using TF-IDF to Determine Word Relevance in Document Queries (2003)

[12] Mikolov, T., Sutskever. I , Chen. K , Corrado. G , Dean. J.:Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*. 26 (2013)