サンプルパケット情報を用いたトラヒック測定分析手法

川原　亮一，森　　達哉，滝根　哲哉，浅野正一郎

インターネット上においてネットワークリソースの浪費や品質劣化を引き起こす異常トラヒックをトラヒック測定を通じて検知・制御する技術は，安心で快適な通信サービスを提供するために不可欠となっている．一方，ネットワークの大規模化・高速化に伴い，パケットサンプリングによる測定が注目されている．本稿では，サンプルパケット情報から異常トラヒックを検出するためのトラヒック測定分析手法について，関連研究動向の紹介を交えながら筆者らの研究内容について紹介する．また，各手法の実データ評価結果も示す．

キーワード：パケットサンプリング，トラヒック，測定分析

1．はじめに

インターネットトラヒックの増加とインターネットの利用形態・アプリケーションの多様化に伴い，ネットワークの効率的な運用を支えるトラヒック測定が重要となっている．特に，品質劣化の要因となるネットワークリソースの浪費や，セキュリティ上の問題を引き起こす異常トラヒック（DDoS，ワーム等）を検出できる仕組みへの重要性が増している．

この狙いのため，フローレベルのトラヒック測定が近年注目されており，異常トラヒック検出，ヘビーユーザ特定，トラヒックエンジニアリング等への応用が検討されている．ここでフローとは，発信元IPアドレス（srcIP），着信先IPアドレス（dstIP），発信元ポート番号（srcPort），着信先ポート番号（dstPort），プロトコル（protocol）の5つ組を同じくするパケット群のことを指す．

すべてのフロー統計情報を取得するためには，監視対象ネットワークですべてのパケットをキャプチャして解析する必要があるが，ネットワークの大規模化・高速化によりスケーラビリティの問題が生じるため，パケットサンプリングに基づくフロー測定法が注目されている．図1がサンプリングのイメージである．本図上段のように全パケットをキャプチャすればすべてのフロー情報（どのフローが何パケット送出しているか）を正確に把握できるが，下段のようにサンプリングをすると，処理すべきパケット数やフロー数は削減できるが，その一方で必要な情報が失われる可能性がある．そこで，サンプリングの影響を考慮して異常トラヒックを適切に検出できる仕組みが必要となる．

本稿では，ネットワーク上で収集されたサンプルフロー情報（どのフローから何パケットサンプルされたか）を用いて異常トラヒックを検出する方法について，関連研究動向の紹介を交えながら筆者らの研究内容について解説する．

以下，2節でサンプリングが異常検出精度にどのような影響を及ぼすか分析し，3節で，異常検出精度向上のためのトラヒック分析手法について述べる．4節で，2節ならびに3節の分析で必要となるサンプルフロー数の平均分散特性について述べる．5節で，フロー長（フロー当りパケット数）分布推定に基づく異常

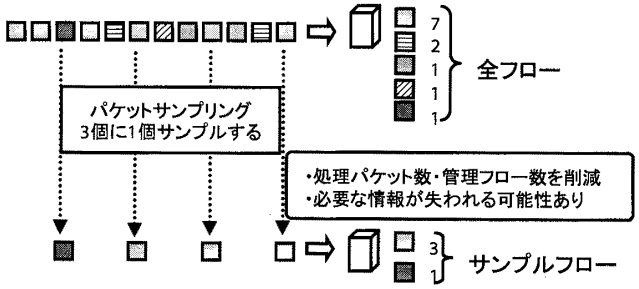
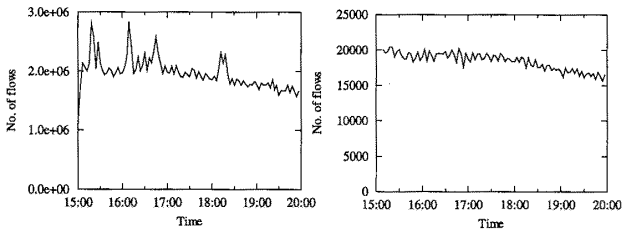


図1　サンプリング測定

かわはら　りょういち，もり　たつや
NTTサービスインテグレーション基盤研究所
〒180-8585 武蔵野市緑町3-9-11
たきね　てつや
大阪大学 大学院工学研究科
〒565-0871 吹田市山田丘2-1
あさの　しょういちろう
国立情報学研究所
〒101-8430 千代田区一ツ橋2-1-2



(a) $p=1$　　(b) $p=10^{-3}$

図2　フロー数の時系列[2]

検出法について紹介する．

2. サンプリングの異常検出精度への影響評価

2.1 実データ分析

サンプリングが異常検出にどのような影響を及ぼすか，まず実データによる評価結果を交えて説明する．ここでは，インターネットバックボーン回線上においてキャプチャしたパケットトレースデータを用いて，フロー数の時系列を分析した．その結果を図2に示す．図2(a)が元のフロー数，図2(b)がパケットサンプリングレート p を 10^{-3} に設定したときのサンプルフロー数の時系列を示す．ここで，サンプルフロー数とは，少なくとも1パケットサンプルされたフローの本数である．これより，元のフロー数ではいくつかのスパイクを目視できるが，サンプルフロー数ではそれらが目視できなくなっている．パケットトレースデータを詳細に分析したところ，15：00-17：00のスパイクは，あるsrcIPから，いくつかのdstIPに向けて，ポート番号を変えながらUDPパケットを送出し続けるUDP floodingであることが分かった．また，18：10-18：20におけるスパイクは，あるsrcIPがあるTCPポート番号めがけてdstIPを変えながらパケットを送出するネットワークスキャンであった．このように，小さなフロー（少ないパケットで構成されるフロー）を大量生成するような異常トラヒックを対象とした場合，元のフロー数では検知可能であってもサンプルフロー数では検知できなくなる場合がある（詳細な分析は2.2節）．

なお本稿では，本節でのデータ分析のように，サンプルフロー情報からトラヒック量（e.g.，フロー数）に関する時系列データを作成してトラヒック量の急激な変化の有無をチェックし，変化を検出したらその要因となる異常トラヒックを特定する，というトラヒック分析の流れ[1]を前提とする．

2.2 異常検出精度影響評価

サンプリングの異常検出精度への影響を以下のモデルで評価した（詳細は文献[2]を参照）．各フローからのパケットは確率 p でランダムにサンプリングされるとし，一定周期ごとに測定区間 t でのサンプルフロー数 $N_t(p)$ を測定し，$N_t(p)$ がある閾値（後述）を超えたら異常トラヒックが発生したと判定する．なお，$N_t(p)$ は，正常時は $N_t(p)=Nn_t(p)$，異常発生時は $N_t(p)=Nn_t(p)+Na_t(p)$ で与えられるとする．ここで，$Nn_t(p)$ はサンプルされた正常フロー数，$Na_t(p)$ はサンプルされた異常フロー数とする．また，$Nn_t(p)$ は，平均 $m_n(p)$，分散 $\sigma_n(p)^2$ の正規分布に従うとする．ここで，$m_n(p)$ は元のフロー数 $m_n(1)$ を用いて，

$$m_n(p)=\sum_x\{1-(1-p)^x\}f(x)m_n(1) \quad (1)$$

で計算した．ここで，$f(x)$ は正常フローのフロー当りパケット数が x である確率を表す（実データによる経験分布を使用）．一方，$\sigma_n(p)^2$ は，

$$\sigma_n(p)^2=\phi\times m_n(p)^c \quad (2)$$

で近似した（本近似式については4節で述べる）．

一方，異常発生時に加わるフロー数を d（固定値），異常フロー当りパケット数はすべて1とし，サンプル異常フロー数 $Na_t(p)$ を，平均 dp，分散 $dp(1-p)$ の二項分布でモデル化する（以下，正規分布で近似）．なお，ここで異常フロー当りのパケット数を1とした理由は，ネットワークスキャンやSYN floodingのような小さなフローを大量生成する異常トラヒックを対象としているためである．以上のモデルを用いて，異常検出閾値を $T_{th}(p)=m_n(p)+\alpha\sigma_n(p)$（$\alpha$ はあらかじめ定めるパラメータで本節では $\alpha=3$ に設定）としたときの異常を見逃す割合であるFNR（false negative ratio）を以下の式で評価する（図3参照）．

$$FNR(p)=\Pr[Nn_t(p)+Na_t(p)<T_{th}(p)] \quad (3)$$

一方，誤って異常を検出してしまう割合FPR（false positive ratio）は，

$$FPR(p)=\Pr[Nn_t(p)>T_{th}(p)] \quad (4)$$

で与えられ，$Nn_t(p)$ が正規分布に従う場合にはサンプリングレート p によらず $FPR(p)$ は一定となり，$\alpha=3$ のとき $FPR(p)=0.13\%$ となる．

公開トラヒックデータ[3]（1分周期でのパケットトレースデータ3時間分）を用いて，サンプリングレートを変えたときのFNRを評価する．$m_n(1)=287,122$[flows/min]（実データでの平均フロー数），$d=6\sigma_n(1)=51,701$ と設定したときのFNRを図4に示す（図中“psamp”）．なお，$d=6\sigma_n(1)$ の意味は以

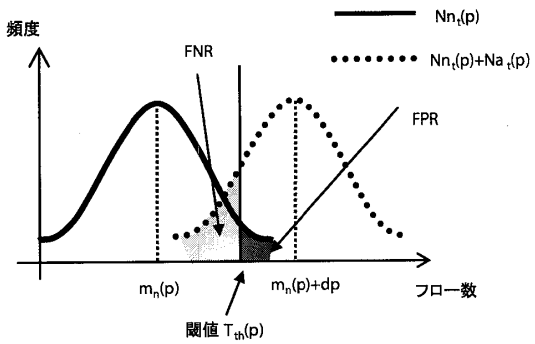


図3　FNR と FPR

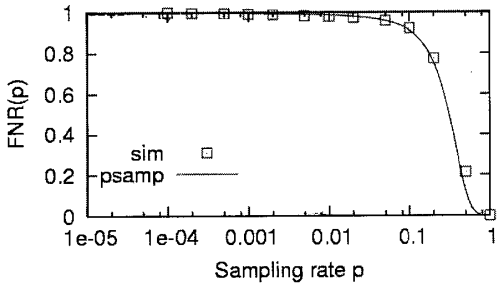


図4　FNR vs. サンプリングレート[2]

下の通りである．$d=6\sigma_n(1)$，$p=1$，$\alpha=3$ を式(3)に代入すると，$FNR=0.13\%$となる．つまり，サンプル前には$1-\epsilon$（$\epsilon=0.13\%$）の確率で検知可能な異常であることを意味する．図4より，pが小さくなるとFNRが著しく劣化していることが分かる．これは，パケットサンプリングを行うと，今回評価対象としているような1パケットのみから成る異常フローは複数パケットからなる正常フローに比べてサンプルされにくいためである．つまり，サンプルフロー数全体の平均(i.e., $m_n(p)+dp$)に占める異常サンプルフロー数の平均(i.e., dp)の割合が減少し，異常フローによるトラヒックの増分の検知が困難になるためである．

参考として，実データのサンプルフロー数時系列の任意時点に異常トラヒックの付加を模擬した場合の結果も示す（“sim”）．これより，解析結果がシミュレーション結果の傾向を捉えていることが確認できる．

なお，ここではパケット数の少ないフローを大量生成する異常トラヒックを対象に評価しているのに対し，文献[4]においては，パケット数の多い異常トラヒック検出においてもパケットサンプリングが影響を及ぼすことを示している．具体的には，サンプリングに起因する分散の増分がFNRを劣化させることを指摘している．また関連研究として，文献[5][6]においても，サンプリングの異常検出精度への影響をケーススタディ等を通じて評価している．

3.　トラヒック分割監視法

前節での問題を回避するために，トラヒックを複数のグループに分割監視する方法を考える．もし正常トラヒックを全グループに分配しつつ異常トラヒックをあるグループに集約できれば，そのグループでの異常フロー数の占める割合が増加し，異常検出確率を向上できると期待される．例えばsrcIPをキーとしてトラヒックを分割することにより，ある発信元ホストから多数の着信先ホストや着信先ポート番号にパケットを送出するスキャンのような異常トラヒックをある特定のグループに集約できることが期待される．一方，ある着信先ホストへトラヒックが集中するDDoS攻撃に対しては，dstIPをキーとして分割することが効果的であると考えられる．

ここで，適切な分割数に関する評価結果を示す．図4の評価で用いたものと同じデータを対象に，$p=10^{-3}$，$d=51,701$に設定し，M_g個のグループに分割後のFNRの値を評価した．ここでは，ランダムパケットサンプリングにより得られるサンプルフロー数時系列を生成し，そのサンプルフロー数時系列をM_g分割（$M_g=1, 2, 4, 8, 16, 32, 64, 100$）して，各グループの任意の時点に異常トラヒックを付加するシミュレーションを実施し，各グループごとのFNRを計算した．各グループのFNRと平均フロー数をプロットした結果を図5の“sim”に示す．また参考として，前節と同様に正規分布を仮定してFNRとグループ内平均フロー数の関係を計算した結果についても図中“model”に示す．ここでは，まず，トラヒックをM_g分割したときのグループj_{M_g}の正常サンプルフロー数の平均を$m_n(p, j_{M_g}, M_g)$，分散を$\sigma_n(p, j_{M_g}, M_g)^2$とし，式(2)と同様に分散と平均の関係を以下の式で近似した（詳細は4節で述べる）．

$$\sigma_n(p, j_{M_g}, M_g)^2=\tilde{\phi}m_n(p, j_{M_g}, M_g)^{\tilde{c}} \tag{5}$$

次に，平均$m_n(p, j_{M_g}, M_g)$（$=\tilde{m}$とおく）が与えられたときのFNRを以下で計算した．

$$FNR=F(\tilde{m}+\alpha\sqrt{\tilde{\phi}\tilde{m}^{\tilde{c}}}, \tilde{m}+dp, \tilde{\phi}\tilde{m}^{\tilde{c}}+dp(1-p)) \tag{6}$$

ここで，$F(x, m, \sigma^2)$は平均m，分散σ^2の正規分布に従う確率変数Xがx以下となる確率を表す．

これより，サンプルフロー数が40 [flow/min]程度になるように分割すればFNRが0.13%以下に抑えられることが確認できる．

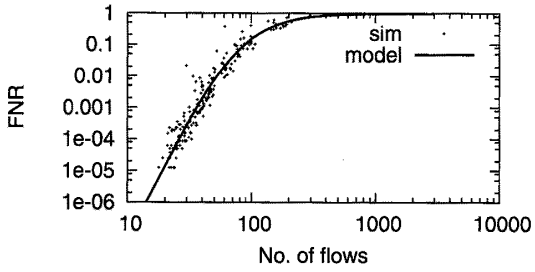


図5　FNR vs. グループ内フロー数[2]

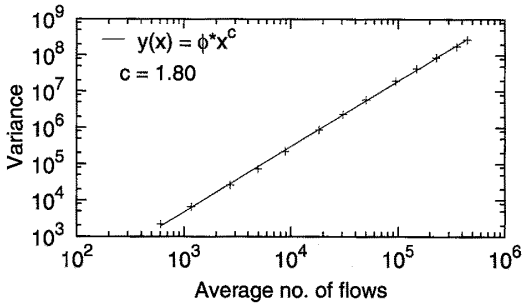


図6　サンプルフローの分散 vs. 平均

4.　平均分散特性

本節では，サンプルフロー数の平均と分散の関係について評価した結果を述べる．まず，2.2節と同じデータを用いて，サンプリングレート p を1から 10^{-4} の範囲で変えたときのフロー数の平均と分散をプロットした結果を図6に示す．これより，平均分散の関係を式(2)で近似できることが期待される．

上記関係式について以下で考察していく．サンプリングレートが p のとき，フロー長 X が k パケットであるフローがサンプルされる確率は $1-(1-p)^k$ となる．よって，フロー数 $F=n$ が与えられたとき，あるフローがサンプルされる確率 $q(n)$ は，

$$q(n)=\sum_{k=1}^{\infty}\Pr[X=k|F=n][1-(1-p)^k] \tag{7}$$

となる[1]．さらにサンプル後のフロー数の平均と分散は以下で与えられる．

$$m_n(p)=\sum_{n=1}^{\infty}\Pr[F=n]nq(n) \tag{8}$$

$$\sigma_n(p)^2=\sum_{n=1}^{\infty}\Pr[F=n]n(n-1)q(n)^2 + m_n(p)(1-m_n(p)) \tag{9}$$

2.2節と同じデータから得られる $\Pr[X=k|F=n]$ を元に，この式に従って平均，分散を計算し，両者の関係を式(2)で近似したところ，$(\phi, c)=(0.01715, 1.809)$ となり，図6のときの結果 $(\phi, c)=(0.01856, 1.802)$ と非常に近い値が得られた．一方，フロー数 F とは独立に同一の確率 q で各フローがサンプルされると仮定して計算すると，$(\phi, c)=(0.003914, 1.911)$ となり分散は小さめに見積もられる．ここでの相関構造の解明・モデル化等については現在検討中である．

[1] 厳密には $q(n)$ は本節での議論のようにフロー長分布を通じてフロー数に依存した形となるが，式(1)では，フローがサンプルされる確率は簡単のためフロー数とは独立として計算しており，今後詳細な分析をしていく．

また式(5)についても，上記と同様の分析を実施しており（詳細は省略），また他のネットワークでも同様の傾向が報告されている[7]．

5.　分布情報を用いた異常検出法

2節のように単にフロー数やパケット数の時系列を観測する代わりに，Lakhinaら[8]は，異常トラヒックによる通信パターンの変化を特徴づける方法について検討している．具体的には以下で定義されるエントロピーを観測することにより異常検知を行う．エントロピーは，ある特徴量 i の出現回数 $n_i(i=1, \cdots, N)$ に対するヒストグラム $X=\{n_1, n_2, \cdots, n_N\}$ を用いて

$$H(X)=-\sum_{i=1}^{N}\left(\frac{n_i}{S}\right)\log_2\left(\frac{n_i}{S}\right) \tag{10}$$

で定義される．ここで，$S=\sum_{i=1}^{N}n_i$ である．特徴量として例えばdstIP i の出現回数（発生パケット数）などが考えられる．エントロピーは，ある特徴量 i にトラヒックが集中している場合には0に近づき，分散している場合はその最大値である $\log_2 N$ に近づくため，$H(X)$ の時間変化を観測することにより異常検知が可能になる．また，文献[9]では，上記手法のオンライン化についても論じている．

5.1　フロー長分布推定に基づく異常検知法

筆者らは，文献[10]において，サンプリングがこのエントロピーを用いた手法に対して影響を及ぼすことを実データを用いて示している．またその解決方法として，サンプルフロー情報から元のフロー長分布を推定し，その推定された分布を用いてエントロピーを監視することにより，サンプル前に検知できた異常を検知可能であることを示している．図7に評価結果を示す．ここでは文献[11]のデータを用い，5分周期ごとにエントロピー $s=-\sum_i p(i)\log\{p(i)\}$ をプロットした．ここで $p(i)$ はフローが i パケット持つ割合であり，サンプル前（図中“original”），サンプリングレートを1/100，1/1,000にした場合，それぞれについて評

価した．本左図は元のフロー長分布 $p(i)$ を推定した場合，本右図はサンプルフロー長分布をそのまま用いた場合の結果である．なお分布推定法は5.2節で述べる．これより，サンプリングレートが小さくなると，サンプルフロー情報をそのまま用いた場合には，サンプル前に顕著であった時刻21：20におけるエントロピーの急減が目視しづらくなっていることが分かる．一方，元のフロー長分布を推定した場合には，サンプル後もエントロピーの変化を適切に把握できていることが分かる．なお，21：20のエントロピー減少の原因を調べたところ，いくつかのdstIPに対してspoofされたたくさんのsrcIPからSYN floodingが行われていた．

5.2 フロー長分布推定法

フロー長分布は多くの場合においてパレート分布などのベキ型の分布に従うことが知られている．そこでフロー長分布を離散パレート分布によってモデル化し，サンプルデータを用いて元の分布をパラメトリックに推定する．具体的には，あるフローのパケット数 X に関して，$\Pr[X=k]$ を以下でモデル化する．

$$\Pr[X=k]=p(k\,;\,\theta)=k^{-\theta}-(k+1)^{-\theta} \tag{11}$$

一方，サンプリングレート p でパケットをランダムにサンプルした場合，あるフローの元のパケット数 X が $X=k$ である条件の下でそのフローからのサンプルパケット数 Y が $Y=i$ である確率は二項分布 $q(i|k)=\binom{k}{i}p^i(1-p)^{k-i}$ で与えられる．したがって，あるフローからのサンプルパケット数が $Y=i$ となる確率は $r(i\,;\,\theta)=\sum_{k=i}^{N}q(i|k)p(k\,;\,\theta)$ となる．なお，N は母集団におけるパケット総数である．

ここで，観測されたサンプルフロー情報から θ を推定することを考える．今，$n_i(i=0,1,\cdots,y_{\max})$ を，サンプルパケット数が i であるフローの数，とする．ここで，$y_{\max}$ はフロー当りサンプルパケット数の最大値を指す．未観測データ n_0 を含めた完全対数尤度は，

$$\log L_c(\theta)=\log\prod_{i=0}^{y_{\max}}r(i\,;\,\theta)^{n_i}=\sum_{i=0}^{y_{\max}}n_i\log r(i\,;\,\theta) \tag{12}$$

となる．これを最大化するような θ を求め，推定値とすればよい．ここで2つの方法が考えられる．一つは，EMアルゴリズムにより推定する方法[10]，もう一つは，未観測データ n_0 あるいはその推定値を何らかの方法で取得できたとして式(12)に対して最尤推定を行う方法である．図7の左図は，後者の方法によって推定された分布を用いてエントロピーを計算した結果である．なお，文献[13]で，元のフロー数を推定する方法を述べており，それを元に n_0 を推定する，といったアプローチも考えられる．また，文献[14]ではサンプルされたTCP-SYNパケット数を利用して，フロー長分布を推定している．

他の手法として，文献[12]では有限測定時間内では分布の裾が切断されていることを考慮して

$$\Pr[X=k]=p(k\,;\,\theta)=\frac{k^{-\theta}-(k+1)^{-\theta}}{1-(\nu+1)^{-\theta}} \tag{13}$$

でモデル化している．ここで ν は X の上限である．このモデルに対して，観測値を用いてまず ν を $y_{\max}/p$ で推定する．次に，$\boldsymbol{n}=(n_1,n_2,\cdots,n_{y_{\max}})$ が観測されたという条件の下での尤度

$$L(\theta\,;\,\boldsymbol{n})=\prod_{i=1}^{y_{\max}}\left(\frac{r(i\,;\,\theta)}{1-r(0\,;\,\theta)}\right)^{n_i} \tag{14}$$

を最大にする θ を求める．ここで，この尤度関数は，

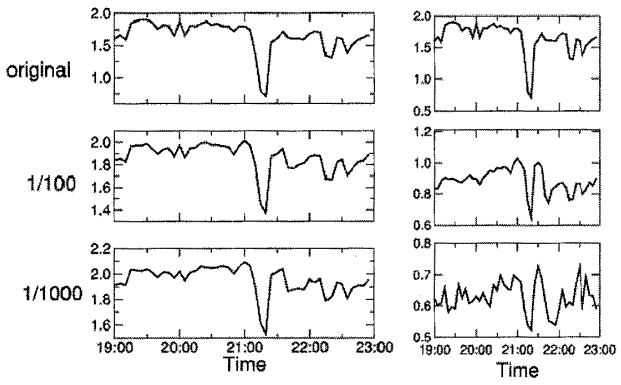


図7 エントロピーの時間変化（左）元のフロー長分布推定時，（右）サンプルフロー長分布使用時[10]

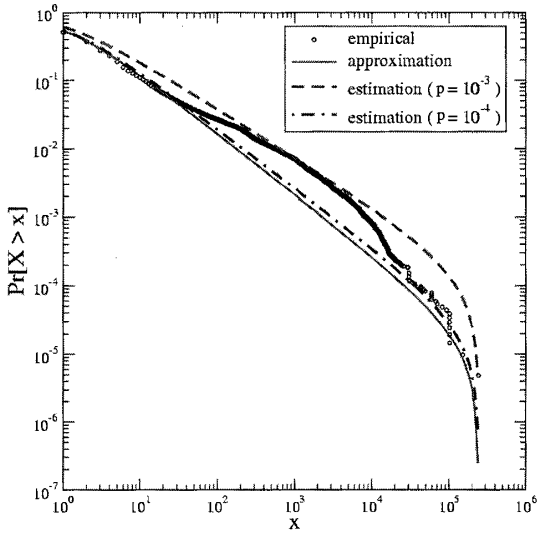


図8 フロー長分布推定結果[12]

y 個 ($y=1, 2, \cdots, y_{max}$) のパケットがサンプルされたフロー数が n_y である確率を意味する．評価結果を図8に示す．ここで“approximation”は，サンプル前の情報を用いて式(13)に fit させた結果である．

6. まとめ

本稿では，サンプルパケット情報を用いて異常トラヒックを検出するためのトラヒック測定分析手法について紹介した．トラヒック変化を自動的に検出するための時系列解析手法や，変化を引き起こした異常トラヒックの特定手法の研究も国内外で進められており，また，単純なパケットサンプリングではなく例えばスケッチと呼ばれるデータ集約手法を応用した検討[15][16]等）もなされており，今後も多面的に研究を進めていきたい．

謝辞 本稿作成にあたり，貴重なコメントを頂いたNTT研究所石橋圭介氏に謝意を表します．本稿で紹介した研究の一部は，総務省委託研究課題「次世代バックボーンに関する研究開発」の成果である．

参考文献

[1] 川原, 森, 原田, 上山, 近藤, 石橋, “異常トラヒック測定分析手法,” NTT技術ジャーナル, Vol. 20, No. 3, pp. 21-25, 2008年3月.

[2] R. Kawahara, K. Ishibashi, T. Mori, N. Kamiyama, S. Harada and S. Asano, “Detection accuracy of network anomalies using sampled flow statistics,” IEEE Globecom 2007, Nov. 2007.

[3] http://pma.nlanr.net/Special/cesc 1.html

[4] K. Ishibashi, R. Kawahara, T. Mori, T. Kondoh and S. Asano, “Effect of sampling rate and monitoring granularity on anomaly detectability,” 10 th IEEE Global Internet Symposium 2007, May 2007.

[5] D. Brauckhoff, B. Tellenbach, A. Wagner, M. May and A. Lakhina, “Impact of packet sampling on anomaly detection metrics,” ACM SIGCOMM IMC, 2006.

[6] J. Mai, C. -N. Chuah, A. Sridharan, T. Ye and H. Zang, “Is sampled data sufficient for anomaly detection?,” ACM SIGCOMM IMC, 2006.

[7] A. Gunnar, M. Johansson and T. Telkamp, “Traffic matrix estimation on a large IP backbone—A comparison on real data,” ACM SIGCOMM IMC, Oct. 2004.

[8] A. Lakhina, M. Crovella and C. Diot. “Mining Anomalies Using Traffic Feature Distributions,” Proc. ACM SIGCOMM 2005, September 2005.

[9] T. Ahmed et al., “Multivariate Online Anomaly Detection Using Kernel Recursive Least Squares,” Infocom 2007, 2007.

[10] T. Mori et al., “Inferring original traffic pattern from sampled flow statistics,” IEEE SAINT 2007 Workshop, Jan. 2007.

[11] MAWI Working Group Traffic Archive, http://tracer.csl.sony.co.jp/mawi/

[12] T. Mori, T. Takine, J. Pan, R. Kawahara. M. Uchida and S. Goto, “Identifying Heavy-Hitter Flows From Sampled Flow Statistics,” IEICE Trans. on Commun., Voll. E 90-B, No. 11, pp. 3061-3072, November 2007.

[13] N. Duffield, C. Lund and M. Thorup, “Properties and Prediction of Flow Statistics from Sampled Packet Streams,” ACM SIGCOMM IMW, 2002.

[14] N. Duffield, C. Lund and M. Thorup, “Estimating Flow Distributions from Sampled Flow Statistics,” ACM SIGCOMM, pp. 325-336, 2003.

[15] B. Krishnamurthy et al., “Sketch-based change detection: methods, evaluation, and applications,” ACM IMC 03.

[16] H. Zhao et al., “A Data Streaming Algorithm for Estimating Entropies of OD Flows,” in IMC 2007, Oct. 2007.