時空間地理情報を用いたマルウェア配布元の傾向分析

岩崎信也† 山口崇志‡ 布広永示‡

†東京情報大学大学院
265-8501 千葉県千葉市若葉区御成台 4-1
tyamagu@rsch.tuis.ac.jp

‡東京情報大学総合情報学部
265-8501 千葉県千葉市若葉区御成台 4-1
tyamagu@rsch.tuis.ac.jp

あらまし サイバー攻撃の種類は多様になり，マルウェアにおいては常に改良が行われセキュリティにおける脅威となっている．またマルウェア配布元の大多数は海外にあることがわかっている．マルウェア配布元の特徴や傾向を掴むことはサイバー攻撃に対する対処として重要である．本研究ではCCC DATAset 2009-2011の3年間のデータにおけるマルウェア配布元の地理空間情報をヒートマップにより可視化することで直感的に配布元の地理空間情報の分布・密集具合を掴み，長期的な時間的推移を考慮した配布元の傾向分析を行った．

Trend analysis using spatio-temporal geographical information of the Malware distribution

Shinya Iwasaki† Takashi Yamaguchi‡ Eiji Nunohiro‡

†Tokyo University of Information Sciences， Graduate School of Informatics.
4-1 Onaridai， Wakaba-ku，Chiba， 265-8501 Japan
tyamagu@rsch.tuis.ac.jp
‡Tokyo University of Information Sciences， Department of Informatics.
4-1 Onaridai， Wakaba-ku，Chiba， 265-8501 Japan
tyamagu@rsch.tuis.ac.jp

Abstract Recently, the risk of Cyber-attacks is increased because the technologies of attacks or malware are diversified and complicated.
Therefore, it is important to clarify the characteristics of malware and the distributor for the appropriate security measures.
In this paper, we investigate the spatio-temporal changes of malware distributor on the heat-maps that is a visualization of kernel density estimation from there CCC data set of 2009 to 2011.

1　はじめに

近年，サイバー攻撃は様々な形で多様化かつ増加の一歩を辿っておりセキュリティ上の脅威となっている [1]. サイバー攻撃においてはマルウェアが利用されることが多く，様々なマルウェアが日々作成・改良され猛威を振っておりその配布元の大多数が日本国外であり地理的な関連があることがわかっている. このマルウェア配布元を分析することはマルウェアの感染の広がりや感染限の特定，未知のマルウェアに対する早期発見が行えるとされている [2]. またマルウェア配布元の可視化においては，複数のマルウェアを配布しているマルウェア配布元の可視化[3]や地理的可視化を用いたマルウェアの統合解析 [4]などの既存研究が行われている. しかしながら既存研究における可視化においては短期的な時間的推移の分析や分布の可視化に留まり，長期的な時間的推移の分析や密度の可視化等は行われていない.

本研究では CCC DATAset 2009-2011 [5] の 3 年間のマルウェア配布元データにおける地理的状況の分布・密集地帯の実態を明確にした上で時間的推移を考慮した傾向分析を行った. このためにマルウェア配布元の地理空間情報を密度推定における可視化手法の一つであるヒートマップとして可視化するシステムを実装した. ヒートマップにより可視化を行うことで直感的な密度の実態把握を行える.

2　提案手法

本研究ではマルウェア配布元の地理空間情報を直感的に理解しやすい形で可視化することで長期的なマルウェア配布元の傾向を分析する. 具体的には 2008 年から 2010 年の 3 年のマルウェア配布元の分布や密度における時間的推移を分析した. このためにマルウェア配布元の地理空間情報を密度推定における可視化手法の一つであるヒートマップとして可視化するシステムを実装した.

2.1　地理空間情報の可視化

地理空間情報とは住所や座標等の位置情報に関連づけられた様々な情報を指す [6]. 地理空間情報の可視化には一般的に地理情報システム(GIS) が利用される. 地理空間情報の可視化によって地理に紐づけられたデータを容易に理解し解釈することが可能となる. 近年では GoogleMap [7]などの Web 地図サービスの普及によって地理空間情報の利用促進が行われた. 同時にセキュリティ分野においても様々な形で地理空間情報の可視化が行われている.

2.1.1　GIS

GIS は地理空間情報を含む様々な情報を作成，管理，分析，可視化，共有するためのシステムである. GIS によって地理的問題発見や課題解決へのアプローチなどの検討が可能である [8]. GIS の種類は大きく検索や可視化のみに特化した地図サービスと解析や分析・管理等も可能な高機能 GIS の二種類に分けられる.

地図サービスは地図を閲覧・検索することに特化し容易に利用可能なシステムである. そのため Web システムであることが多く，汎用的な地図サービスとして図 1 の GoogleMap [7]などがあげられる. また東京都の風速情報を可視化した東京風速 [9]のように防災情報や気象現象・インターネット通信等の一部の用途に特化したサービスもある.

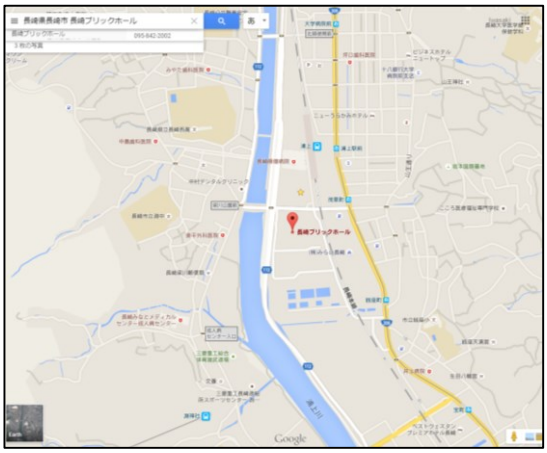

図 1　汎用的な地図サービスの例
GoogleMap [7]

高機能 GIS はデータの管理や高度解析が可能であり，気象現象等の様々な複雑なデータの分析が可能であり専門家に利用され図 2 の ArcGIS for Desktop [10]や QGIS [11]があげられる.

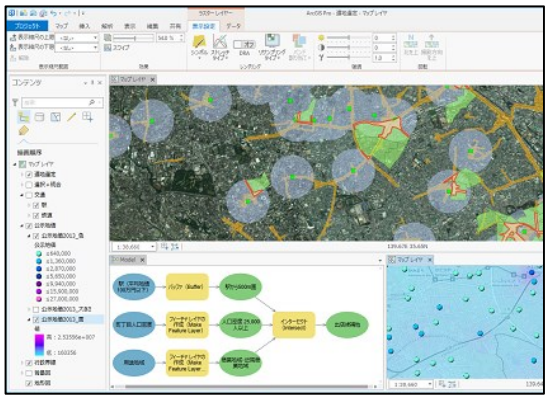

図 2 高機能 GIS の例
ArcGIS for Disktop [10]

また，近年では高機能な端末の普及によって高機能 GIS を Web システムとした Web GIS の提供が進んでいる.

2.2 可視化手法

地理空間情報の可視化手法には様々な手法がある. 地理空間情報には大きく分けて特定の地点を指すポイントデータと特定の範囲を示すテクスチャデータがあり分析をする際は容易さや手法への適応のし易さを考慮しポイントデータが利用されることが多い. ポイントデータの可視化手法で比較的利用される手法としてはポイントマップを基本に，ポイントデータの集計や分布等を行った上で可視化するヒートマップ，クラスタマップなどがある.

2.2.1 ポイントマップ

ポイントマップはポイントデータを可視化する際の位置情報において地図上にポイントをプロットすることで可視化する最も単純な手法である. ポイントマップでは色や形等を変更することでポイントごとのデータの違いを表現することが可能であり，詳しい位置情報等を適切に理解することが可能である. しかしながらポイントデータが多数になるとポイントデータの分布の傾向などを大域的に理解するのは難しいとされる.

2.2.2 ヒートマップ

ヒートマップはカーネル密度推定を利用しポイントやデータの密集具合によって色を塗り分けて表現する手法である [12]. ヒートマップではカーネル密度推定を利用することでポイントデータにおける外挿が可能となり密度の特定の地理空間情報に対しての大域的な地理的分布を容易かつ直感的に理解することが可能である. 図 3 はヒートマップの例であり関東圏の避難場所の密集具合を色によって表現している. 本研究ではマルウェア配布元の大域的な密度傾向等の実態を把握するためヒートマップを利用する.

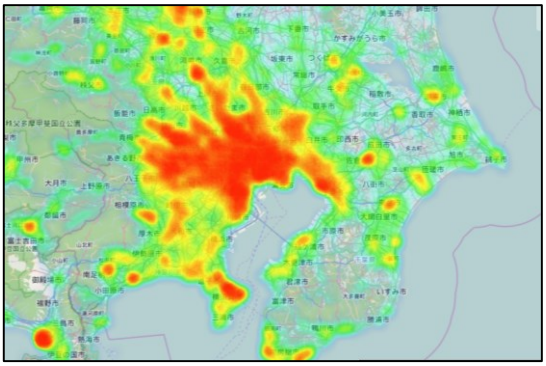

図 3 ヒートマップの例
関東圏の避難場所のヒートマップ

また同様の密度推定を利用するマップとしてプリズムマップやクラスタマップがある. プリズムマップは行政区画や特定のエリアごとにポリゴンを作成し，ポイント内のポイントデータの密度によって高さや色を表現した手法でありヒストグラムの理論を応用している. クラスタマップは周辺のポイントの集計を行う簡略化した可視化手法である. 密集している位置を中心とし周辺のポイントデータの数を大きさにした円で表現し可視化する.

2.3 セキュリティにおける地理空間情報

サイバーセキュリティにおける地理空間情報はサイバー攻撃などのインターネット通信の通信元や通信先の分析の際に利用されており通信の

可視化を行う地図サービスなども公開されている. 例としてNORSEのNorse Attack Map [13]や情報通信機構のNicter [14], カスペルスキー社のCYBERTHREAT(図 4)などがあげられる.

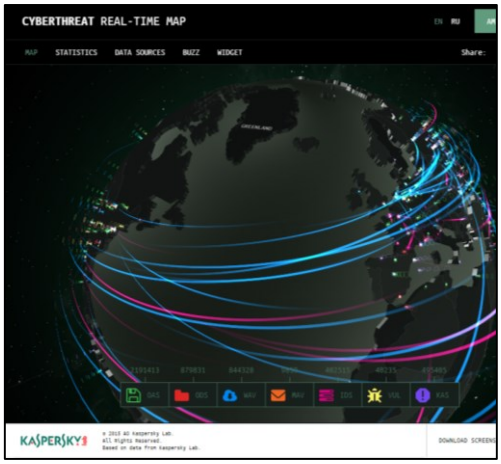

図 4 サイバー攻撃の可視化例
CYBERTHREAT REAL-TIME MAP [15]

2.3.1 IP アドレスと地理空間情報の紐づけ

インターネット通信の地図上への可視化の際には一般的に IP アドレスを地理空間情報に変換する必要がある. IP アドレスと地理空間情報の対応データベースは無料のサービスから有料のサービスまで様々あるが本研究では無料かつ全域を比較的網羅しているGeoLite2 [16]を利用した.

2.4 カーネル密度推定

ヒートマップ等の密度の可視化を行う際には密度推定が必要になり手法の一つにカーネル密度推定がある [17]. カーネル密度推定とはデータの標本に対して外挿を行う推定手法である. データの標本値をぼやかして確率密度関数を求めることができ他の密度推定手法であるヒストグラムに比べて領域内を連続で推定することが可能であり, かつ領域の分割を行う必要がないため密度推定の際は利用されることが多い. 地理空間におけるカーネル密度推定を利用した可視化手法の一つがヒートマップである.

カーネル密度推定における確率密度 p(x)は n 個の標本を式(1)で示すようなベクトルの集合とした時, 式(2)で導かれる.

$$\mathbf{x}_i = \{x_1, x_2, \ldots, x_n\} \tag{1}$$

$$p(\mathbf{x}) = \frac{1}{n \cdot h} \sum_{i=1}^{n} K\left(\frac{\|\mathbf{x} - \mathbf{x}_i\|}{h}\right) \tag{2}$$

$K(x)$は標準的なガウス関数を利用することが多く本手法でもガウス関数を利用した. $\|\mathbf{x} - \mathbf{x}_i\|$はベクトル間の距離を示し, ユークリッド距離を用いた. hはバンド幅であり大きく設定すれば広域的な傾向が, 小さく設定すれば局所的な傾向が抽出できる. バンド幅の値については標本データにより最適な値が異なるため複数の関数や値を試すとよいとされる. またバンド幅においてはすべての標本において固定値とする固定カーネル密度推定と標本の周辺の密度によって変更する可変カーネル密度推定がある. 一般的にヒートマップのバンド幅では固定カーネル密度推定を利用する.

3 実装

本研究では CCC DATAset のマルウェア配布元データの地理空間情報をヒートマップとして可視化するシステムを実装する. 可視化の手順は図 5 のとおりである.

ヒートマップの可視化においては前研究である統合的地理情報解析プラットフォーム(Open Gaia System)を拡張する形で実装した [18]. Open Gaia System は様々な時空間地理情報を Web 上で統合的に収集・管理・解析・可視化・共有が可能な Web GIS であり簡単な解析やや地図上での可視化, 時系列ごとの地理情報の表示・操作等が可能である. 本研究ではこのシステムを拡張しカーネル密度推定処理やヒートマップによる可視化処理を実装した. この上でカーネル密度推定におけるバンド幅値をヒートマップを確認しながら変更できるようすることで容易な解析を可能とした.

Open Gaia System は通常 Web システムで

あるが本研究では CCC DATAset の利用規定に基づきローカルシステムとして実装した.

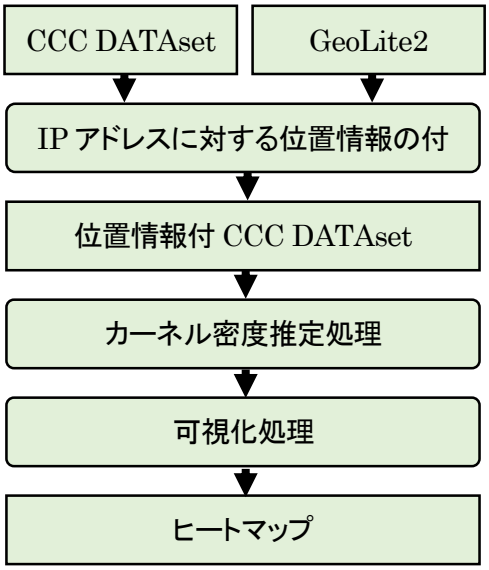

図 5 ヒートマップによる可視化の手順

4 検証

提案手法により CCC DATAset 2009-2011 のマルウェア配布元のデータを 1 か月毎にヒートマップ化し傾向を分析する. 本稿ではそのうち長期的な時系列遷移の実態を可視化するため 008 年 12 月, 2009 年 12 月, 2010 年 12 月のデータを利用した. CCC DATAset のデータカラムは表 1 となる.

表 1 CCC DATAset 2009-2011 のカラム

項目	例(一部マスク)
取得時刻	2009-04-01 00:01:58
送信元 IP アドレス	honey035
送信元ポート番号	1034
宛先 IP アドレス	**.215.1.206
宛先ポート番号	80
TCPorUDP	TCP
検体ハッシュ値	**8af718797c91
検体名称	WORM_**.CZU
ファイル名	C:¥**¥ptkj.exe

本研究ではこの内取得時刻及び配布元 IP アドレスである宛先 IP アドレスのみを利用した. また本稿で利用するデータのデータ件数が表 2 である.

表 2 CCC DATAset の配布元のデータ件数

データ取得日付	件数
2008 年 12 月	147338 件
2009 年 12 月	116674 件
2010 年 12 月	17418 件

ヒートマップ化し地球全域を可視化したものの例が図 6～図 8 である. バンド幅においては複数の値を試行し 391km とした。

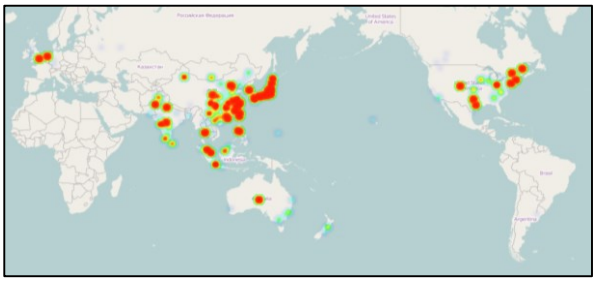

図 6 マルウェア配布元のヒートマップ

2008 年 12 月

図 7 マルウェア配布元のヒートマップ

2009 年 12 月

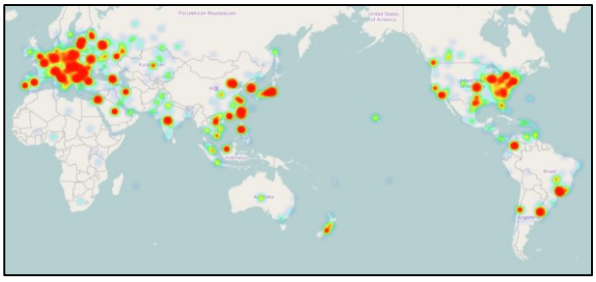

図 8 マルウェア配布元のヒートマップ

2010 年 12 月

それぞれ 2008 年 12 月, 2009 年 12 月, 2010 年 12 月のマルウェア配布元のヒートマップでありマルウェア配布元の密集度, 分布が変化していることが直感的にわかる. 2008 年 12 月では

東アジアを中心としていたマルウェア配布元が 2010 年 12 月にはヨーロッパ，アメリカが中心に変化した．これは 2008 年 12 月においてはマルウェアの配布サーバが東アジアに用意されマルウェアを配布していたのに対し，2010 年になるにつれ正規の Web サイトやルータなどがマルウェアに感染し再配布元になり増加したのではないかと考えられる．また南米においては 2008 年 12 月に配布元がなかったのに対し 2010 年 12 月には配布元の密集地帯が存在している．

5 まとめ

本研究では，マルウェア配布元をヒートマップによる可視化し時間的推移を考慮した人間による直感的な理解，傾向分析を行った．ヒートマップにより可視化することでマルウェア配布元の直感的な分布や密集地帯の傾向・特徴を把握することが可能であることがわかった．

今後としてはマルウェアごとの配布元の分布変化の分析や他のインフラ普及率や人口データ等の地理空間情報との組み合わせによる傾向の分析を検討している．

6 謝辞

本研究では MWS Datasets の内 CCC DATAset を利用しています．提供していただいた CCC 運営連絡会及び関係機関の皆様には深く感謝いたします．また本研究では MAXMIND 社の GeoLite2 データベースを利用しました．

参照文献

[1] 警視庁，“平成 26 年中のサイバー空間をめぐる脅威の情勢について，” http://www.npa.go.jp/kanbou/cybersecurity/H26_jousei.pdf.

[2] 大井俊介，“今後脅威となりうるマルウェア配布元ホストの早期発見に関する一考察，” WMS2009, 2009.

[3] 松木隆宏，“CCC DATAset 2009 によるマルウェア配布元の可視化，” CSS2009, 2009.

[4] 金子博一，“地理的可視化を用いたマルウェアの統合解析，” Computer Security Symposium 2011, 2011.

[5] 畑．充弘，“マルウェア対策のための研究用データセットとワークショップを通じた研究成果の共有，” MWS2009, 2009.

[6] 柴崎亮介，地理空間情報活用推進基本法入門，日本加除出版, 2008.

[7] Google ，“Google Map，” https://www.google.co.jp/maps.

[8] ESRI，“GIS（地理情報システム）とは，” http://www.esrij.com/getting-started/what-is-gis/.

[9] C. Beccario，“東京風速，” http://air.nullschool.net/.

[10] ESRI，“ArcGIS for Desktop，” http://www.esrij.com/products/arcgis-for-desktop/.

[11] QGIS，“QGIS，”http://qgis.org/ja/site/.

[12] R. Maciejewski, S. Rudolph , R. Hafen，“A Visual Analytics Approach to Understanding Spatiotemporal Hotspots，” Purdue University, 2010.

[13] Norse，“Norse Attack Map，” http://map.norsecorp.com/.

[14] 情報通信研究機構，“NICTERWEB，” http://www.nicter.jp/nw_public/scripts/index.php.

[15] Kaspersky，“Cyberthreat real-time map，” https://cybermap.kaspersky.com/.

[16] MAXMIND，“GeoLite2，” https://dev.maxmind.com/ja/geolite2/.

[17] O. Lampe , H. Hauser，“Interactive visualization of streaming data with kernel density estimation，” Pacific Visualization Symposium, 2011.

[18] 岩崎信也，“Knowledge presentation using geographical maps on Web GIS, 20th International Symposium on，” AROB2015, 2015.