社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－CSEC－27　(10)
2004／12／20

ウイルスの拡散過程と感染国数の関係について

小　泉　芳[1]　　小池英樹[2]　　安村　通晃[3]

慶應義塾大学　大学院政策・メディア研究科[1]　電気通信大学　大学院情報システム学研究科[2]　慶應義塾大学　環境情報学部[3]

本研究では、ウイルスの拡散過程において、感染国数に注目した解析を行う. 特にIPアドレスを基にターゲットを選ぶウイルスを対象として, ターゲット選択方法が地理的な拡散にどのような影響を及ぼしているかをシミュレーションにより分析する. また感染国数を急増させることを目的とした新しいターゲット選択方法であるハブリストスキャンの効果についても述べる.

The Relationship Between Virus Spread Process and The Infected Number of Countries

Kanba KOIZUMI[1] , Hideki KOIKE[2]
and Michiaki YASUMURA[3]

Graduate School of Medea and Governance, Keio University[1] Graduate School of Information Systems, University of Electro-Communications[2] Faculty of Environmental Information, Keio University[3]

In this paper, we analysis the relationship betweeen viruses spread process and infected number of countries. We investigates the dynamism of geographical spread of viruses by target selection mechanism. We also investigates the effect of new scan method, named hub list scan.

1. はじめに

コンピュータウイルスの対策は、ウイルスの発生後の事後解析に追われているのが現状である. 従来研究においてネットワークサービスのセキュリティホールを狙う自己増殖型ウイルス（ワーム）の解析は主に感染ホスト数 (IP アドレス数) により、増殖過程の分析が行われている. これは多くの感染ホストを獲得することが主なウイルスの挙動であると考えられているからである. しかし、ウイルス作成者によっては、感染ホスト数よりも感染国数に着目する可能性がある. これは多くの国（地域）への地理的な拡散を目的としたウイルスである. 本研究ではこのような地域的な拡散を目的としたコンセプトウイルスを想定し、その拡散過程の解明を行いウイルス対策への先行的な情報提供を目的としている.

本研究では特に感染国数に注目し、IP アドレスを基に選ぶウイルスのターゲット選択方法が, 地理的な拡散にどのような影響があるかを分析する. 当初、サイバー空間（IP アドレス）と地理空間（IP 保有国）の関係について統計的な特徴抽出をする. 次にその統計的特徴及びウイルスのターゲット選択方法が感染国数に与える影響についてシミュレーションによる分析をする. 特に多くの国が密集する上位 16 ビットごとの集合であるハブサイトについて説明し、ウイルス拡散に与える影響を分析する. また新たに登場が予想されるターゲット選択方法として、感染国数を急速に増加させるハブリストスキャンを説明し、その拡散速度についても調査する.

以下の構成は、2章で関連研究、3章でIP空間と国の関係、4章でシミュレーション、5章で考察である.

2. 関連研究

ウイルスはメール型とセキュリティホール型に大別される. セキュリティホール型ウイルスはネットワークサービスのセキュリティホールを介して自己増殖するタイプであり、ワームとも呼ばれる. この場合、IP アドレスを基にターゲットが選択される. メール型ウイルスはメールの添付ファイルを実行することで感染するタイプである. この場合メールアドレスを基にターゲットが選ばれる. 本研究では特に、セキュリティホール型に焦点を当てて議論する.

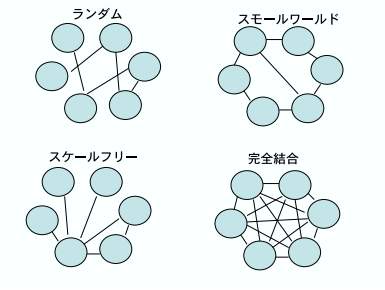

図 1　各ネットワークトポロジー

次にウイルスの増殖過程についての先行研究を説明する. 主に数理モデルやシミュレーションにより増殖または拡散過程の解析を行っているものを取り上げる. 拡散過程の解析の視点として、ネットワークトポロジーとウイルス感染台数の増加速度がある. 以下それぞれ説明する.

2.1　ネットワークトポロジー

始めに想定するネットワークトポロジーごとの研究に分けて説明する. ネットワークトポロジーとはホストをノードとして、ノード間の接続関係を線分で表した場合の特徴を示す. 代表的なネットワークトポロジーであるランダム、スモールワールド、スケールフリー、完全結合ネットワークを図 1 に示す.

ラ ン ダ ム

ランダムネットワークを用いた例として White[1] らの古典的研究がある. これはホストとホストがランダムに接続してるランダムネットワークをウイルス増殖の基盤としている. しかし、ランダムネットワークはウイルス感染の観測データを適切に説明できないとされている. この研究では、電子媒体（フロッピーなど）を介して広まる古典的ウイルスが主な対象であった.

スケールフリー

Barabasi は度数分布がべき乗則になるさまざまな現象を文献[12] にまとめている. べき乗則は典型的なノードが存在しないことからスケールフリーと呼ばれる. 保有メール数、友人関係、論文の引用回数、ルーターやウェブのリンクなどがスケールフリーネットワークに従うと報告されている. べき乗則とは度数を k、度数 k のノードの出現確率を P(k) とすると $P(k) \sim k^{-\gamma}$ の関係が成り立つ場合である.

スケールフリーネットワークを用いたウイルス解析の結果、ハブと呼ばれるリンク数の多いノードに感染すると増加速度が速まるとされている. 極端にリンクの多いハブの存在はスケールフリーネットワークの特徴である. またハブを優先的に免疫するなどの対策を施すことにより効果的にウイルスの封じこめができると報告されている[7]. スケールフリーネットワークはメールを介したウイルスを対象としてる.

スモールワールド

メール型ではその他に、スモールワールドネットワークが想定されている. これはノードの大多数は近傍と接続しているが、ショートカットと呼ばれる遠くのノードと接続してるノードのおかげで、ランダムネットワークよりも短い距離で他のノードにリンクできるという特徴を持つ[11]. ウイルスの感染シミュレーションにおいては、ショートカットをもつノードに感染すると拡散速度が速まるとされている[8)9)].

完 全 結 合

2001 年に登場した Nimda を初め SQL Slammer などさまざまなセキュリティホール型ウイルスが登場している. これらのウイルスのターゲット選択方法は IP アドレスをランダムに選ぶため全てのノードが結合していると仮定する. このようなウイルスのモデル化として Staniford[2], Zou[4] らが感染の数理モデルやシミュレーション結果について報告している.

ランダムスキャンのウイルスモデルでは完全結合のネットワークが想定されている.

また、CodeRed, Blaster などのウイルスはランダムスキャンの他に、ローカルスキャンも含んでいる. この場合サイト（上位 16 ビットごとの集合）ごとに最初に登場したウイルスがサイトの脆弱なホストに感染していくというシミュレーション結果も報告されている[3)4)].

2.2　ウイルス増殖速度

もう１つの解析の視点は、ウイルスの増殖速度に注目する場合である. これは特に完全結合ネットワークにおいて、どのようなターゲット選択方法が最も速くウイルスの感染台数を増加させるかを解明することを目的としている. 先行研究としては、Staniford[2)13)], Zou[4] らによりウイルスのスキャン方法と感染台数の関係が検討されている. これらの研究よると最も急速に感染台数を増加させるには、初期状態でのウイルス感染数が多いこと、脆弱ホストのターゲットリストを持つことなどが主な方法とされる.

2.3　問　題　点

従来研究では IP アドレスを基にした感染台（ホスト）数でウイルスの流行状態の分析がなされているが、地域的な拡散との関係については十分な検討がなされていない. 本研究では、地域的な拡散を解析するために感染国数に注目した分析を行う. 特に IP アドレスは国と対応しているため、スキャン方法ごとに地理的な拡散に特徴があると予想される. IP 情報と地理情報との関連性について次章にて詳しく述べる.

3.　IP 情報と地理情報の関係

本章では IP アドレスと IP 保有国の統計的特徴についての調査結果を説明する. 調査項目は、IP アドレスの国ごと使用率, IP アドレス上位 16 ビットごとの

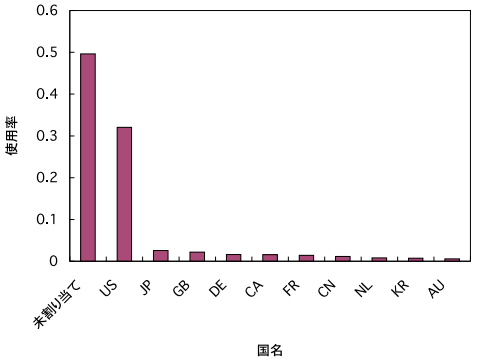

図 2　IP アドレスと国の利用率（上位１０か国）

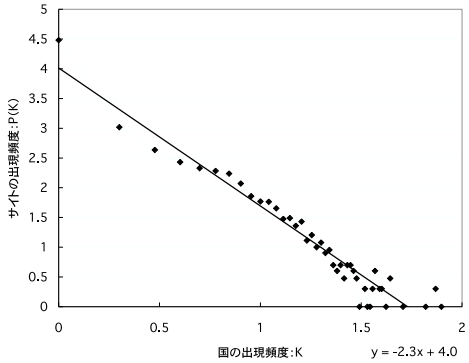

図 3　サイトと国数の度数分布 (両対数表示)

集合（以下、サイトと呼ぶ）に含まれる国数である. 前者はランダムスキャンにおける国ごとのスキャンを受ける確率であり、後者はローカルスキャンの単位であるサイトごとに含まれる国数である.

3.1　IP アドレスの国ごと使用率

IP アドレスと国の割り当て率について、図 2 に示す. これは IP アドレスと国の対応関係のデータベースである GeoIP[10] を用いて 160 万 IP を標本抽出した結果である. アメリカが大半を占め、次いで日本、イギリスなどが続いている. また多くの IP は未使用のため、ランダムに IP アドレスを選んだ場合は未使用領域であるかアメリカ所有である確率が非常に高いといえる. また効率的にスキャンをするに、未割り当て領域はターゲットに選ばないウイルスもある [5].

3.2　サイトごとの国数

次にサイトごとに含まれる国の数の度数分布を調査した. つまりローカルスキャンの単位である上位 16 ビットごとの集合に含まれている国の度数を計測した. 計測の結果、サイトごとの国数 k と国数が k であるサイト数 P(k) の関係は、べき乗則:$P(k)$〜$k^{-2.3}$ になっていることが解った (図 3). これは合計 33099 サイト、サイトごとの平均国数は 1.4 である. これからサイトの国数 k と国数が k であるサイト数 P(k) はスケールフリーの特性を持つことが解った. これは大多数のサイトでは少数の国しか含まれていないが、少数のサイトに非常に多くの国が含まれているという特徴を持つ. このような非常に多くの度数を持つ少数のノードであるハブの存在がスケールフリーの特徴である. このようなハブはランダム、スモールワールドなどの他のネットワークトポロジーでは出現しない.　本研究ではこのような、含まれる国数が非常に多い（25 か国以上の）サイトを、ハブサイトと呼ぶことにする. 最も国数が多いのは、194.117 であり、79 か国がこのサイトを分割して利用している. 194.117 に含まれる国を表 1 に示す.　データベースの解析の結果ハブサイトは 44 サイト存在することが解った.

表 1　ハブサイト:194.117 に含まれる国

国名（ISO 表示）
PT,DE,FR,GB,SE,RU,US,JP,NL,CH,ES,IT,AU
AT,SG,BE,CA,DK,HU,TR,MY,BR,GR,PL,AR,LU
NO,IN,SK,NZ,AE,TW,CN,MX,RO,SA,IL,ZA,IE
CY,VE,BG,PE,BM,CL,CZ,IS,NA,KE,KW,LB,DO
FI,KR,HK,MU,LI,PH,MZ,ZW,ID,SI,KZ,TH,AO,CU
MP,BH,AN,UA,VI,PR,HR,QA,AD,CO,BN,RE,MA

3.3　特徴のまとめ

この章では、IP アドレスと国数の関係について、統計的特徴（IP 使用率やハブサイトの存在）を説明した. 次章において、この統計的特徴が国数を感染の単位とした場合のウイルスの拡散過程にどのような影響があるのかについて分析を行う.

4.　シミュレーション

ここでは、ウイルスの拡散過程において、ウイルス拡散シミュレーションを通じて感染国数の分析をする. 分析の焦点はウイルスのスキャン方法や前章で調査した統計的特徴の影響を明らかにすることである.

4.1　実 験 準 備

4.1.1　シミュレーションの想定

シミュレーションでは、32 ビットの IP 空間でノードが全て接続しているランダムネットワークを想定する. 各ノードは IP アドレスと国情報を持ち、感染ホストの国情報を調査することで感染国数を計測する. IP アドレスと国の対応は 3 章と同じく GeoIP を利用する. また IP アドレスの未割り当て領域にはウイルスは発生しないこととする. 初期状態ではウイルス感染ホストは一台存在する. 単位時間（ステップ）ごとに感染ウイルスは一回スキャンを行うものとする. ターゲットホストが脆弱である場合はそのホストはウイルスに感染し、次のステップから感染行為を行う. 脆弱ホストの分布は各国一定と仮定する.

4.1.2　分 析 項 目

分析項目を以下に述べる.

- 単一ウイルス感染ホストのスキャン回数とスキャン受信国数の分析
- スキャン方法ごとのウイルスの感染国数の分析
- スキャン方法ごとの感染台数と感染国数の分析
- ローカルスキャンを含む場合の感染が起きたハブサイトと感染国数の分析

4.1.3 スキャン方法

シミュレーションではランダムスキャン、ローカルスキャン、BGP スキャン、ハブリストスキャンを取り上げる. 以下に詳細を説明する.

ランダムスキャン

32 ビットの IP アドレスからランダムにターゲットを選ぶ方法である.

BGP スキャン

ランダムスキャンの拡張として BGP テーブルを利用して、実利用されている IP アドレスの範囲からランダムに選ぶ [5]. ウイルスは BGP テーブルをプログラム内に保有している.

ハブリストスキャン

ハブリストスキャンは実際のウイルスには含まれていない想定上のスキャン方法である. ウイルスがハブサイトのリストを参照して上位 16 ビットを選び、下位ビットを切り替えてスキャンを行う方法とする. これはウイルスが意図的にハブサイトを狙うという方法である. ハブリストスキャンはウイルスの個体数の増加ではなく、地域的な拡散を狙ったスキャン方法である. ウイルスは感染時にはハブサイトリスト（プログラムなどで参照可能な表）を持っていると仮定する.

ローカルスキャン

IP アドレスのうち、上位 16 ビットを固定し、下位ビットを変更させる. 通常ランダムスキャンと組み合わせて利用される.

また今回取り上げないがウイルスがターゲットリストを持つヒットリストスキャンは研究者らが先行的にスキャン方法を想定したものであり感染台数の増加速度を早めるとされている [2]. ヒットリストスキャンには事前に脆弱なホストのリストアップが必要であるため対象外とする. これらのスキャン方法により感染台数の増殖過程（増加速度）に差があることが従来研究で報告されている [2][4]. 本研究はこのようなスキャン方法の違いが感染国数にどのような影響を与えるか, またどのスキャン方法が最も早く拡散速度が速いかを分析する.

4.2 シミュレーション解析結果

ここではシミュレーション結果について述べる.

4.2.1 スキャン数と受信国数

ここでは単一ウイルスの場合の解析結果として、スキャン方法ごとのスキャンの受信国数について述べる. 用いるスキャン方法はランダム, BGP, ハブリストスキャンである. これらのスキャン方法は、どの IP アドレスで感染が起きてもターゲット選択結果に影響はない. 図 4 に 5 0回試行した平均を示す. ここでは 10000 回ごとにまとめて表示している. 横軸はスキャン回数を表しており, 縦軸はスキャンの受信国数を表している.

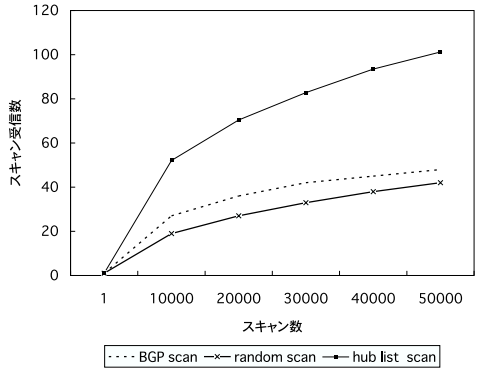

図 4 単一ウイルスにおける各スキャン方法とスキャン受信国数

ウイルスの単位時間のスキャン回数は TCP 型であれば１分間に 358 回程度といわれている. 従ってこの結果は TCP を用いる一台のウイルスが約 2 時間程度スキャンを行った場合に相当する. これよりスキャンの受信国数はハブリストスキャンが最も多く、BGP スキャンが次いで多いという結果である. この方法はどちらもウイルスに内臓される表を参照するため効率のよい結果となっている. またランダムスキャンやや少ない受信国数であった. これは前章で述べた通り、ランダムスキャンでは大部分は未使用領域か US にスキャンが行われるため受信国数は増加しないと考えられる. また BGP スキャンは未使用領域はスキャンをしないためやや効率の良い結果となっている. さらにハブリストスキャンは国が密集した領域を集中的にスキャンするため最も効率の良い結果となったと考えられる.

4.2.2 スキャン方法と感染国数

次にウイルスの増殖過程が行われる場合のシミュレーション結果について述べる. シミュレーションでは国ごとに平均して 10 個ホスト中に 1 台脆弱ホストが存在し、全体で脆弱ホストは約 1000 万台存在すると仮定する. また計算量の制約があるため、完全に蔓延するまえにシミュレーションを中断し、約 150 ステップまでの結果で解析する. ウイルスのスキャンのターゲットが脆弱なホストであればウイルスに感染し、ウイルスに感染したホストの IP アドレス情報から感染国数を計測する. 用いるスキャン方法は BGP, ランダム、ハブリスト、BGP とハブリストを組み合わせたものである.

図 5 に各スキャン方法ごとの感染国数のシミュレーション結果を示す. これは５０試行の平均を示している. 横軸はシミュレーションでのステップ数であり、縦軸は感染国数を表す. これより従来ウイルスでよく用いられるランダムスキャンでは感染国数は低い結果

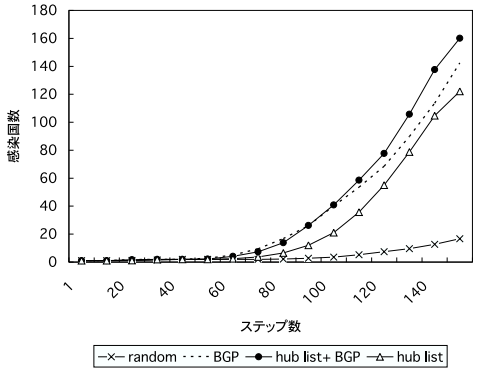

図 5　ウイルスの拡散過程における
スキャン方法ごとの感染国数

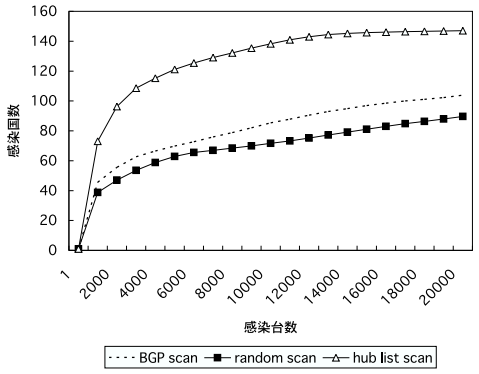

図 6　ウイルスの拡散過程における
感染国数と感染台数の関係

になっている. また BGP スキャンやハブリストスキャンは急激に感染国数を増加させていることが解る. 更に BGP スキャンとハブリストスキャンを組み合わせた方式は最も感染国数が多いという結果であった. なおハブリストと BGP スキャンの比率は 2 対 8 である.

4.2.3　感染台数と感染国数

次に感染国数と感染台数の関係の分析を行った. 感染台数ごとの感染国数を図 6 に示す. 横軸は感染台数であり、縦軸は感染国数である. これよりハブリストスキャンは感染台数あたりの感染国数が他のスキャン方法よりも多いことが解る. つまりハブリストスキャンは感染台数は少なくても感染国数が他のスキャン方法よりも多いといえる. BGP スキャンが感染国数を増加させているのは、国ごとの感染効率が高いわけではなく、ホストごとの感染効率が良いため福次的な効果として感染国数も増加していると考えられる.

4.2.4　ハブサイトとローカルスキャン

最後にハブサイトの影響を分析するために、スキャン方法がローカルスキャンを含む場合の分析を行う. ローカルスキャンは上位 16 ビットを固定して下位ビットを変化させるため、ハブサイトで感染した場合には感染国数に大きな影響を与えると予想される. ローカルスキャンは感染ホストの IP アドレスに依存（上位 16 ビットを利用）してターゲット選択するため感染場所に影響される. シミュレーションでは初期感染ホストはすべてハブサイト（194.117）で固定的に出現させた.

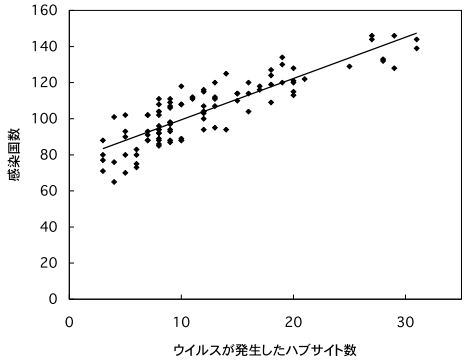

図 7　ローカルスキャンを含む場合の
ハブサイトの感染と感染国数の関係

ローカルスキャンを含むウイルスのシミュレーション結果を図 7 に示す. 横軸はウイルスが発生したハブサイトの数であり、縦軸は感染国数である. 完全に蔓延する前の途中結果である. これよりローカルスキャンを含む場合、ハブサイトの感染数と感染国数は比例関係にあるといえる. ハブサイトの感染が多く生じた場合は感染国数が増加している. またハブサイトへの感染が少ない場合感染国数は少ない結果となっている. これより、ハブサイトへの感染により感染のダイナミズムが変わり、地理的な拡散が速まる（感染国数が増加する）と考えられる. またハブサイトの感染数が少ない (0 から 10 の) 場合は感染国数も少ない（地域的な拡散が遅い）といえる. しかし、20 サイト以上ハブサイトにウイルスが発生した場合はいずれも多く（120 か国以上）の感染国数になっている. この極端な例がハブリストスキャンでありさらに感染国数を増加させることが前々節にて明らかになっている. 地域的な拡散速度はハブサイトが初期に集中的に狙われた場合（ハブリストスキャンなど）は、事後の封じこめなどの対策をより困難にすることが予想される.

5. 考　　察

ここでは、ハブサイトスキャンの効果とハブサイトの影響について考察する.

5.1　ハブリストスキャンの効果

ここではスキャン方法、特にハブリストスキャンが感染国数に与える影響についてまとめる. 本研究で想定したハブリストスキャンはウイルスの地域的な拡散

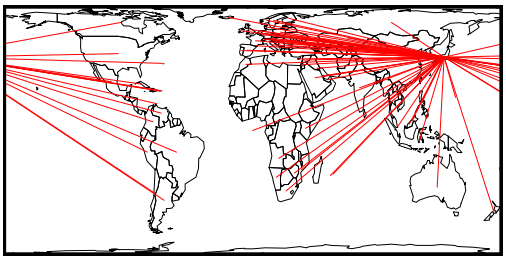

図 8　ハブサイト：194.117 へのハブリストスキャン
日本で感染した場合のイメージ図

を急増させるターゲット選択方式であることがシミュレーションによる分析で明らかになった. 特に BGP スキャンと組み合わせることで最も急速に地域的な拡散が起きることが解った. このようなハブサイトへのスキャンは世界中にウイルスを拡散させることから地域的な封じこめ対策などが難しくなると考えられる. ハブサイトに対するハブリストスキャンのイメージを図 8 を示す. これはハブサイト 194.117 に対して日本でウイルスが発生しスキャンを行っている視覚化図である. これは IP 空間的には同じ上位 16 ビットへのスキャンであったも、地理的にはグローバルにスキャンを行っている. ハブリストスキャンを行うウイルスはまだ登場していないが、ハブサイトである組織は何らかの対策が必要である.

5.2　ハブサイトの影響

メール型ではハブというリンク数の多いノードに感染すると拡散速度が速まるとされていた. ローカルスキャンを含むウイルスの解析の結果、ウイルスが発生したハブサイト数が多いほど感染国数が比例して上昇していることが解った. つまりハブサイトでウイルス感染が起きると拡散速度を速めるという結果であった. この結果はメール型ウイルスにおいて、ハブ（保有アドレス数が多いユーザ）が感染すると拡散が速まるのと同様の効果が、ローカルスキャンを含む場合のハブサイトの感染にもあてはまるといえる. つまり、感染のダイナミズムは少数しか存在しないハブの存在に影響されるといえる.

ウイルス対策への応用として、ハブサイトの集中免疫をした場合の効果などを今後検討する予定である.

6. おわりに

本研究では、感染国数に注目してセキュリティホール型ウイルスの拡散速度の特性について解析した. 当初予備調査では IP アドレスの上位 16 ビットごとに含まれる国の度数分布がべき乗則に従うことを示した. またスキャン方法を変化させた場合の拡散速度の変化をシミュレーション実験により分析した. 解析の結果、地域的な拡散を目的としてウイルスにおいて、ハブサイトを集中的に狙ったハブリストスキャンを用いると拡散速度 (拡散速度) は急増することが解った. またローカルスキャンを含むウイルスの場合は、多くのハブサイトでウイルス感染が起きると感染国数が上昇することが解った. 以上よりウイルスの地理的な拡散は多くの国を含むハブサイトに大きく影響されることが明らかになった. このような地理的な拡散はウイルスの封じこめを困難にすると考えられるため、ウイルス対策方法を今後検討する予定ある.

参　考　文　献

1) J.O.Kophart, S.R.White, Directed-Graph Epidemiological Models of Computer Viruses, http://www.research.ibm.com/antivirus/SciPapers/Kephart/VIRIEEE/virieee.gopher.html
2) S.Staniford, V.Paxson, N.Waver,How to 0wn the Internet in Your Spare Time, Usenix 11th 2002.
3) T.Vogt, Simulating and optimising worm propagation algorithms, http://web.lemuria.org/security/WormPropagation.pdf.
4) C.C.Zou, D.Towsley, W.Gong, On the Performance of Internet Worm Scanning Strategeis, Technical Report TR-03-CSE-07, Nov, 2003.
5) C.C.Zou, D.Towsley, W.Gong, S.Cai, Routing Worm: A Fast, Selective Attack Worm based on IP Address information, Umass ECE Technical Report TR-03-CSE-06, November, 2003.
6) R.Albert, A.L.Barabasi, Statistical mechanics of complex networks, Reviews of Modern Phisics, Vo 74, Jan 2002.
7) Z.Dezso, A.L.Barabasi, Halting Viruses in scale-free networks, Phys. Rev. E 65, 055103 (2002).
8) C.C.Zou, D.Towsley, W.Gong, Email Virus Propagation Modeling and Defense, 13th International Conference on Computer Communications and Networks (ICCCN'04), October 11-13, Chicago, 2004.
9) M.Garetto, W.Gong, D.Towsley, Modeling Malware Spreading Dynamics, The 22nd Annual Joint Conference of the IEEE Computer and Communications Societies.
10) MaxMind, Geolocation IP Address to Country, http://www.maxmind.com.
11) Watts.D.J, S.H.Strogatz, Collective dynamics of 'small-world' networks, Nature 393:440-42.
12) Barabasi、新ネットワーク思考, NHK 出版.
13) S.Staniford, D.Moore, V.Paxson, N.Weaver, The Top Speed of Flash Worms, Workshop of Rapid Malcode 2004.