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

2006－CSEC－34 (41)
2006／7／21

インターネットの分散観測による不正侵入者の探索活動のマクロ・ミクロ解析

福野 直弥† 小堀 智弘† 菊池 浩明† 寺田 真敏†† 土居 範久†††

† 東海大学電子情報学部情報メディア学科
259-1292 平塚市北金目 1117
†† 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
††† 中央大学理工学部情報工学科
112-8551 東京都文京区春日 1-13-27
E-mail: †{fukuno,nopay,kikn}@cs.dm.u-tokai.ac.jp

あらまし ワームやウィルス，スパイウェアなどに感染したホストはトラップドアを仕掛けたり，新たな侵入先を求めて定期的なポートスキャンを実行したりしている．これらの振舞いは一様ではなく，特定のアドレスブロックを集中的に探索する不正者やアドレス空間全域をランダムに探索する不正者などが混在している．そこで，本研究では，JPCERT/CC により運営されている定点観測システム ISDAS のデータを解析し，インターネット上での探索活動を明らかにすることを試みる．まず，探索者全体の集合をマクロに見て統計的な情報を解析し，次に個々の探索者に着目してミクロに解析を行う．これらの解析結果を元にして，不正者の振舞いを近似する数学モデルを提案する．

Macro and Micro Analysis on Vulnerability Scanning Activities via Distributed Observation over the Internet

Naoya FUKUNO†, Tomohiro KOBORI†, Hiroaki KIKUCHI†, Masato TERADA††, and Norihisa DOI†††

† Couse of Information Engineering, Graduate School of Engineering Tokai University
1117 Kitakaname, Hiratsuka, Kanagawa 259-1292
†† Hitachi, Ltd. Hitachi Incident Response Team (HIRT)
890 Kashimada, Kawasaki, Kanagawa 212-8567
††† Dept. of Info. and System Engineering, Faculity of Scienece and Engineering, Chuo University 1-13-27
Kasuga, Bunkyo, Tokyo 112-8551
E-mail: †{fukuno,nopay,kikn}@cs.dm.u-tokai.ac.jp

Abstract Computer virus and worms perform randomly spyware and port-scanning to find a vulnerability in the Internet. The fraction of malicious behaviors varies, e.g, some host performs scan contentionally and some host scans uniformly over the IP address blocks. In this paper, First, we analysis a set of source addresses observed by distributed sensors in ISDAS from a "macro" view point. Second, we examine behaviors of from "micro" perspective. Finally, we study a new mathematical model for malicious hosts based on these analysis.

1. はじめに

現在，インターネット上には多種にして大量のワームやウィルスが存在している．それらの不正ホストは新たな侵入先を求めて定期的にポートスキャンを実行したりしている．しかし，これらの振る舞いは一様ではなく，特定のアドレスブロックを集中的に探索したり全 IP アドレス空間を探索するものなど，多種多様である．

この感染活動は不正ホストの種類によって異なるため一つ一つ特定がされていない．例えばワームの代表的なものとしてSasserがある．[2]によれば，32%の割合でスキャン対象のアドレスを完全にランダムに，23%でIオクテット以外をランダムに生成しては，TCPコネクションの確立を試みる．また，別のワームにWittyがある．[3]によるとそのワームは送信先IPアドレスをランダムに生成するが，その生成法は不完全で全てのIPアドレスブロックを探索することはない．また，観測した不正ホストの内，47%が感染してから5日以内であることが明らかになっている．このように，不正ホストの振る舞いを解析するには次の二種類が考えられる

マクロ解析　不正ホストを集合としてとらえ，そのスキャン等の行動パターンについて，全体の何割が実行しているかを明らかにする．不正ホスト集合を大極的にとらえて，モデルを作り，統計的な行動パターン量を予測できる．[1]や，[4]の閾値ランダムウォーク等がある．

ミクロ解析　特定の不正ホストに注目し，その時間軸上での各センサの観測結果を合わせて，特異な行動パターンそのものを分析する．[3]のように，ウィルスの献体を逆エンジニアリングしたりして解析したりする手法が該当する．

我々はJPCERT/CC [5]が運営する定点観測システムISDASのデータを解析し，インターネット上での探索活動を明らかにすることを試みる．まず，探索者全体の集合をマクロに見て統計的な情報を解析し，次に個々の探索者に着目してミクロに解析を行う．これらの解析結果を元にして，不正者の振舞いを近似する数学モデルを提案する．

2. 基本定義

不正ホストとは，ウィルスやネットワークワームなどにより他のホストへのスキャン（ポートスキャン等）を仕掛けるホストである．センサとは，不正ホストからのスキャンを観測する正規ホストであり，決して感染しない．このときセンサ台数をmとした時センサの集合を$S=\{s_1,s_2....s_m\}$とする．有効な全グローバルアドレスの数をn_0，不正ホストの数をnとする．x台の独立したセンサで期間$[0,t]$で観測できる異なるセンサの累積数をユニークホスト数と呼び，$h(x,t)$で表し，その一日毎の平均を$\Delta h(x)$[回/日]とする．

また，一台の不正ホストが期間[0,t]にスキャンするセンサの種類数をビジット数kとし，ある期間[0,t]にセンサs_iをスキャンする総回数をC_iで表す．また，センサs_iについて不正ホストjによってスキャンされる間隔を$\omega_{i,j}$と書き，不正ホストの区別をしない時のスキャン間隔をω_iで示す．また全センサSについてのにスキャンする間隔をω_Sとする．

2.1 定点観測データ　ISDAS

ISDAS(Internet Scan Data Acquisition System)は，JPCERT/CCが運用しているセキュリティに関するトラフィックの分散観測システムである[5]．単位時間当たりの主要なポート(135,445,139,1026,80)のトラフィックなどを毎週更新している．我々の解析十分な観測期間における独立した複数のセンサの観測データを必要とする．そこで，JPCERT/CCの協力の下で，2004年9月1日～2005年9月30日の間の独立した12台($m=12$)のセンサのログデータを解析する．それぞれのセンサの情報を表1に示す．

表1　ISDASによる各センサの観測データ

センサ	総スキャン数	$h(x)$	$\Delta h(x)$[/日]
s_1	268024	97102	245.8
s_2	153310	63198	160.0
s_3	154126	60755	153.8
s_4	137848	40315	102.1
s_5	168191	62881	159.2
s_6	173566	47809	121.0
s_7	17167	10066	25.5
s_8	164078	54865	138.9
s_9	10667	9046	22.9
s_{10}	170417	24394	61.8
s_{11}	30898	13200	33.4
s_{12}	143725	53716	136.0

今回の使用したデータでビジット数kの取りうる範囲は$1\leq k\leq 12$となる

3. マクロ解析

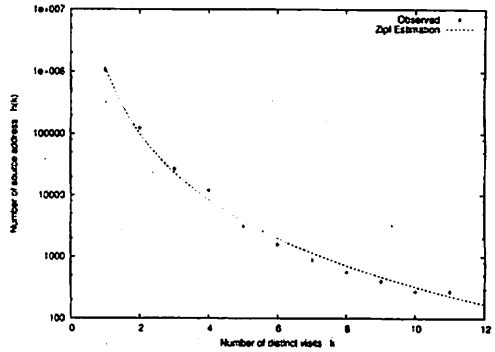

図1　観測されたビジット数kについてのユニーク発信アドレス数とその近似

不正ホストには積極的に多くのアドレス空間を渡り歩くものがあるが，ほとんどはごくまれにしか観測されない．この「渡り」の度合いを定量化するために，分散された（12台の）センサのうちそのアドレスを観測したセンサの台数をビジット数kと呼び，ビジット数kについての異なる発信アドレス数$h(k)$の実測値は図1のような分布をしている（Y軸が対数軸になっていることに注意）．ここで，全体の86%にあたる105万台の不正ホストはビジット数$k=1$であり，第2位の$h(2)$は12万台と続き，$k=12$となるホストはわずか293台しかいない．すなわち，不正ホストのスキャンの振舞いは一様ではないことがわかる．

この不正ホストの集合をマクロに考えて，その不正スキャン行為の振舞いの定式化を試みよう．ここでは，様々な自然現象や社会行動のモデルとして知られているジップの法則(Zipf's

law) [6] を適用する．ジップの法則はウェブページへのアクセス頻度や単語の使用頻度と順位の関係などに広く適用できることが知られており，k 番目の部分集合の割合は $1/k$ に比例するという経験則である．すなわち，ビジット数 k の不正ホスト数 $h(k)$ は，

$$h(k) = \frac{h_0}{k^s}$$

で表せる．ここで，s は定数であり，オリジナルのジップの法則では $s = 1$ である．k は順位ではないが，この例のように十分大きな集合を $k \leq 12$ 程度の小さな集合へ分割するとき，順位に一致する．そこで，s を同定するために，$h(k)$ の対数を取り，実測値について最小二乗法を適用すると $h_0 = 1174433, s = 3.5616$ が得られた．こうして得られた近似式を図 1 の上に重ねて示す．実測値との強い一致が観測できる．

4. ミクロ解析

4.1 不正ホストのタイプ

不正ホストを集合ではなく，個々のホストに注目してミクロに見ると，定期的にスキャンを繰り返す等の特異な行動を観察できる．不正ホストはその振る舞いから，周期型，集中型，ランダム型の 3 種類に分類出来る．周期型とは，不正ホストが全センサを規則性を持ってスキャンするタイプである．集中型は一部のセンサを集中的にスキャンをする不正ホストである．そして，周期型，集中型でないものをランダム型とする．

この 3 タイプの分類は，経験を積んだ技術者ならば曖昧なく行うことが出来る．ここでは，各タイプの特徴に注目し，この分類を自動化できないか考えよう．

まず，周期型は全センサで同じ間隔でスキャンをすることが多い．すなわち，各センサの観測数は一定になりその到着間隔と観測回数の分散は少ないはずである．一方，集中型の場合，特定のセンサのみが頻繁にスキャンされるため，各センサ間の観測回数の分散は逆に大きくなることが期待できる．全てのセンサにスキャンしているためその他のセンサにスキャンする際の観測間隔が大きくなる．

4.2 解析方法

対象とする不正ホストのアドレスは 105 万台あり，ミクロにすべて解析するには多すぎる．そこで，ここでは 3. 章で示した観測された不正ホストの内，ビジット数 $k = 12$ となる部分集合 A に着目する．この部分集合の大きさは，$h(12, 2004/9/1 - 2005/9/30) = 293$ である．まず，この 293 個のアドレスを人手で 3 種類のタイプへ分類する．次にこの作業を自動化するため以下のパラメータを定義する．不正ホストを 3 種類のパターンに分けることを目的とする．

まず，A_12 の全不正ホストによるスキャン数 $C_1 ... C_m$ の m 台のセンサ平均 $\mu(C_\bullet)$ と分散 $\sigma(C_\bullet)$ は，

$$\mu(C_\bullet) = \frac{1}{m}\sum_{i=1}^{m} C_i, \sigma(C_\bullet) = \frac{1}{m-1}\sum_{i=1}^{m}(C_i - \mu(C_\bullet))^2 \quad (1)$$

で表される．不正ホストがセンサ s_i をスキャンする間隔 ω_i の全スキャン平均を

$$\mu(\omega_i) = \frac{1}{C_i - 1}\sum_{j=1}^{C_i - 1} \omega_{i,j} \quad (2)$$

と定める．ただし，ここで，$\omega_{i,j}$ はセンサ s_i における j 番目と $j+1$ 番目の巣間の時間間隔を表している．こうして定められた各センサの到着間隔を用いて m 台についてのスキャン到着間隔 $\omega_\bullet$ のセンサ平均，分散を次式で定める．

$$\mu(\omega_\bullet) = \frac{1}{m}\sum_{i=1}^{m}\mu(\omega_i), \sigma(\omega_\bullet) = \frac{1}{m}\sum_{i=1}^{m}(\mu(\omega_i) - \mu(\omega_\bullet))^2 \quad (3)$$

この値から，各センサ毎の到着間隔のばらつきをみることができる．つまり，この値が小さいということは各々のセンサが一定間隔でスキャンされていると考えられる．

最後に，センサ集合 S を統合して単一のセンサと見なした時のスキャン到着間隔を ω_S で表し，その平均と分散を

$$\mu(\omega_{S,j}) = \frac{1}{mC_\bullet}\sum_{j}^{mC_\bullet}\omega_{S_j} \quad (4)$$

$$\sigma(\omega_{S,j}) = \frac{1}{mC_\bullet - 1}\sum_{j=1}^{mC_\bullet}(\omega_{S,j} - \mu(\omega_S))^2 \quad (5)$$

と表す．ただし，$\omega_{S,j}$ は全センサにおけるスキャンを時系列にした時の j 番目と $j+1$ 番目のスキャンの間隔を表しており，その総数は $mC_\bullet$ で表せられる．

4.3 解析結果

人手で主観的に分類した，A_12 における不正ホストのタイプ別の個数を 2 に示す．A_12 においては全体の半分がランダムであり，1/3 が周期型であった．この周期型タイプの割合は k が小さくなるに従って小さくなることが予想される．実際に，分類された各タイプについて代表的なホストを選び出し，それらの統計値を求めると 3 のようになった．この考察の通り，V_1 は $\sigma(\omega_\bullet), \sigma(C_\bullet)$ の値が小さく周期型で V_2 は $\sigma(\omega_\bullet)$ が小さいが $\sigma(C_\bullet)$ の値が大きい集中型であった．しかしながら，これらの 6 種類の統計値の閾値を設定するのは容易ではない．そこで，このデータにクラスタリングアルゴリズム C4.5 を適用する．C4.5 は，情報量利得に基づいて，識別を行い決定木を生成するアルゴリズムであり，連続値の閾値も対応している．4 に，A_12 を C4.5 にかけた結果を示す．ここで，木の節定が識別の値であり，枝が閾値，葉がタイプを表す，葉に示されている値は，(正しく分類された個数)/(誤分類数) である．木を簡単かするために (m=4,c=90) で枝狩りをしている．また，この時のクラスタリングの精度と結果を 4 に示す．

図 2 に振る舞いに周期性を持つ不正ホスト V_1 のスキャン先の推移を示す．2005 年 4 月 1 日前後を境に規則的にセンサをスキャンしているのがわかる．図 3 に 12 台についてのセンサが V_1 からのスキャンの到着間隔の分布を示す．

次に，スキャンを集中して行う不正ホスト V_2 のスキャン先の推移を図 5 に示す．V_2 が s_3 と s_{10} を集中的にスキャンをしている．また，各センサの間隔の分布を図 6 に示す．その時の s_3 と s_{10} のスキャン到着間隔が他と比べて狭いことがわかる．

表 2　不正ホストタイプの頻度と割合

	周期型	集中型	ランダム型	合計
頻度	90	23	180	293
割合	0.3	0.1	0.6	1

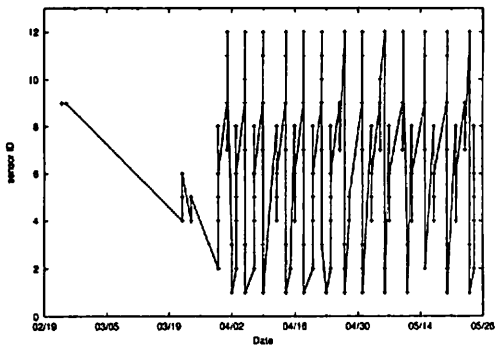

図 2　スキャン先の時間推移 (発信元 218.26.191.182 , V_1)

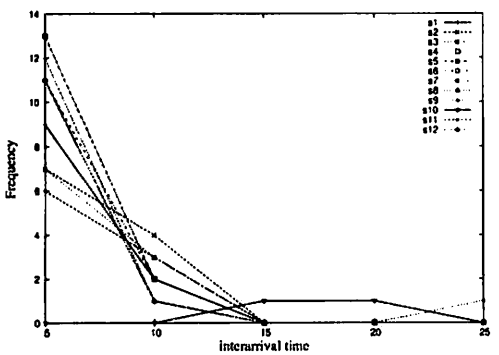

図 3　到着間隔の分布 (発信元 218.26.191.182 , V_1)

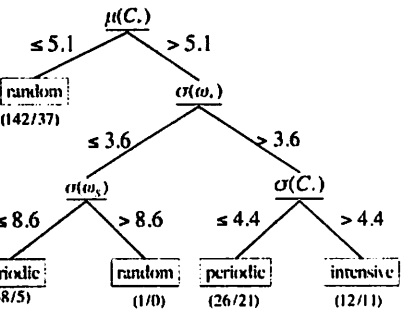

図 4　ID3tree

スキャン先をランダムに決定する不正ホスト V_3 のスキャン先の推移を図 7 に示す．スキャン回数は今まで例に挙げた V_1, V_2 よりも少ないが，特定のセンサを集中的にスキャンをしたりセンサを周期的にスキャンはしていない．V_3 のスキャン間隔の分布を図 8 に示す．これまであげた 2 例とは違い，様々なところで値を取っている．

不正ホスト V_4 は C4.5 によって誤判別の例である．クラスタリングに用いた値が V_1 に近いことがわかる．

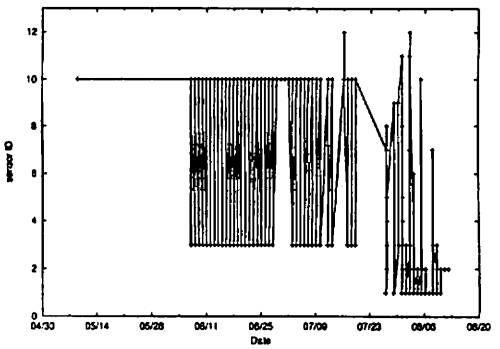

図 5　スキャン先の時間推移 (発信元 61.172.240.137 , V_2)

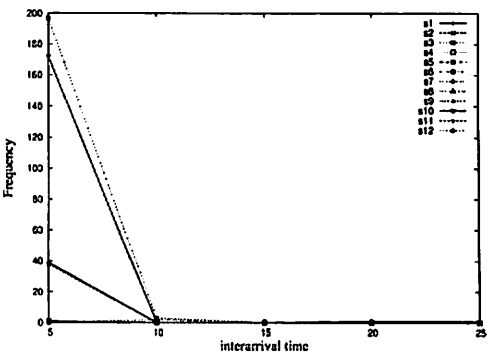

図 6　到着間隔の分布 (発信元 61.172.240.137 , V_2)

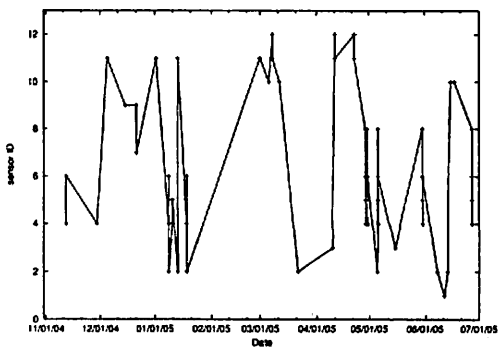

図 7　スキャン先の時間推移 (発信元 218.12.40.134 , V_3)

各タイプの代表的ホスト V_1, V_2, V_3, V_4 における各種等計量を表 3 に示す．集中型の V_2 はスキャン数の分散 $\sigma(C_*)$ が極端に大きく，特定のホストのみにスキャンが集中していることがわかる．4 のクラスタリングの結果に基づいて，分類した不正ホストの分布を図 11,12 に示す．まず，第一の識別は木の根に相当する属性である平均スキャン数 $\mu(C_*)$ であり，これ

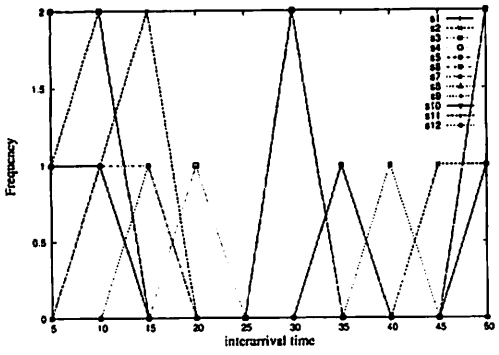

図 8　到着間隔の分布 (発信元 218.12.40.134 , V_3)

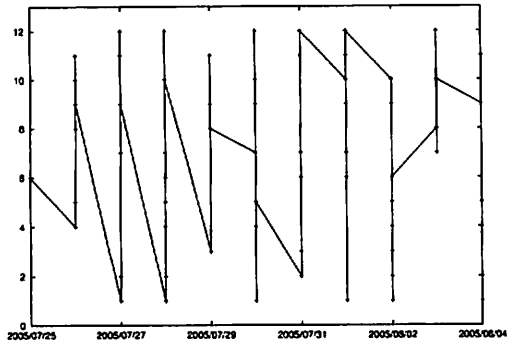

図 9　スキャン先の時間推移 (発信元 218.64.55.25 , V_4)

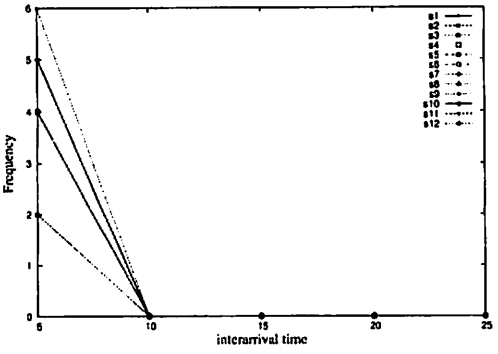

図 10　到着間隔の分布 (発信元 218.64.55.25 , V_4)

を x 軸において A_{12} をプロットしたのが図 11 である．更に次の識別属性である到着間隔の分散を y 軸にしている．閾値の $\mu(C_*) = 5.1$ を境にして左側がランダム，右がその他である．次に，図 12 はこの図 11 の右側の部分集合だけを更に到着間隔の分散 $\sigma(\omega_*)$(x 軸) とスキャンの分散 $\sigma(C_*)$(y 軸) においた時の散布図である．$\sigma(\omega_*)$3.6，$\sigma(C_*)$4.4 を超える部分 (図の右上) が集中型，それ以外が周期型である．

以上のように，C4.5 によるクラスタリングの有効性とその誤差が視覚的に示された．

最後に，これまで議論してきた k_{12} の不正ホストの使用ポートについて述べる．まず，不正ホスト一台が使用したポートの

表 3　型別の各種パラメータ

	$\mu(C_*)$	$\sigma(C_*)$	$\sigma(\omega_*)$
V_1	12.33	3.73	5.94
V_2	39.35	71.16	3.69
V_3	4.83	3.10	29.07
V_4	5.25	1.22	2.05

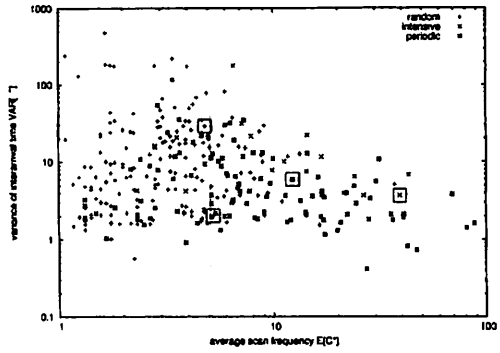

図 11　平均スキャン数 $\mu(C_*)$ と到着間隔の分散 $\sigma(\omega_*)$ についての分布

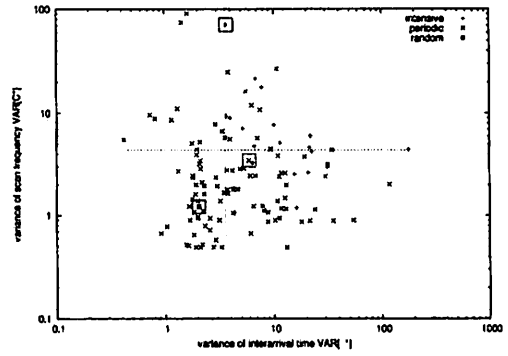

図 12　到着間隔の分散 $\sigma(\omega_*)$ とスキャン数の分散 $\sigma(C_*)$ についての分布

表 4　決定木 C4.5 による解析結果と精度

ID3 評価	周期型	集中型	ランダム型	合計	再現率
周期型	63	7	32	102	61%
集中型	2	12	6	20	60%
ランダム型	23	5	143	171	83%
合計	88	24	181		
適合率	71%	50%	79%		

内訳を数を図 13 に示す．単一のポートを用いた不正ホストが約 70% 以上あり最も多い．次に，単一のポートをスキャンの内，そのポート番号について詳細を図 14 に示す．最も多いポートは 1434 であり，約 70% を占めている．それら全ては UDP であった．これは SQLSlammer [7] であると考えられる．

5. 結　　論

ISDAS の 12 台のセンサによる 1 年間のスキャンデータについて，マクロとミクロの 2 通りの解析を行った．その結果，ビジット数と不正ホストの間には，$h(k) = h_0/k^{3.56}$ で示される

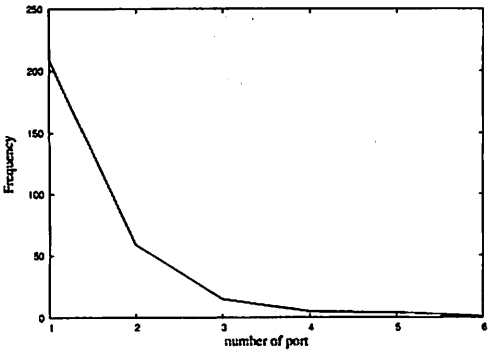

図 13　ポート種類数についてのスキャン頻度

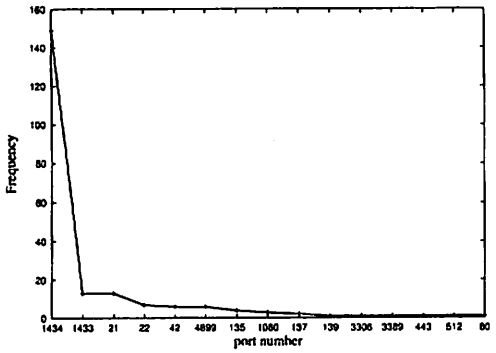

図 14　使用ポートについてのスキャン頻度

Zipf の法則が成立することがわかった．

また，ミクロに解析することにより，不正ホストのスキャンパターンが周期型，集中型，ランダム型に分類でき，その比率は 3 : 1 : 6 であることがわかった．クラスタリングの結果により，平均スキャン数が 5.1 未満のものはランダム型であり，到着間隔の分散が 3.6 以上で，かつ，スキャン数のセンサ分散が 4.4 以上のものが集中型であり，それ以外が周期型に分類できることが分かった．ただし，ISDAS の $k = 12$ のデータにおける平均再現率は 68%，適合率は 67% である．

謝　　辞

本研究を遂行するにあたり，データの解析を協力していただいた中央大学の杉山 太一氏，定点観測データ提供し，議論いただいた JPCERT/CC の竹田 春樹氏，中谷 昌幸氏，鎌田 敬介氏に感謝する．

文　　献

[1] 菊池 他，ネットには何台の不正ホストがいるのか?，情報処理学会，CSS 2005, pp.421-426, 2005.

[2] 寺田，高田，土居，ネットワークワーム動作検証システムの提案，情報処理学会論文誌，Vol. 46, No. 8, pp. 2014-2024, 2005.

[3] A.Kumar, V.Paxson, and N.Weaver,"Exploiting Underlying Structure for Detailed Reconstruction of an Internet-scale Event", Internet Measurement Conference 2005.

[4] J. Jung, V. Paxson, A. W. Berger, and H. Balakrishnan, "Fast Portscan Detection Usig Sequential Hypothesis Testing", proc. of the 2004 IEEE Symposium on Security and Privacy (S&P'04), 2004.

[5] JPCERT/CC,ISDAS,(http://www.jpcert.or.jp/isdas, 2006 年 2 月参照)

[6] George K. Zipf, Human Behavior and the Principle of Least-Effort, Addison-Wesley, Cambridge MA, 1949.

[7] IPA,(http://www.ipa.go.jp/security/ciadr/vul/ 20030126ms-sql-worm.html)