ウィルスっていったいどれ位感染しているのかな？

研究指導　菊池浩明　教授

東海大学　電子情報学部　情報メディア学科

3ADM1110 小堀　智弘

第1章　序章

1.1　はじめに

近年のインターネットの普及に伴い、インターネット上ではワームやウィルスなど猛威を振るっている。特に近年で問題となっている個人情報の漏洩などは、このワームやウィルスといったものが原因である。これらに感染すると他の PC に対して、感染活動を行うようになる。このようにして、近年では Witty ワームのように短時間で数万台という PC に感染するようになった。このような感染後の振る舞いはワームやネットワーク環境によってことなる。

1.2　定点観測システムについて

ネットワーク上のパケットを観測するシステムとして、定点観測システムがある。これは、ネットワーク上に複数の装置を設置してアクセスログを集めることを目的としている。

現在の日本では警視庁や JPCERT/CC などが運営している。

今回の実験では、JPDERT/CC より 2004 年 9 月 1 日～2005 年 9 月 30 日までの提供されたデータを使用する

第2章　研究の定義と新規性・問題点

2.1　基本定義

本研究では、ウィルスやワームなどにより他のホストへポートスキャン（以後スキャン）を仕掛けるホストを不正ホストと呼び、不正ホストの動きを観測する悪意のないホストをセンサとする。センサはアドレス空間上に独立になるように分散しているものとする。また、センサで観測された異なるアドレスをユニークホストと定義する。

j 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_j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カウント数c_iで表す。不正ホストは、ワームなどの感染によりスキャン活動を開始し、ウィルスが検知されて駆除されることで終了する。このサイクルをラウンドと呼び、その開始から終了までの長さを感染期間dで表す。いったん駆除しても再び別のワームに感染することがしばしばある。このラウンドの回数をrとする。また、ラウンドとラウンドの間の全くパケットを観測できない期間をtで表す。

2.2　新規性

本研究では、ウィルスに感染している期間はどれくらいなのかを K=6 について、ネットワーク上に設置したセンサのログデータより、感染期間を推定することを目的とする。また、K=6 以外のときの感染期間を推定し、K の値によってどれくらいの誤差が生じるのかについても検証する。

具体的な目的として、以下の 3 つについて実験した

1. 一度感染したら平均で何日間ウィルスに感染しているか
2. 年間平均で何回ウィルスに感染するか
3. K＝3～12 の時の感染期間の日数について

2.3　問題点

12 台の分散センサの観測結果に基づいて不正ホストの平均から、ラウンド数 r と，感染期間 d を求めることを目的とする。しかし、この研究を行う上で以下の２つの問題が存在する。

（1）　パケット数の多さ

JPCERT/CC より提供されたログは 100 万個以上のアドレス（表１にユニークホストとカウント数を表す）が存在し、これらすべてを手動で解析することは困難である

表１　ユニークホストとカウント数の関係

K	カウント数（回）	ユニークホスト数（個）	平均パケット数（パケット）
1	3627084	1050309	3.45
2	356412	123348	2.89
3	163297	26984	6.05
4	110590	12087	9.15
5	44031	3108	14.17
6	24249	1586	15.29
7	20376	878	23.21
8	20137	566	35.58
9	15580	400	38.95
10	9880	271	36.46
11	27723	275	100.81
12	26938	293	91.94

（２）　ラウンド識別の困難さ

不正ホストによって活動する期間や間隔は異なる。例を図１に示す。この例では不正ホストの活動が 2 回に分かれて行われていると考えられる。その活動の期間をそれぞれd_1、d_2とする。d_1では 4 ヶ月間にわたり不正なパケットを観測でき、その後のtの期間では 100 日間の全く観測できない期間が存在し、またd_2の期間で 2 ヶ月間パケットを観測している。このように長期間パケットが観測できない場合は、その前後のパケットは違うウィルスに感染したと考えるのが妥当である。しかし、単に観測できなかった可能性、プログラムによって活動していない可能性などが否定できない。また、使用ポートによって振る舞いが違うため、同様な解釈を他の不正ホストに対して行うことが難しい。そこで、本稿では次の 3 つの方法によって不正ホストの感染期間を推定してみる

A、　サンプリング

B、　固定閾値

C、　適応閾値

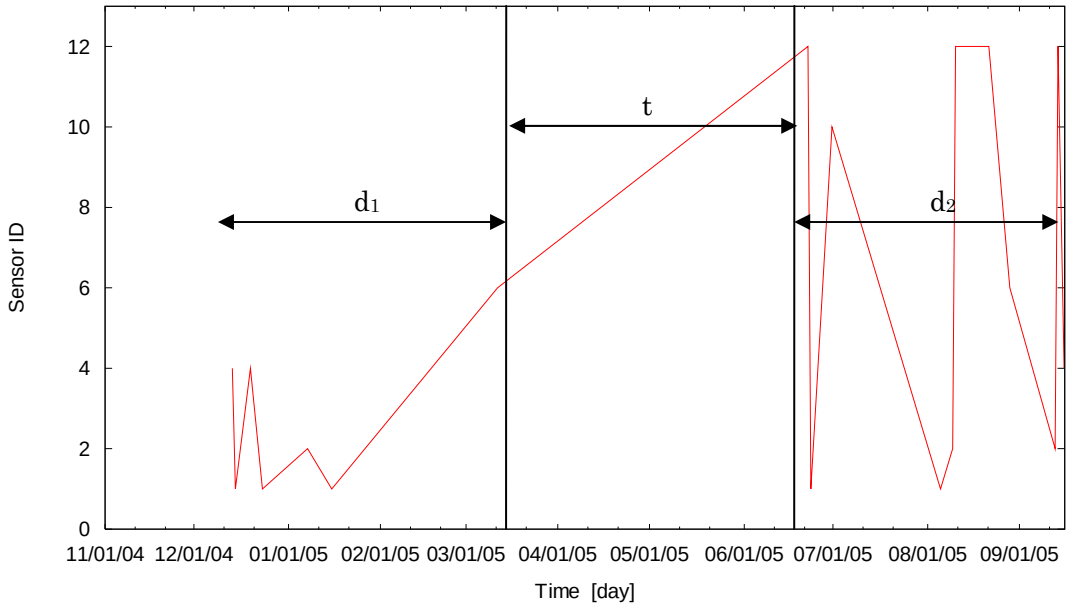

図１　ユニークホストの観測パケット

第3章　提案方式

3.1　サンプリングについて

不正ホストの K=6 の特徴をつかむために、不正ホストをランダムに 100 個抽出し、閾値を T=１、14、30 として時の平均の d と r を推定する。また、主観評価によって、同様の不正ホスト 100 個の活動期間を求める。サンプリングの手順を以下に示し、サンプリングの例を表 2 に、グラフを図 3、図 4 にそれぞれ示す。

１、Java プログラムによって、K=6 の不正ホスト 100 個を抽出する
２、抽出したホストをエクセルのグラフで可視化し、時系列に動きを見る
３、主観評価によって、活動の期間を推定する

表 2　サンプリング例

IP	開始	終了	ポート	ラウンド数	スキャン数 C（回）	ビジット数 k（台）	感染日数（日）	平均感染日数
222.168.200.198	2004/12/16	2004/12/16	4899	1	1	1	1	15.3
	2005/4/26	2005/6/22	ICMP 80	2	10	3	58	
	2005/8/30	2005/8/30	ICMP	3	1	1	1	
	2005/9/15	2005/9/15	ICMP 80	4	2	1	1	
222.122.10.105	2004/10/28	2004/10/29	21	1	6	6	2	2.0

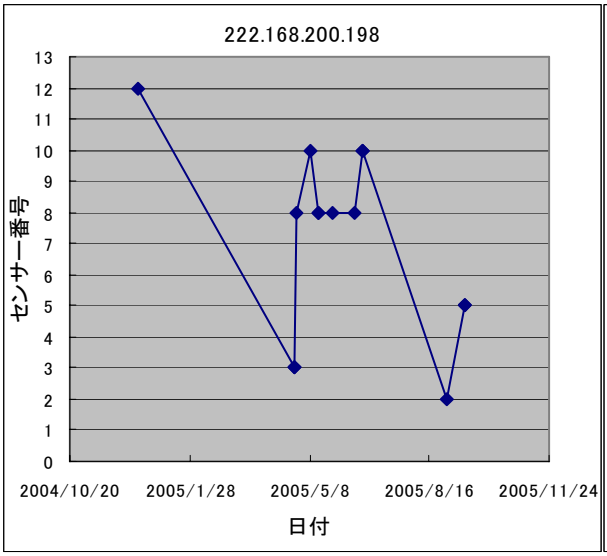

図２　サンプリング例１

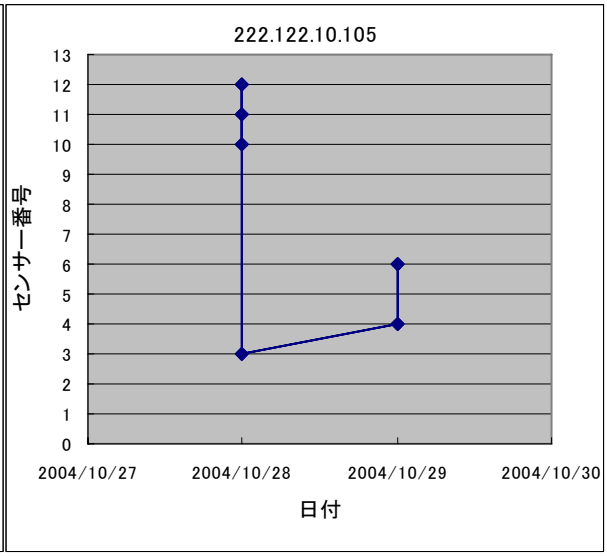

図３　サンプリング例２

図 2、図 3 は明らかにホストの活動に違いがある。図 2 は約 1 年間にわたりパケットを観測できるが、図 3 では 2 日間しか観測することが出来ない。カウント数も図 2 では 14 回観測されているのに対し、図 3 では 6 回と K=6 の集合では最小である。このように、不正ホストの活動はそれぞれに異なる。この特徴をくみとり、ホストの活動を推定するために、固定閾値と適応閾値の 2 つの手法を行う。

3.2　固定閾値について

不正ホストについて，ラウンド間隔 t が定めた閾値 T を超えたときを 1 ラウンドの終結と判断する。ただし、最適な T を同定するのは難しいので T を 0～395 日に変化させた時の r、c、k、d を求めるプログラムによる解析を行う。感染期間はポートやプロトコルに依存することが予想される。この影響についても解析する.

3.3　適応閾値について

ラウンドの間隔はその不正ホストにおけるスキャンパケット数に依存して決められるはずである。そこでパケットの到着間隔がポアソン分布に依存して決められると仮定し、期間 t で全く届かない確率を同定する。

ポアソン分布とは、一定時間に何通のメールが届くのか、交差点に何台の車が通過するのかなど、自然現象に適応したモデルである。各パケットがランダムに独立に発生しているときもポアソン過程であると考えられている。

平均λ回到着するパケットが、単位時間内に k 回到着する確率は

$$P(N=k)=\frac{e^{\lambda}\lambda^{-k}}{k!}$$

で与えられる。

ただし，N はパケット数を取る確率変数で、λはそのホストにおける平均パケット到着率であり、本稿では、単位ホストあたりの年間平均で、

$$\lambda=\frac{c}{d_0}$$

と定義する。cはそのホストの総カウント数，d_0は観測期間における最初と最後のパケットの時間差（日）である。この時、パケットの到達間隔tが閾値Tを超える確率は、

$$P(t>T)=e^{-\lambda T}$$

で与えられる指数分布に従う。逆に、ある閾値T^*を超える間隔でパケットが 1 つも到着しない確率を 1%で与えると、

$$P(t>T^*)=0.01=e^{-\lambda T^*}$$

両辺に対数を取り、適応閾値

$$T^*=\frac{\ln(0.01)}{-\lambda}$$

を定めることが出来る。即ち，ある連続したパケットの間隔がT^*より広いならば、それは 1%未満でしか発生しないごく珍しい事象が起きたことを表しており、その前後を独立したラウンドと考える。

第4章　実験 1

4.1　基本統計量

本稿の解析は、サンプリングデータのK_6の集合 1586 について行う。すでに表 1 にて、K_1～K_{12}のユニークホスト数とカウント数についての動向を示した。これより、ユニークホスト数が多ければカウント数が少なく、ユニークホスト数が少なければカウント数が少ないことが分かり、K_6が平均的な振る舞いであることが考えられる。ただし、K_{12}においてはK_{11}よりカウント数が少なくなっていることに注意がいる。K_{12}にはすべてのセンサに定期的にスキャンする不正ホストが多く含まれており、こういった現象が起きていると考えられる。

まず、K_6全体の統計量として、図 3 にパケット量の多いポート 10 個について示す。特殊なウィルスはポート 135 と 445 を行き来するが、そのような場合もそれぞれでカウントする。また、IPアドレス(/8)の分布を図 4 に示す。これより、クラスBとCが主流であるこが分かる。

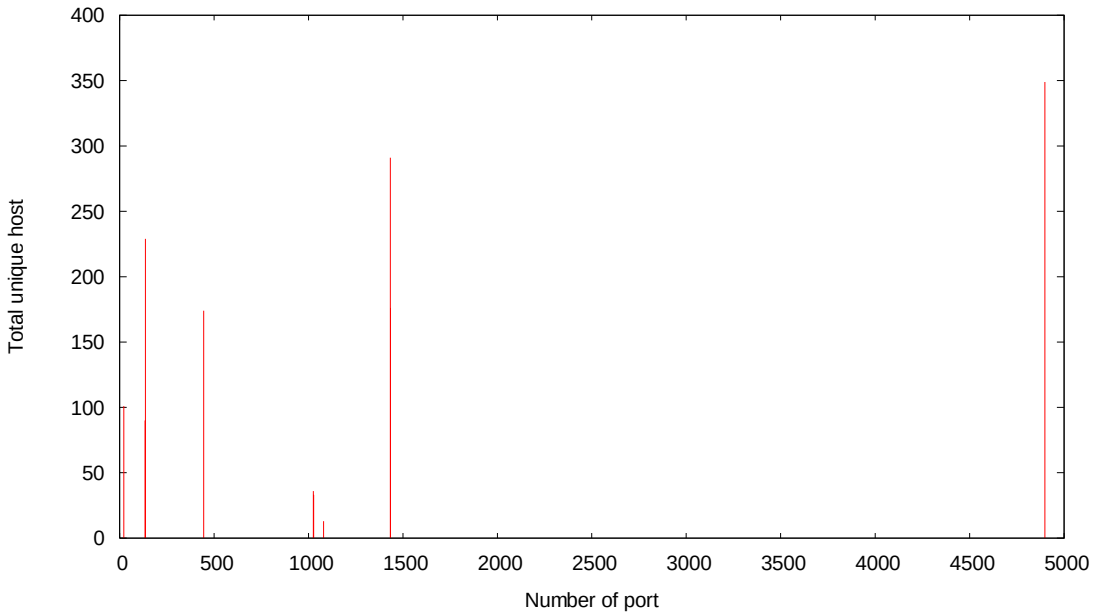

図 3　K_6におけるポートとユニークホスト数

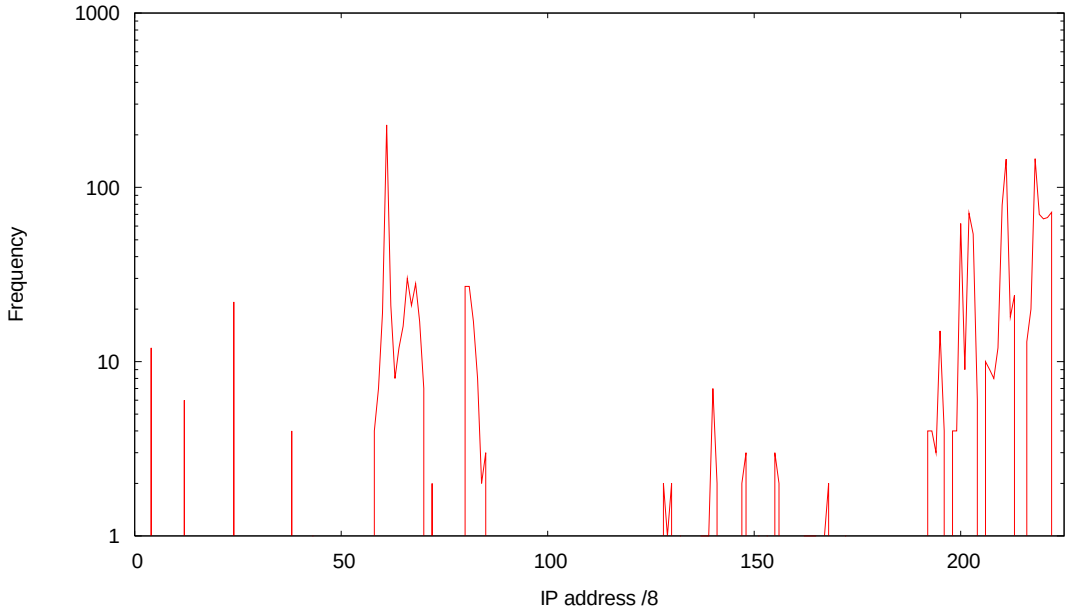

図 4　K_6におけるIPアドレス/8 の分布

4.2　実験結果

4.2.1　サンプリングの解析結果

K_6におけるIPアドレス 1586 において、ランダムに 100 個のIPアドレスを抽出したr、c、k、dのそれぞれの平均と分散を表 3 に示す。

表 3　ランダムサンプリングによる平均と分散

	r ［ラウンド/ホスト］	c ［パケット/ラウンド］	k ［センサ/ラウンド］	d ［日/ラウンド］
平均	1.49	8.72	4.36	24.6
分散	0.81	11.57	1.99	40.79

4.2.2　固定閾値の解析結果

図 5 は閾値 T を 1 日ずつ増加させた時のホストあたりの平均のラウンド数とラウンドごとの感染期間を示している。感染回数は閾値を増加させることによって減少し、最終的には 1 になる。

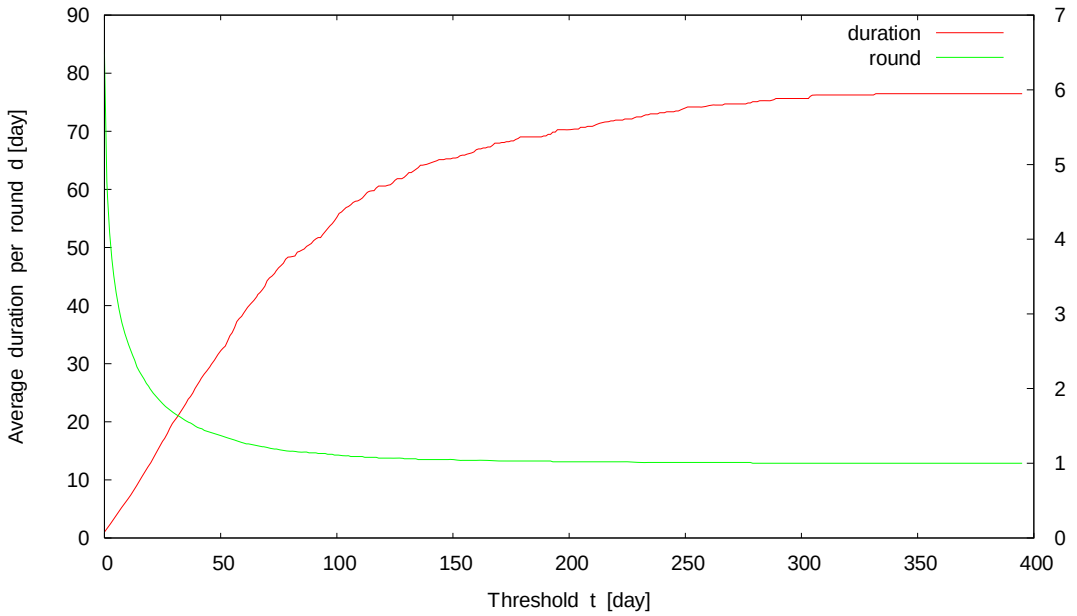

図 5　固定閾値 T についての平均のラウンド数と感染期間

図 6 はK_6の主要ポートにおける閾値を 30 日とした時のポート毎の感染期間を表す。この結果から、主要ポートの感染期間の推移が同じような動きをしていることが分かり、ポートの違いによる感染期間への影響が薄いことが言える。したがって、固定閾値や適応閾値を使って解析をする上では、ポートによる誤差を考慮せずに解析することができることが言える。

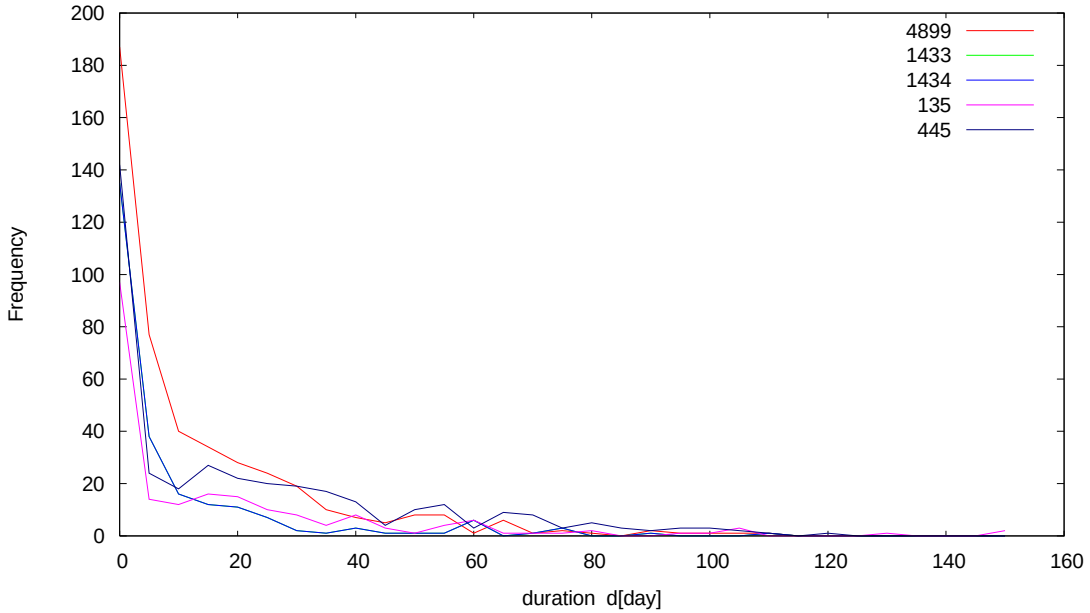

図 6　ポート毎の感染期間の頻度

4.2.3　適応閾値の解析結果

2.4 節の原理に基づき、各ホストに応じたパケット到着確率に適応した閾値を同定する。まず、パケット到着がポアソン過程になっているかを実データで検証する。

図 7 は総カウント数のヒストグラムを表す。

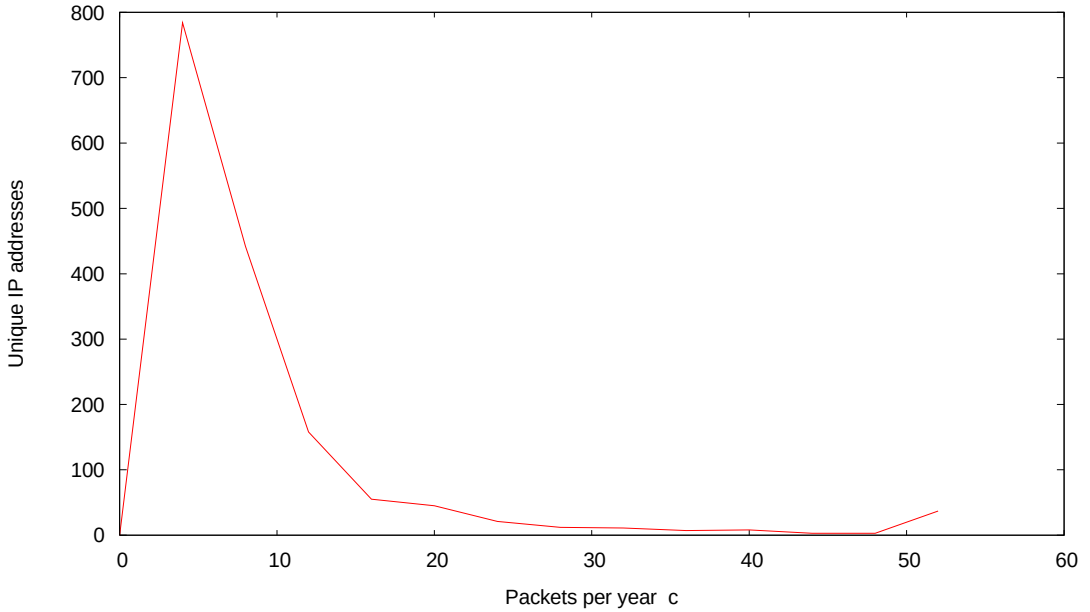

図 7　スキャンのカウント数の分布（K_6）

ビジット数が k=6 のため、最低でも 6 回はスキャンしたことになる。この図より k=6 の場合のスキャン頻度はほとんどが 20 回以下であり、平均は表 2 より、8.72 だが 95%の信頼区間で 6.86～10.57 に分布していることが分かる。しかし、50 回を超えるスキャンをするホストも数は少ないが存在する。これらのことから、ホストによってパケットの到達間隔は一様ではないことが分かる。

そこで、図 8 に代表的なセンサとしてs_{04}についてパケットの到達間隔と指数分布に従った到着間隔を表す。ただし、指数分布に従った到着間隔

$$P(t) = ae^{-\lambda}$$

で表されるので, 定数 a、λ は最小二乗法により求めた a=165、λ=0.005 を用いている。

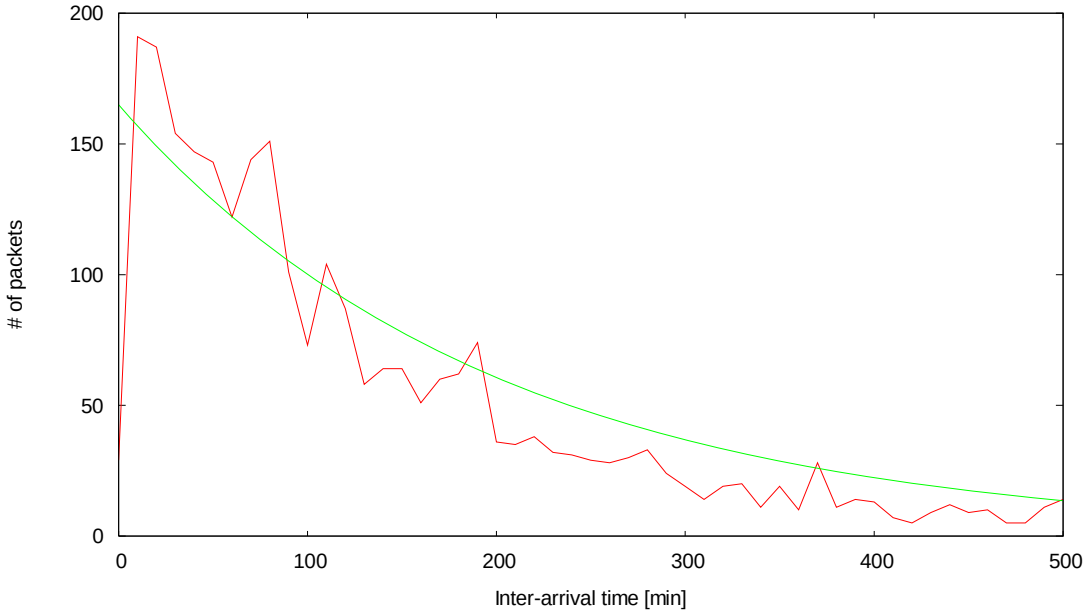

図 8　パケット到着間隔の分布

赤の属性がポアソン分布を表し、緑の属性は指数関数を理論的に近似したことを表す。次のパケットが到着するまでの時間が 10 分以内となるのが最も多く、時間が経過すればするほど到着するパケットの量が少なくなってくる。そして、パケットの到着する確率が 1%を切った時、即ち、99%の確率でパケットが到着したとする時間を過ぎて次のパケットが到着した場合に次のラウンドとなるように閾値を定める。この確率と分布が近似できたことより、理論値と実測値の適合性が分かり、ポアソン分布の過程が成立することが裏付けられる。

この結果を元に、ポアソン過程をすべてのホストに適用し、平均到着率 λ を求めた結果を図 9 に示す。

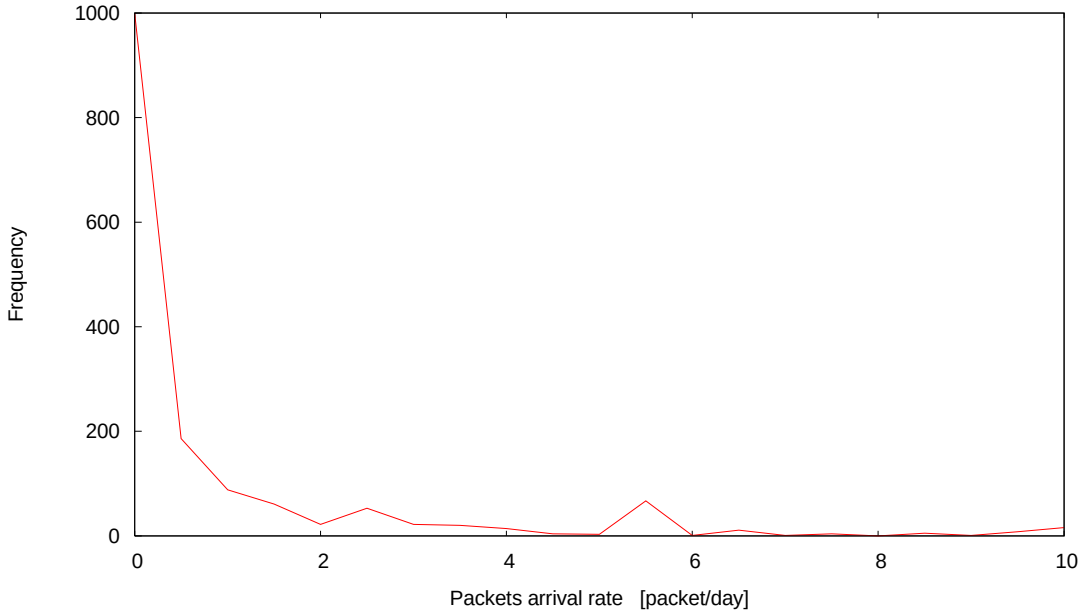

図 9　パケットの平均到着率

この結果より、1000 近くのパケットの到着率が 1 以下であることが分かる。これは、図 7 で示したように、ほとんどの不正ホストのカウント数が 10 前後に集中していることなどから出たと考えられる。しかし、図 7 でも 50 以上のパケットが存在したように、到着率が 5〜6 の間で増えていることから、図 9 にも同じような現象が発生している。

以上の結果から、適応閾値T^*と固定閾値T=30 の時の平均感染期間の対比を図 10 に、ラウンド数の対比を図 11 に示す。

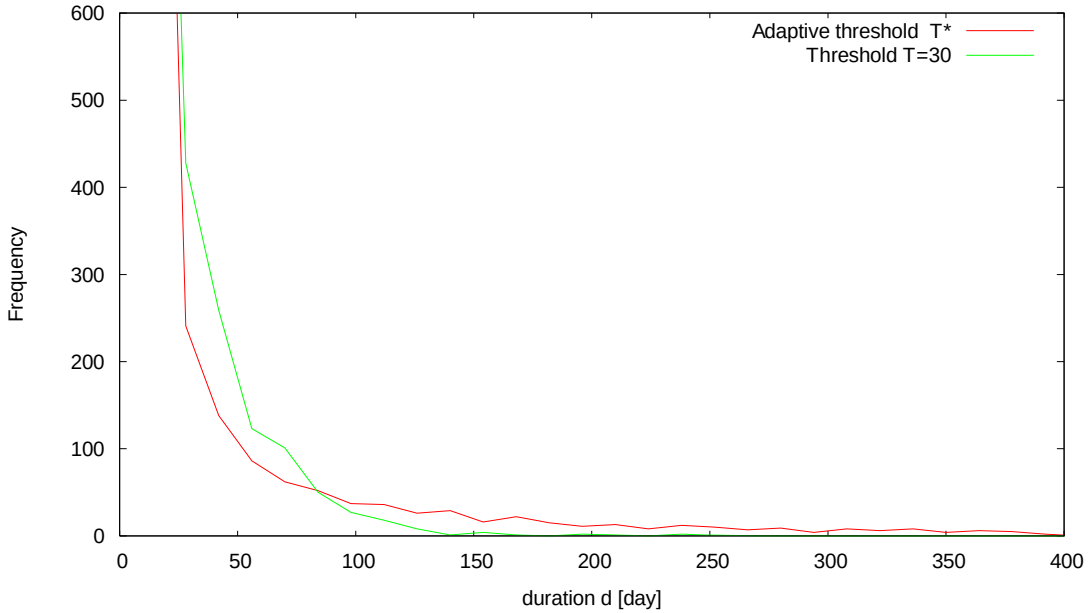

図 10　固定閾値と適応閾値による感染期間 d の分布

図 10 より、0～80 日までの間は固定閾値の方が感染期間の頻度が多いのに対し、それ以降は適応閾値の方が多くなっている。これは、80 日を境にして短いラウンドが減り、長いラウンドが増えたことを表す。これは、適応閾値が活動期間とパケット数から算出した各ホストの活動密度から干すとの特徴を取った結果であると判断でき、固定閾値に比べ、全体的にホストの評価が行えた結果であると考えられる。

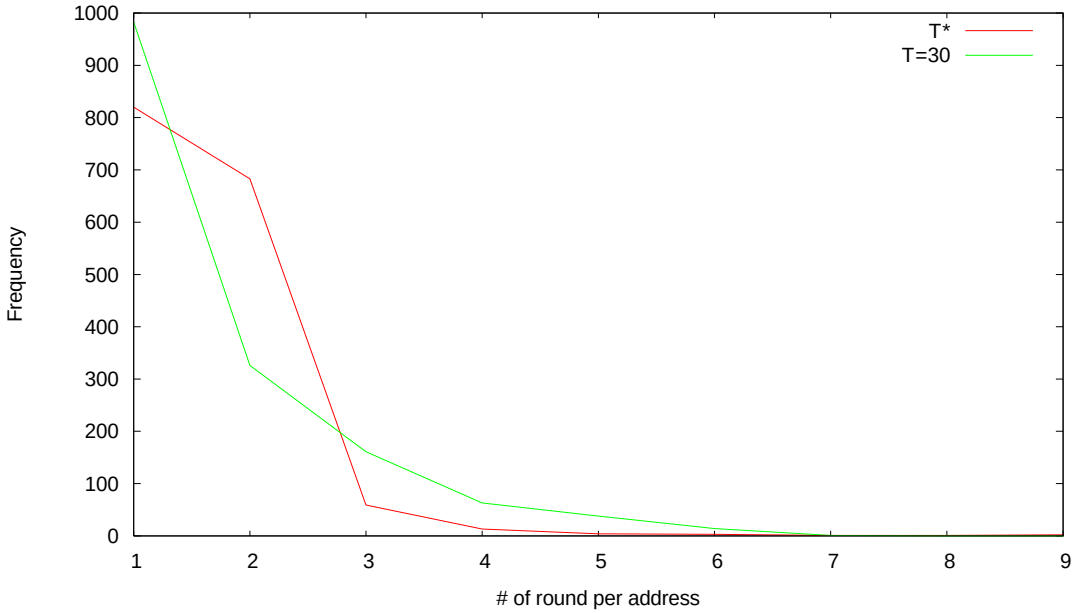

図 11　固定閾値と適応閾値によるラウンド数 r の分布

図 11 では、固定閾値で算出した場合は、適応閾値に比べてラウンド数が 1 と 3 以上の場合に多くなっている。これは固定閾値で算出した場合は、30 日以内に次のパケットを観測した場合には同じラウンドとするが、適応閾値で算出した場合にはたとえ 30 日以内であっても確率密度が高ければ、ラウンド数が増える。逆に 30 日以上経ってパケットが到着した場合でも、確率密度が低ければラウンド数が少なくなる。また、適応閾値で求めた場合は、ほとんどの場合にラウンド数が 1、2 回となり、その数はほぼ等しい。これらの結果からも、固定閾値より適応閾値の方が不正ホストの特徴を組んでいることが言え、単一の閾値ですべてのホストを分けることが難しいことが分かる。

また、ポアソン過程をK_6以外の集合についても適応した結果を図 12 に示す。ただし、解析結果にK_1、K_2は含まれない。

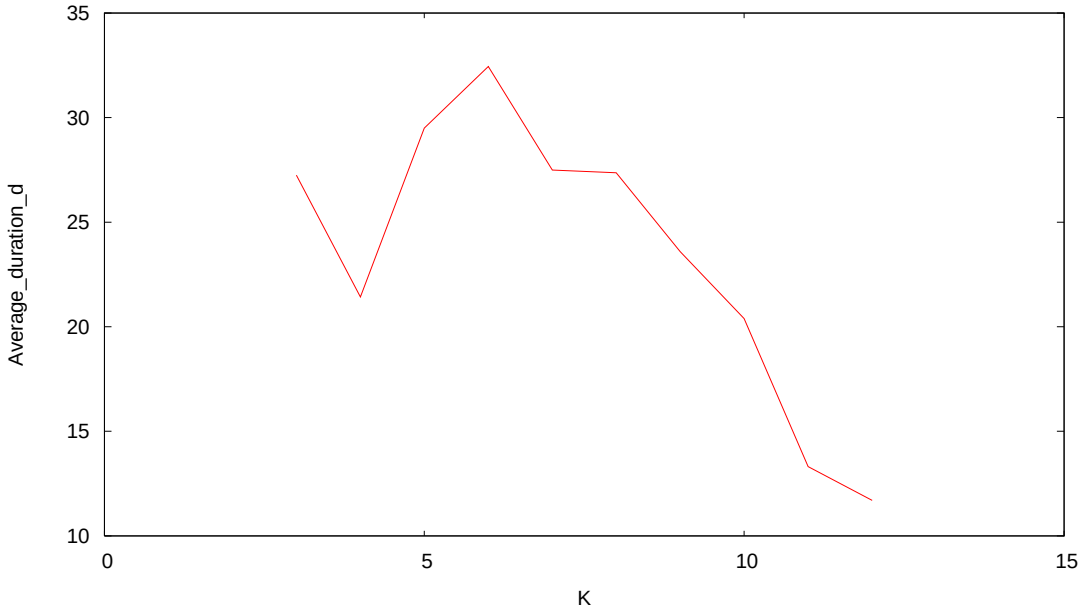

図 12　適応閾値を適応した時の K*の平均感染期間

本来の予想では、Kの値が大きければ大きいほど感染期間が長いと考えていたが、適応閾値によって、解析した結果K_6が最も感染期間が長かった。この結果から、センサの台数とパケットの多さになんらかの関係があるのかもしれない。

また、今回の実験で明らかになった感染期間を表 4 に示す。この結果より、感染期間は 32 日間の寿命があり、ウィルス検出により駆除されるまでに 32 日間かかっている。しかし、平均で 1.5 回感染を繰り返していることが分かった。

表 4　閾値と平均感染期間

	ラウンド r	カウント c	ビジット k	感染期間 d
	μ_r	μ_c	μ_k	μ_d
（1）サンプリング（人手）	1.49	8.72	4.36	24.6
（2）固定閾値	1.67	9.15	3.13	18.2
（3）適応閾値	1.57	9.75	4.32	32.3

第5章　実験 2

5.1　実験方法

学内にセンサを 4 台設置し、そのログデータから一番パケットの多いセンサの設置場所を一番危険な研究室だと判断する。センサの設置場所と IP アドレスを表 5 に示す。ログデータは 7 月の 1 ヶ月間のログデータから日ごとの総カウント数をプログラムより算出し、グラフ化する。また、1 ヶ月間のログデータの総スキャン数が一番多い研究室を、一番危険な研究室とする

表 5　センサの設置場所

センサ名	設置場所	IP アドレス
sensor103	E 棟	150.7.62.108
sensor104	F 棟	150.7.63.35
sensor105	G 棟	150.7.64.155
sensor106	5 研	150.7.50.31

5.2　実験結果

図 13 に 7 月のパケットの流れを示す。7 月 13 日のみすべてのセンサのログデータが多くなっている。これは APIPA による通信が行われていたことが原因である。APIPA は DHCP サーバとの通信が行えないときに自動で PC に割り当てられる IP アドレスである。従って、このときはネットワーク上に通信の以上が起きていたと考えられる。また、7 日は極端にパケットが少なくなっているが、これはシステムが落ちていたのが原因である。

これらのことを踏まえた上での結論として、表 6 に示すように G 棟が一番パケットの量が多く、危険であると判断することが出来る。

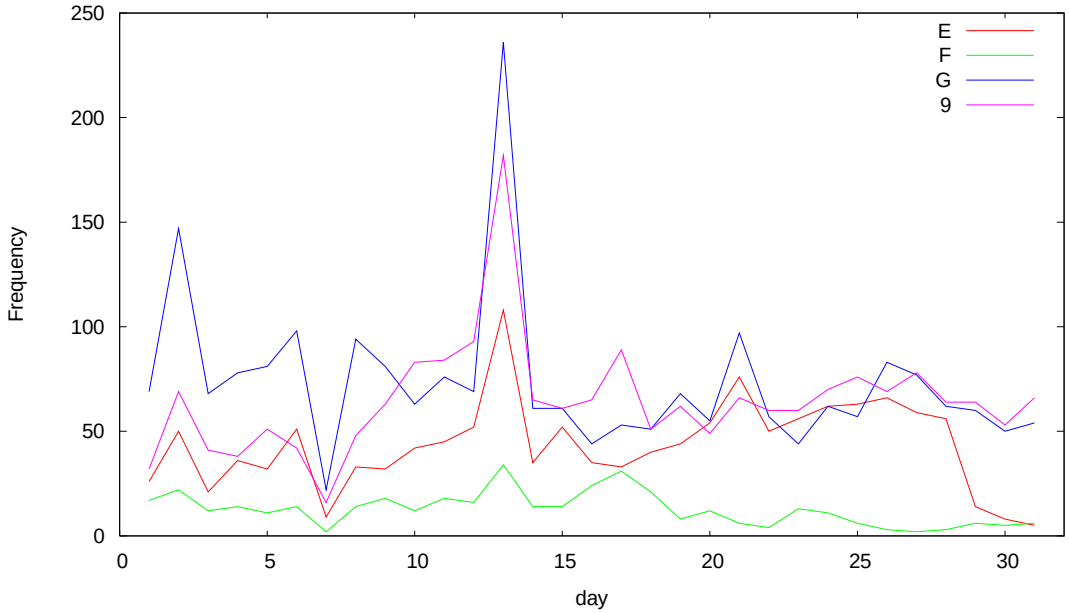

図 13　7 月の各月のカウント数

表 6　7 月の総カウント数

場所	カウント数（回）
G 棟	2278
5 研	2010
E 棟	1345
F 棟	393

第6章　結論・課題

6.1　結論

観測データの解析により、最適な感染期間はそれぞれのIPアドレスによって　異なることが分かり、ポアソン分布における閾値T*によってこの違いを考慮して、最適な感染期間を求めた。

最適な感染期間はそれぞれのユニークホストによって異なり、一概に固定の閾値による分類は当てはまらない。また、フィッティングを行った場合は適応閾値 T の時とは違い、ポートによる影響が大きいことも分かり、それぞれのセンサによっても観測期間に差があることが分かった。

6.2　今後の課題

今回の実験では、適応閾値の確率を 1%で行った。しかし、これは数学的モデルによって求めたものではない。従って、数学的モデルに基づいて、到着確率を求めることが必要である。また、その値がサンプリング値に近いことも必要である。

現在は、プログラムによって、到着確率が 2～3%の時にサンプリング値に近い値があるというところまで分かった。その結果を、表 7 に示す。

今後の課題は、その値を数学的なモデルから求めることである。

表 7　到着確率と感染期間

ln（x）	ラウンド数	カウント数 C	ビジット数 K	平均感染期間
1%	1.57	9.75	4.32	32.30
2%	1.73	8.84	4.01	26.06
3%	1.83	8.34	3.85	23.24
4%	1.93	7.91	3.70	20.37
5%	2.03	7.54	3.58	18.32

参考文献

[1]小堀, 他, ISDAS 分散観測 : ウィルスの平均寿命はいくらか？　情報処理学会, CSS2006, pp.519-524，2006.

[2]戸田，他，ISDAS: Internet Scan Data Acquisition System　情報処理学会，CSS2004, pp.199-204，2004.

[3]福野，他，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　情報処理研報，Vol.2006，NO.81，2006-CSEC-34，pp,299-304，2006.

[4] 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，コンピュータセキュリティシンポジウム(CSS 2005)，pp,421-426，2005.

謝辞

本研究を行うにあたり、暖かいご指導を受け賜りました東海大学電子情報学部情報メディア学科菊池浩明教授に深甚なる感謝を申し上げます。

また，常に貴重な助言をして頂いた菊池研究室大学院生の福野直弥様に厚く感謝を申し上げます。共同研究者として，研究にお力を貸して頂いた寺田真敏様，土居範久様，中央大学土居研究室の杉山太一様には厚く感謝を申し上げます。