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

2007－ＣＳＥＣ－38 (36)
2007／7／20

不正ポートスキャンパケットの直交展開

小堀 智弘† 菊池 浩明† 寺田 真敏††

† 東海大学工学研究科情報理工学専攻部 〒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
E-mail: †{nopay,kikn}@cs.dm.u-tokai.ac.jp

あらまし インターネットの上に観測されているポートスキャンには，時間，ポート番号，発信と宛先アドレスなど多くのパラメータが絡んでいる．多くの不正コードから狙われる有名なポート番号がいくつか知られているが，ポート番号と不正コードの関係は明らかにされていない．そこで，本研究では，複数のセンサで分散観測されたポートスキャンの特徴を取り出す新しい方式を提案する．提案方式の特長には次の特徴がある．1) 可能性のあるすべてのポート番号を考慮するのではなく，直交展開された少数の重要な成分からのセンサの分析，2) 直交成分の線形結合による観測パケットデータの圧縮．3) ポート番号間の統計上の相関関係から，任意のセンサの任意のポートのスキャン数の近似．また，実測したパケットデータについて提案方式の精度評価する．
キーワード 直交展開，ポートスキャン，不正アクセス数の復元

Orthogonal expansion of port-scan packets

Tomohiro KOBORI†, Hiroaki KIKUCHI†, and Masato TERADA††

† Couse of Information Engineering, Graduate School of Engineering Tokai University 1117 Kitakaname, Hiratsuka, Kanagawa, 212-8567 Japan
† Hitachi, Ltd. Hitachi Incident Response Team (HIRT) 890 Kashimada, Kawasaki, Kanagawa, 212-8567 Japan
E-mail: †{nopay,kikn}@cs.dm.u-tokai.ac.jp

Abstract Observation of port-scan packets performed over the Internet is involved with so many parameters including time, port numbers, source and destination addresses. There are some common port numbers to which many malicious codes likely use to scan, but a relationship between port numbers and the malicious codes are not clearly identified. In this paper, we propose a new attempt to figure characteristics of port-scans observed from distributed many sensors. Our method allows 1) analysis of sensors with few significiant factors extracted from an orthogonal expansion of port-scan packets, rather than taking care of all possible statistics of port numbers, 2) compression of packets data, computed by linear combination of limited number of orthogonal factors, and 3) approximation of number of scanning packets at arbitrarly specified sensor and ports, made from statistical correlation between port numbers. We also evaluate the accuracy of our proposed approximation algorithm based on actually observed packets.
Key words orthogonal expansion, port-scan, reconstruction of malicious access

1. まえがき

不正ホストによるポートスキャンは頻繁に起こっている．それらは人，ウィルス，ボットなど多種多様な方法によって行われいている．これらの不正パケットの挙動を知る手段として定点観測システム[2]がある．定点観測データから不正アクセスの全域的な振る舞いを正しく分析するためには，次に挙げるいくつかの課題がある．

(1) センサは属するネットワークの影響を多く受けるので，センサのアドレスが十分に分散している必要がある．例えば，石黒らは，ポートスキャンに感染したホストを中心した局所性があることを指摘している[3]．あて先ポートの偏りは，不正ホスト数を同定する際に大きな誤差の原因になる[4]．

(2) 不正ホストのアクセスポートはワームの種類によっても

異なる．Blasterのように，主にローカル内でスキャンを繰り返すものもあればSasserのように，ローカルとグローバルをランダムに行き来するワームも存在する．ポートとワームの関係も自明ではなく，135や445のように多くのワームが用いる一般的なポート番号や逆にポート番号からワームが特定できる特殊なものもある[8]．ポート番号空間も独立ではなく，135ポートを狙うワームは，445ポートも対にしてスキャンすることが多いなど，いくつかのポートには強い相関があることが知られている．従って，やみくもに16bitのすべてのポートを見る必要はなく，主要なポートに焦点を絞って考えればよい．例えば，ISDAS[2]では，経験に基づいて80，135，137，139，445，ICMPの6つのポートのパケット数のみを提示している．[1]で福野らは，TF-IDF値[5]について主要8ポートを選んでいる．

(3) ネットワーク管理者はポートフィルタリングを行うなどの対策をしている現状である．ネットワークインシデントの軽減は図れるが，定点観測の立場では，パケット数が人工的に激減してしまい，正しい統計データを得ることが難しくなる．

そこで，本論文ではセンサ間で生じるポート番号ごとの観測パケット数に注目する．攻撃されるポートはランダムではなく，相関があることを仮定し，各ポートの線形結合の分散を最大化するパケット数の固有ベクトルを求める．そして，定点観測システムによって得られたログデータの直交基底を求め，その応用方法を3つ提案する．

(1) 観測値のスペクトラム解析
直交展開した係数がそのセンサの大きな特徴を表していることを利用し，センサの分布や時刻による変化，アドレスとの相関などを解析する．

(2) 観測ベクトルの近似
画像に対する離散コサイン変換とjpegへの応用と同じ直交展開の原理で，観測ベクトルの圧縮や近似ができることを示す．

(3) 欠損データの補完
パケットフィルタリングなどにより，欠損するポートのデータに対して，他のポートの観測値から予測を行う

本論文ではまず，第二章で，本研究を行う上での基本定義と研究の提案方式を示す．第三章で，研究の目的とセンサの近似・補完を行った結果を示す．第四章で，結論と今後の課題について述べる．

2. 提案方式

2.1 基本定義

不正ホストとは，ウィルスやワーム等に感染して，他のホストに不正パケットを飛ばすホストを言う．センサとは，不正パケットを観測する正規ホストである．センサの台数をn，センサの集合は$S=\{s_1,s_2,\ldots,s_n\}$とする．

観測されたログデータは，あて先ポートと送信元IPアドレス，観測日時，センサID，プロトコルから成る．全センサで観測された不正パケットの総和の上位nポートをポート集合$P=\{p_1,p_2,\ldots,p_m\}$とする．また，IPアドレスは4つのオクテットから成っているが，本研究では上位2つのオクテットまでを解析の対象とする．

2.2 直交基底と展開

各センサにはそれぞれに異なった特徴が存在する．そのため，その違いを加味した区分けをする必要がある．そこで，本研究ではセンサごとの特徴を抽出するために，正規直交基底によって，センサの識別を試みる．

センサの集合Sと宛先ポートの集合Pについて集計した観測パケット数を$S\times P$行列

$$A=\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}=(\boldsymbol{a}_1,\ldots,\boldsymbol{a}_n)$$

で表す．例えば，センサs_iがポートp_jで観測したパケット総数がa_{ij}である．また，$\boldsymbol{a}_i$はセンサs_iによるPの上の列ベクトルを表しており，センサs_iの観測ベクトルと呼ぶ．

ポートp_jにおけるn台のセンサの平均を$\bar{a}_j=\frac{1}{n}\sum_{i=1}^{n}a_{ij}$と置き，全ポートまとめて平均列ベクトル

$$\boldsymbol{g}_0=\begin{pmatrix} \overline{a_1} \\ \vdots \\ \overline{a_m} \end{pmatrix}$$

で表す．Aの各要素を$\boldsymbol{g}_0$から引いたベクトルを$\boldsymbol{a}_j'=\boldsymbol{a}_j-\boldsymbol{g}_0$をまとめて，$A'$とする．これは，平均からの差異を表している．

次にA'を用いて行列

$$\tilde{M}=\sum_{j=1}^{m}\boldsymbol{a}_j'\cdot\boldsymbol{a}_j'^{\mathsf{T}}$$

を定義する．ただし，$\boldsymbol{a}_j'^{\mathsf{T}}$は$\boldsymbol{a}_j'$の転置である．$\tilde{M}$の固有値を降順にソートしたものを$\lambda_1,\ldots,\lambda_m$，それに対応する単位固有ベクトルを$\boldsymbol{g}_1,\ldots,\boldsymbol{g}_m$とする．これらは対称行列の固有ベクトルから作られているので，正規直交基底になる．すなわち，任意の$i\neq j$となる$\boldsymbol{g}_i$と$\boldsymbol{g}_j$の内積が

$$(\boldsymbol{g}_i,\boldsymbol{g}_j)=0$$

(直交性)を満たし，すべての$\boldsymbol{g}_i$が単位ベクトルであり(正規性)，任意のm次のポートについての観測ベクトル$\boldsymbol{a}$が，$\boldsymbol{g}_1,\ldots,\boldsymbol{g}_m$の線形結合

$$\boldsymbol{a}=\boldsymbol{g}_0+c_1\boldsymbol{g}_1+\cdots+c_m\boldsymbol{g}_m \qquad (1)$$

で一意に表すことができる(基底)．この式(1)を，観測ベクトル$\boldsymbol{a}$の基底$\boldsymbol{g}_0,\boldsymbol{g}_1,\boldsymbol{g}_2,\ldots$による直交展開という．これは，各センサを平均センサで近似し，正規直交基底で表れた特徴を補正していくことを表す．$\boldsymbol{a}$に対する最小二乗誤差

$$(\boldsymbol{a}-(\boldsymbol{g}_0+c_1\boldsymbol{g}_1+\cdots+c_m\boldsymbol{g}_m))^2$$

を最小化する係数c_iは$i=1,\ldots,m$について，

$$c_i=(\boldsymbol{a},\boldsymbol{g}_i)$$

で与えられる．

2.3 観測ベクトルのスペクトル解析

基底の中でも，全観測ベクトルとの誤差 $|\sum_{i=1}^{n} a_i - cg_j|$ を最小化するものは，M の最大固有値に対する固有ベクトル，すなわち，g_1 であることが知られている．従って，式 (1) の直交展開における係数 $c_1, c_2, \ldots$ は c_1 から順に識別の重要度の高い成分である．特に影響のある g_1,g_2 基底の係数 c_1, c_2 が，そのセンサの大まかな特徴を表していると考えられる．そこで，c_1 , c_2 の 2 軸についてセンサの散布を表現する．例えば，センサの設置環境が ISP，大学，ケーブルテレビとそれぞれ異なるプロバイダであるとき，観測ベクトルに生じる差を c_1，c_2 で判別できることが期待できる．

2.4 観測ベクトルの近似

直交展開されたすべての成分 $c_1, \ldots, c_m$ を用いると，任意の観測ベクトル$\boldsymbol{a}$が誤差なく近似できる．ここで，高次の成分の影響が大きいことを考えれば，一部の成分のみによる近似で原センサの特徴が十分に表れていると考えられる．観測ベクトル $\boldsymbol{a}$ の第 k 近似は，$\boldsymbol{a}$ の直交展開

$$\boldsymbol{a}^{(k)} = \boldsymbol{g}_0 + \sum_{i=1}^{k} c_i \boldsymbol{g}_i \qquad (2)$$

と定める．ここで，c_i は直交基底 $\boldsymbol{g}_0$,$\boldsymbol{g}_1, \ldots, \boldsymbol{g}_k$ による直交展開の m 個の係数である．c_i と $\boldsymbol{a}$ の要素 a_j の大きさが，$|\, c_i \,| \approx |\, a_j \,|$ であると仮定すると，第 $m/2$ 近似では観測データを約半分に圧縮したことを意味する．この考えに基づき，観測ベクトルを任意の割合で圧縮することができる．

2.5 欠損データの予測

すべてのセンサが同一の環境でパケットを観測できているわけではない．例えば，ポート 135 や 445 は脆弱性を持つことが多いのでファイアウォールで遮断している環境は多い．これらの遮断されたパケットを他のポートのスキャン頻度から予測する．

ポート p_j が欠損しているベクトル $\boldsymbol{a}'$ を，観測ベクトル $\boldsymbol{a}$ が与えられている時，

$$\boldsymbol{a}' = (a_1, \ldots, a_{j-1}, 0, a_{j+1}, \ldots, a_m)^{\mathsf{T}}$$

と定式化する．この a_j を残りの $a_1, \ldots, a_{m-1}$ 個のデータから補完しよう．

式 (1) より直交展開できるので，$\boldsymbol{a}'$ は，

$$\boldsymbol{a}' = \boldsymbol{a} - x\boldsymbol{e}$$

と表すことができる．ここで，$x = a_j$ で $\boldsymbol{e}$ は j 要素のみ 1 の単位ベクトル $(0, \ldots, 1, \ldots, 0)^{\mathsf{T}}$ である．基底の直交性より，

$$\begin{aligned}(\boldsymbol{a}', \boldsymbol{g}_1) &= (\boldsymbol{g}_0, \boldsymbol{g}_1) + c_1(\boldsymbol{g}_1, \boldsymbol{g}_1) + c_2(\boldsymbol{g}_2, \boldsymbol{g}_1) + \ldots + x(\boldsymbol{e}, \boldsymbol{g}_1) \\ &= (\boldsymbol{g}_0, \boldsymbol{g}_1) + c_1(\boldsymbol{g}_1, \boldsymbol{g}_1) + x(\boldsymbol{e}, \boldsymbol{g}_1)\end{aligned}$$

同様に，$i = 1, \ldots, m$ について，

$$(\boldsymbol{a}', \boldsymbol{g}_i) = (\boldsymbol{g}_0, \boldsymbol{g}_i) + c_i \parallel \boldsymbol{g}_i \parallel^2 + x(\boldsymbol{e}, \boldsymbol{g}_i)$$

が成立し，$c_1, \ldots, c_m$ を x から成る m 個の関係式が得られる．そこで，

表 1 東海大センサ $s_{(7)}$，$s_{(8)}$ の観測パケット数

ポート	$s_{(1)}$	$s_{(2)}$
135	74508	68553
445	0	0
–	0	0
139	0	0
80	356	800
1026	10025	47864
1433	3168	3524
1027	1578	17648
1434	1627	1461
4899	555	1103
137	0	0
23110	0	0
1025	61	52
22	735	606
1028	25	69
1029	24	45
113	0	0
3389	54	134
1030	16	40

$$(\boldsymbol{a}', \boldsymbol{e}) = (\boldsymbol{g}_0, \boldsymbol{e}) + c_1(\boldsymbol{g}_1, \boldsymbol{e}) + \ldots + x \parallel \boldsymbol{e} \parallel^2$$

よって，x についてこれらの連立方程式を解くと

$$x = \frac{(\boldsymbol{g}_0, \boldsymbol{e}) - \sum_{i=0}^{m}\{(\boldsymbol{a}', \boldsymbol{g}_i)(\boldsymbol{e}, \boldsymbol{g}_i) + (\boldsymbol{g}_0, \boldsymbol{g}_i)(\boldsymbol{e}, \boldsymbol{g}_i)\}}{1 - \sum_{i=1}^{m}(\boldsymbol{e}, \boldsymbol{g}_i)^2} \qquad (3)$$

が得られる．(3) 式は，基底が完全な観測ベクトル $\boldsymbol{a}$ から算出された時は，x が一意に正確に決まることを表している．しかし，欠損した観測データ $\boldsymbol{a}'$ しか与えられている時は，誤差を生じて a_j に近づく．この誤差の大きさは正常なセンサの数に依存して決まる．逆に言うと，十分な数のセンサがあれば，大数の法則で誤差を無視できるほど小さく出来ることが期待できる．

3. 実　　験

3.1 観測データ

不正ホスト動向を解析をするためには十分な観測期間，独立した複数のセンサをネットワークに設置する必要がある．本研究では，JPCERT/CC によって運営されている ISDAS システム [2] によって得られた観測データ (2005 年 10 月 1 日〜2006 年 3 月 30 日、$m = 30$) と，比較対象として，東海大学で分散観測したデータを用いる．後者のデータは，2006 年 11 月 30 日〜2007 年 5 月 2 日，$m = 8$ である．また，観測ベクトルの例を表 1 に示す．図 2 上の (1)，(2) をそれぞれ $s_{(7)}$，$s_{(8)}$ とする．これらのセンサの中にはポートフィルタリングや，DHCP による IP アドレスの変動の影響を受けるものが混在していることに注意が必要である．

3.2 目的と方法

本実験の目的は提案する直交展開によって，センサの特徴が表現でき，欠損などの不完全なデータの近似が可能であることを確かめることである．

そこで，次の方法で実験を行う．

（1） 観測ベクトルの直交基底を算出する．実観測データに，提案方式を適用して直交基底を同定し，各センサの観測ベクトルを直交展開する

（2） スペクトルから，センサの分布や時間推移などの相関を分析する．得られた各基底の係数 $c_1, \ldots, c_m$(スペクトル) を用いて分析を行う．

（3） 観測データを直交展開し，成分からパケット数の近似をできることを示す．成分の一部だけを用いて観測データを合成し，真値との平均二乗誤差を求める．

（4） パケットフィルタリングされて特定のポートの情報が欠損したデータから，観測値を予測する．欠損ポートは，原データから $p_2(445)$ であり，観測ベクトルの a_{i2} を 0 にして，提案方法により予測して真値との比較を行う．

3.3 基　　底

全センサの観測ベクトルから求めた直交基底の一部を表 3 に示す．表 3 より，第一基底 g_1 では，ポート 135，445 が，第二基底 g_2 では，ポート 80，ICMP の影響が各々強いことが分かる．ポート番号はセンサの平均パケット数 ($= g_0$) について並び替えている．ほとんどのセンサでこれらのポート 135，445，80，ICMP のパケットが観測できているので，この基底は観測データの特徴を適切に表していることが言える．多くのワームが 135 と 445 の 2 つのポートを対にして攻撃するという報告 [6] にも矛盾しない．一方，139 は多くの基底で主成分のひとつになっており，ワームによって使い方が分かれていることが予測される．

3.3.1 スペクトルによる分析

a） センサの分布

表 3 の直交基底について展開した c_1, c_2 についてのセンサの分布図を図 1 で示し，その値を．表 3 に示す．同様に，東海大のセンサの散布図を図 2 で表す．

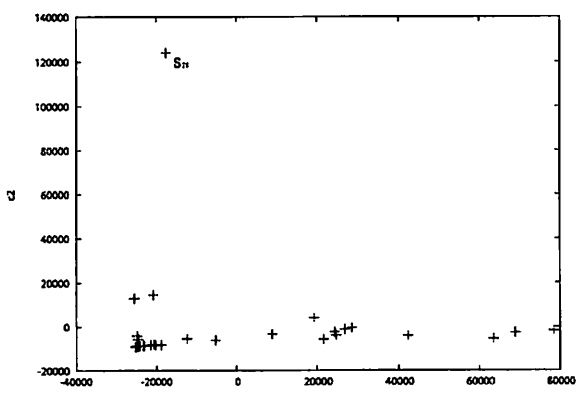

図 1 c_1 と c_2 軸上のセンサの散布図 (ISDAS，$m = 30$)

図 2 には，2 つの商用 ISP からの ADSL による観測データ (1)、(2) と，学内に設置したセンサ群 (3) の 3 つのクラスタが見える．前者 (1)、(2) は DHCP によって不定期にアドレスが変わり，後者 (3) はいくつかの主要なポートがパケットフィルタリングをされている明らかな違いが観測できる．それ

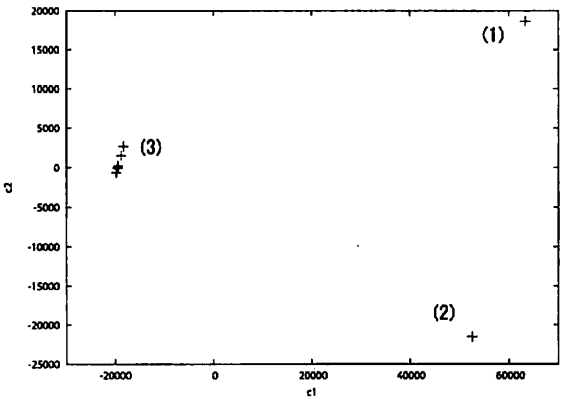

図 2 c_1 と c_2 軸上のセンサの散布図 (東海大，$m = 8$)

に対して図 1 では，左上に 1 点他のセンサ集合から外れたセンサ s_{26} が存在しているが，他の 29 センサは c_1 について広く分布している．g_2 主成分は，ICMP と 80 なので，この例外的なセンサはボットネットなどの集中的な攻撃を受けている可能性がある．

b） 時系列推移

図 3 より，センサの観測するパケットの月ごとの c_1，c_2 成分の推移を表す．時系列を図中の矢印で示している．これより，月ごとの主要なポートの増減が観測できる．例えば s_{01} は 2005 年 10 月から 4ヶ月間単調に c_1 成分 (主に 135，445) を減らしていき，2006 年 2 月から 3 月にかけて急に c_2 成分 (ICMP と 80) が増加している．従って，s_{01} の属するネットワークは,2005 年 10 月と 2006 年 3 月ではまったく違うワームの影響を受けていることが考察できる．

また，s_{01} の c_1〜c_5 の各月の変動を図 4 に示す．図 4 から c_1 の変動が激しく，正から負に大きく振れている．このことからも，ネットワークの環境が変化していることが言える．

c） センサアドレスと c_1 分布

図 5 に，IP アドレス空間における全センサの c_1 成分の大きさを示す．ただし，センサのアドレスは第一オクテットのみの概算値である．ポートスキャンに局所性があるならば，IP アドレスと c_1 成分との間に強い相関が予想されたが，図 5 より，近いアドレスでも正や負の成分が混在し，ほぼ無相関であることが分かった．他の c_2，c_3 成分についても調査したが，ほぼ同様の結果であった．

3.3.2 圧縮と復元

図 6 にセンサ s_{01} の観測ベクトル $\boldsymbol{a}_1$ と第 1 近似 $\boldsymbol{a}^{(1)}$，第 5 近似 $\boldsymbol{a}^{(5)}$ の結果を示す．第 5 近似の時点でかなりの近似ができていることがわかる．しかし，各ポート番号で元データとの誤差が見える．そこで，近似の精度を見るために，各ポートについての近似値を表 4 に示す．更に，式 (2) の k についての原データとの平均二乗誤差 (MSE) を図 7 に図示する．

これらの結果より，誤差は近似の階数に従って単調に減少するが，第 5 近似では十分でないことが分かる．これより，近似をする場合は第 10 近似あたりまで計算する必要があると言

表 2　全センサの基底

ポート		g_0	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}
135	p_1	19499.7	**0.803**	-0.017	-0.103	0.057	-0.131	**-0.457**	**-0.3**	0.111	-0.076	-0.025
445	p_2	15326.5	**0.58**	0.044	0.033	-0.111	0.024	**0.663**	**0.366**	-0.214	0.089	-0.041
ICMP	p_3	6537.4	-0.034	**0.872**	-0.002	-0.114	-0.184	0.001	-0.097	-0.014	0.001	0.036
139	p_4	5778.2	0.069	0.234	-0.224	**0.421**	**0.627**	-0.251	**0.396**	-0.144	0.052	0.035
80	p_5	3865.9	-0.008	**0.332**	-0.135	0.131	0.269	0.292	**-0.488**	0.06	-0.096	-0.074
1026	p_6	3706.0	0.084	0.054	**0.683**	**-0.35**	**0.42**	-0.113	-0.038	0.027	-0.172	0.046
1433	p_7	2423.3	0.045	0.081	-0.004	-0.016	-0.001	0.114	**0.3**	**0.871**	-0.203	0.187
1027	p_8	1268.7	0.03	0.006	**0.326**	-0.124	0.266	-0.082	-0.109	0.015	0.244	-0.159
1434	p_9	1130.9	0.018	0.007	0.003	0.025	0.048	0.026	-0.028	0.039	0.364	-0.069
4899	p_{10}	1007.9	0.015	0.028	0.001	0.025	0.01	-0.025	0.026	0.275	**0.571**	-0.2
137	p_{11}	989.5	0.036	-0.002	0.042	0.023	-0.033	-0.028	-0.144	-0.057	**0.405**	**0.779**
23310	p_{12}	789.6	0.02	0.049	**0.568**	**0.762**	-0.287	0.083	0.013	-0.006	-0.047	-0.023
1025	p_{13}	713.1	0.018	0.008	-0.017	-0.032	-0.04	0.034	0.049	-0.197	**-0.368**	**0.348**
631	p_{14}	552.2	-0.021	0.246	0.12	-0.218	**-0.361**	**-0.338**	**0.446**	-0.167	0.098	-0.1
22	p_{15}	470.5	0.004	0.006	0.003	-0.007	-0.005	0.041	0.06	0.052	0.19	0.136
1028	p_{16}	265.0	0.004	0	0.043	-0.031	0.037	0.012	0.021	-0.012	0.064	0.261
1029	p_{17}	183.4	0.003	-0.001	0.031	-0.021	0.028	0.009	0.015	-0.018	0.04	0.179
113	p_{18}	174.6	0.001	-0.002	-0.063	0.076	0.123	-0.21	0.199	0.015	-0.139	0.085
3389	p_{19}	164.4	0.002	0.003	-0.006	0.002	-0.003	-0.005	0.014	0.082	0.105	-0.065
1030	p_{20}	154.4	0.001	0.003	0.038	-0.02	0.025	-0.009	0	-0.012	0.056	0.12

表 3　ISDAS センサの係数

	c_1	c_5	c_{10}	c_{20}
S_1	78476.01408	-1332.291209	-65.06091315	0.597504165
S_2	42326.3214	-2053.993984	-265.7648826	-0.134574995
S_3	68876.37934	1948.973217	-878.834608	0.027751055
S_4	-23303.95483	-1781.194349	-207.4363615	-3.278791425
S_5	26783.23649	-4491.728714	-7.645881314	0.022692525
S_6	-5218.143885	6276.877551	230.1868143	-0.446194135
S_7	-25539.21478	-8468.61131	-113.5766946	0.216678735
S_8	-24905.4032	-2583.764547	328.3779576	-0.888484215
S_9	24630.03873	-1083.54413	1610.023826	-0.558598995
S_{10}	21600.38661	-2034.629418	-910.9118256	-0.991403865
⋮				
S_{30}	-20838.71956	36303.2633	3129.592697	1304.989657

える．

3.3.3　欠損値の補完

実データのポート p_2(=445) の到着パケット数を欠損しているとみなして 0 とし，式 (3) を適用して，センサ s_{01} において欠損データ x の補完をした結果を図 8 に示す．同様に全センサに適用して求めた実データとの差を図 9 に，それらの誤差のヒストグラムを図 10 に示す．ただし，図 8 における 445 ポートの予測値 $a_{12} = x$ とし，他のポートの値は，$c_1 = \boldsymbol{a}^{(10)}\boldsymbol{g}_i - g_0 g_1$ で近似した c_i から c_{10} までの値を求め，式 (2) で合成して求めている．

図 8 より，欠損したパケット (445 ポートの値) の補完はできている．それは他のポートの誤差と比べて無視できるくらい小さい．s_{01} 以外の観測ベクトルについてもほとんどの場合で十分な補完ができており，図 9 より，30 台のセンサ中 28 台が実データを近似できていると言える．

補完の精度を表 5 に整理する．ただし，前述した明らかな例外センサは除いている．平均誤差は式 (3) より明らかに g_0 へ収束している．その標準偏差より，提案補完値の誤差は $\pm 2\sigma$(±6604) であり，これはポート 445 の平均値 ($\mu(a_{12})$) に対して 43%の大きさである．

3.4　考　　察

大きな誤差が生じているのは，s_{26} と s_{30} であり，他のセンサと比較しても明らかに異常なセンサである．これは，対象としたポート 445 の影響がほとんどなく，他のポートのパケットが大きな影響を与えていることなどの原因が考えられる．センサの観測環境は明らかにされていないが，例えば，ポート 445 のみフィルタリングアウトされている可能性が考えられる．補完された値は他のポートの観測値より算出されていることを考えれば，この誤差の大きさがフィルタリングによって遮断されているパケットの数と我々は考える．

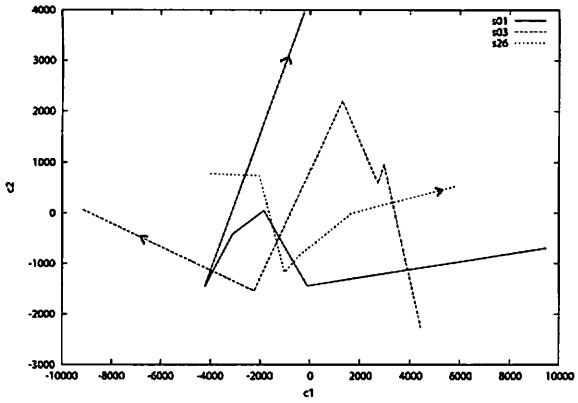

図 3　3 つのセンサの c_1 と c_2 上の推移
2005 年 10 月から 1ヶ月毎

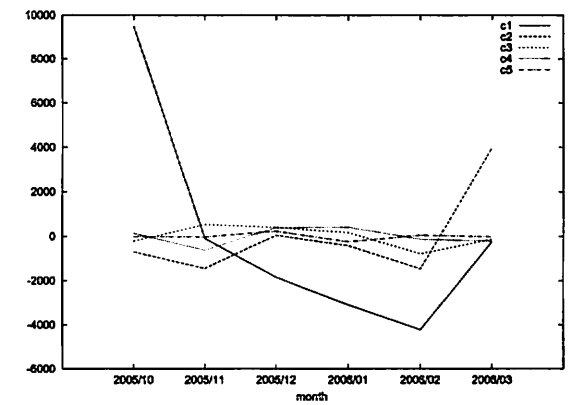

図 4　センサ s_1 の月毎の直交成分 c の推移

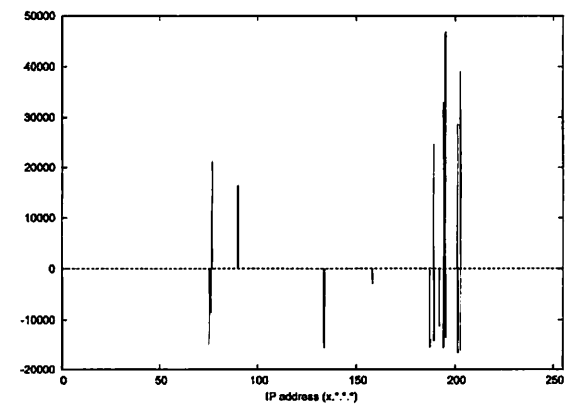

図 5　センサと c1 の関係

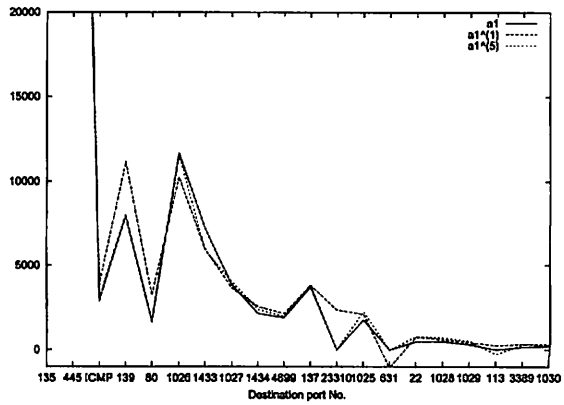

図 6　s_{01} の原観測ベクトルと第 5 近似

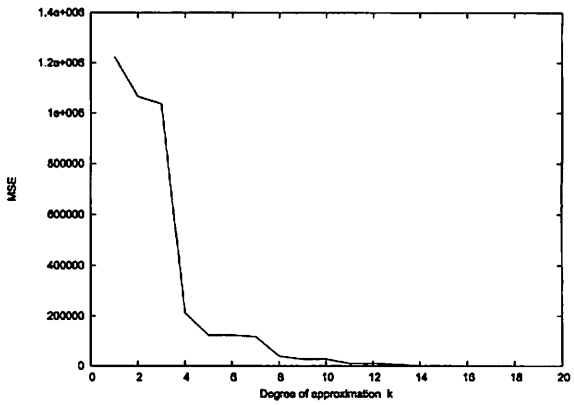

図 7　原データとの平均二乗誤差

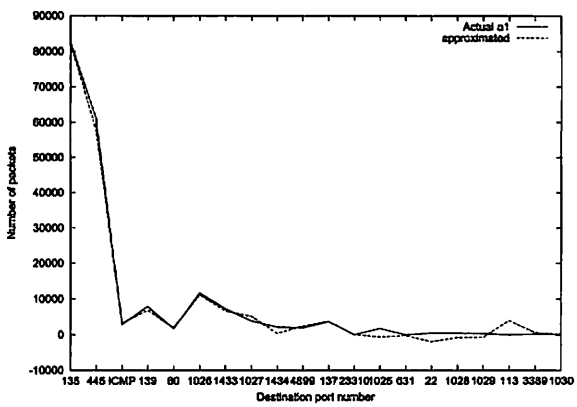

図 8　センサ s_{01} の観測ベクトルと第 10 補完

4. おわりに

本論文では全センサの直交基底を求め，各々の基底に影響のあるポート番号を示した．また，直交基底 g_i と係数 c_i を用いることにより，原観測データを近似できることを示した．さらに，欠損したパケットデータを，少量の誤差によって補完できることを証明した．これらのことより，本研究で示したセンサに対する直交展開の有用性，提案手法の有効性が示せたと考える．

本研究で提案した式 (3) は，欠損したポートが 1 つである場合にのみ有効である．しかし，実際のネットワーク環境では遮

表 4　センサ s_{01} 観測ベクトル a_1 の近似ベクトル

p_i	a	$a^{(1)}$	$a^{(5)}$	$a^{(10)}$	$a^{(20)}$
135	82411	82496	82391	82455	82412
445	61126	60862	61229	61060	61127
ICMP	2921	3888	3050	2994	2922
139	7946	11159	8026	7960	7947
80	1695	3217	1632	1584	1696
1026	11698	10279	11569	11667	11699
1433	7229	5986	5906	7186	7230
1027	3851	3655	4042	3913	3852
1434	2161	2568	2394	2262	2162
4899	1906	2151	1989	2080	1907
137	3714	3822	3807	3440	3715
23310	0	2356	-9	13	1
1025	1759	2105	2260	2186	1760
631	0	-1063	-42	-133	1
22	485	752	777	764	486
1028	476	608	718	663	477
1029	297	430	504	455	298
113	0	244	-275	-125	1
3389	161	321	308	368	162
1030	198	262	335	285	199
MSE		1224006	1066416	27554	1

表 5　p_2(=445) 補完の誤差の統計量

センサ数 m	28
平均 μ	15326
誤差平均 $\mu(x-a_{i2})$	0
標準偏差 $\sigma(x-a_{i2})$	3302
$\mathrm{Max}(x-a_{i2})$	10083
誤差率 $2\sigma/\mu(a_{12})$	0.43

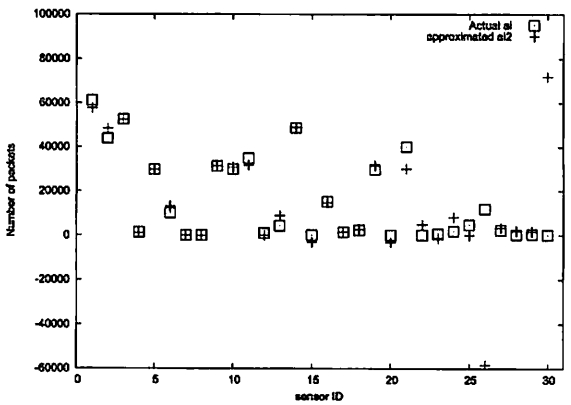

図 9　各センサの欠損ポートとの近似誤差

断されているポートは 1 つとは限らない．また，今回の実験では 2 つのセンサの近似に失敗した．これらのことを踏まえ，今後はこの提案手法を改良し，複数のポートの欠損に対応できるようにする．

謝　辞

本研究を遂行するにあたり，定点観測データ提供し，議論いただいた JPCERT/CC，議論いただいた杉山 太一氏，仲小路 博史氏，鬼頭 哲郎氏，藤原 将志氏に感謝する．

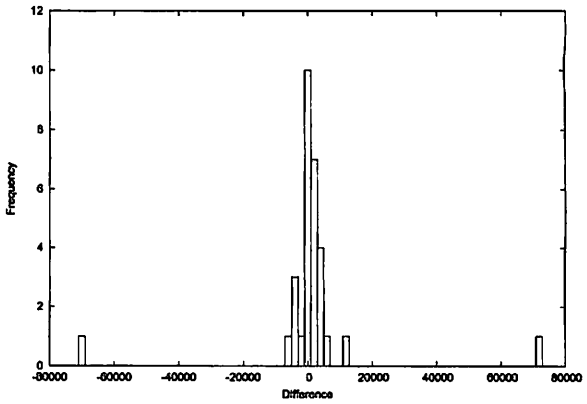

図 10　補完の誤差のヒストグラム

文　献

[1] 福野，菊池，寺田，土居，不正アクセスのトラフィックによるセンサの独立性，CSEC36，pp.95-102，2007.

[2] 戸田，他，ISDAS: Internet Scan Data Acquisition System 情報処理学会，コンピュータセキュリティシンポジウム CSS2004，pp.199-204，2004.

[3] 石黒，伊藤，戸田，鈴木，赤井，村瀬，インターネット上のポート観測による不正パケットの分布に関する特徴分析，コンピュータセキュリティシンポジウム (CSS 2005)，情報処理学会，6A-3，2005

[4] H.Kikuchi and M.Terada, How many scaners are in the Internet, The 7th International Workshop on Information Security Appication(WISA), Springer LNCS, 2006(to appear)

[5] 形態素解析と検索 API と TF-IDF でキーワード抽出 http://chalow.net/2005-10-12-1.html (2007 年 6 月参照)

[6] TCP/UDP Port List Well-known ports http://lists.thedatalist.com/portlist/portlist.htm (2007 年 6 月参照)

[7] Computer Virus Timeline http://www.infoplease.com/ipa/A0872842.html (2007 年 6 月参照)

[8] トロイの木馬と使用するポート http://h-ishida.hp.infoseek.co.jp/troi-house/troi.html (2007 年 6 月参照)