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

2007－ＤＳＭ－45（13）
2007／5／11

異なる間隔を用いたパケットサンプリングにおける差分情報のモデル化

磯崎　裕臣†　　阿多　信吾†　　岡　　育生†

† 大阪市立大学 大学院工学研究科 〒 558-8585 大阪市住吉区杉本 3-3-138
E-mail: †isozaki@n.info.eng.osaka-cu.ac.jp, ††{ata,oka}@info.eng.osaka-cu.ac.jp

あらまし　高速回線においてすべてのトラヒックを収集し、その統計情報を得ることは、必要とされる資源量や収集オーバーヘッドなどの問題により実現が難しい。このため、確率的に選択されたパケットのみで統計情報を収集する、パケットサンプリングが有効な手法であると考えられている。パケットサンプリングを利用することで、必要となる資源量を削減でき、効率良くフローの収集を行うことが可能である。しかし、サンプリングを行った結果、情報の欠落が発生することで統計情報の精度が劣化するという問題が発生する。本稿では、異なる間隔のパケットサンプリングで得られた統計情報の差分を分析することで、パケットサンプリングによるフローのパケット数の変化をモデル化する。導出したモデルに基づいたオリジナルのフロー統計情報をより正確に推定する手法を新たに提案する。さらに、インターネット上で公開されている実トレースデータに提案手法を適用させることで、提案手法が高い精度でオリジナルのフロー統計情報を推定できることを示す。
キーワード　パケットサンプリング、サブサンプリング、フロー統計、差分情報、フロー推定、モデル化

Modeling of Difference Information on Packet Sampling by Different Intervals

Hiroomi ISOZAKI†, Shingo ATA†, and Ikuo OKA†

† Graduate School of Engineering, Osaka City University
3-3-138, Sugimoto, Sumiyoshi-ku, Osaka 558-8585, Japan
E-mail: †isozaki@n.info.eng.osaka-cu.ac.jp, ††{ata,oka}@info.eng.osaka-cu.ac.jp

Abstract　Packet sampling is one of effective ways to reduce the overhead of the traffic measurement in high speed routers, in which the router collects only a part of whole traffic based on the probabilistic theory. Since the statistic information of uncollected packets may be lost, the accuracy of the statistic information of sampled traffic becomes degraded as the increase of the sampling interval. In this paper, we analyze the difference information from statistics obtained by multiple sub-samplings with different sampling interval in order to model the change in the number of packets of flow by packet sampling. We then propose a new method which improves the accuracy of the estimation of flow length distribution based on derived model. We then apply our method to show that our method can estimate original flow length distributions accurately.
Key words　Packet Sampling, Sub Sampling, Flow Statistics, Difference Information, Flow estimation, Modeling

1. はじめに

ネットワークの性質を特徴づけるためには、ネットワーク上でトラヒックを計測し、それらをフロー単位で分析することが有効である。しかしながら、回線がより広帯域となるにしたがい、特にコアネットワークにおけるルータですべてのパケットをモニタリングし、フローごとの統計量を計測することは、必要となる資源量やルータへの負荷を考えるとスケーラビリティの点で問題がある。このため、ルータ上を経由するトラヒックのうち一部のトラヒックの情報のみをモニタリングし、その情報に基づいてフローごとの統計情報を取得するパケットサンプリング [1,2] と呼ばれる手法が提案されている。フローの統計情報とは、宛先 IP アドレス、送信元 IP アドレス、宛先ポート番号、送信元ポート番号、プロトコル番号が等しいパケットの集合であるフローについてパケット数、バイト数、継続時間などの情報を収集したものである。

パケットサンプリングでは、パケット数が多いフローの一部のパケットやパケット数がごく短いフローの全パケットが抽出されないことにより、パケットサンプリングで得られたトラヒックを単純にサンプリング間隔でスケーリングを行うだけで

は、実際のトラヒックとの間に誤差が生じ、正確な統計情報を得ることができない [3]。

この問題を解決するため、パケットサンプリングにより得られたトラヒックの統計情報より、トラヒック全体の統計情報（本稿ではこれを「真の統計情報」と呼ぶ）をより精度よく推定するための手法が検討されている。[4] では、サンプリングにより得られたフローの総数やバイト数にサンプリング間隔を掛けることで、真の統計情報のうちフローの総数と総バイト数を推定する手法が提案されている。また、SYN パケットに注目することで、SYN パケットを含んでいるフローの数にサンプリング間隔を掛けることで SYN パケットが含まれるフローの総数を推定する手法も提案されている。しかしながら、これらの手法では、フローの分布を推定することができない。このため、より確からしい真の統計情報を推定する手法として、パデ近似を用いた手法 [5] や EM (Expectation Maximization) アルゴリズム [6] による最尤推定手法を用いた手法 [7,8] が提案されている。[7] では、推定誤差が特に TCP SYN パケットに依存していることに着目し、推定精度を向上させている。また [9] では、EM アルゴリズムによる推定において、ルータを経由するフローの総数を正しくカウントできれば、精度を大幅に向上できることを示している。しかしながら、TCP SYN フラグを用いた手法では TCP 以外のプロトコルについては精度向上ができない。また、フロー総数のカウントについても一度すべてのパケットのヘッダをチェックする必要や、ルータ自身にカウンタの機能を追加する必要があるなど、実現のための課題が存在する。これらの課題を解決する手法として、異なるサンプリング周期のフローの統計情報を比較し、その差分情報に着目することで、容易に真の統計情報を推定できる手法 [3] が提案されている。しかし、真の統計情報の推定に関数による近似を用いているために、関数による近似が正確に行えない場合、推定精度の誤差が大きくなるという問題が存在する。

そこで、本稿では、異なる間隔を用いたパケットサンプリングにおけるフローの統計情報を比較し、その差分情報を分析することで、パケットサンプリングによるフローのパケット数の変化をモデル化する。そして、導出したモデルに基づいたより簡便に真のフロー統計情報を精度よく推定できる手法を新たに提案する。そして、実際のトレースデータを分析することにより、提案手法が精度よくオリジナルフローの統計情報を推定できることを示す。

以下、2. で異なる間隔を用いたパケットサンプリングの統計情報の差分情報のモデル化について述べる。3. では差分情報のモデル化に基づいた推定手法を提案する。4. で実際のトレースデータに対して提案手法を適用し、オリジナルフローの統計情報を推定することにより、提案手法がより精度よく推定できることを示す。5. で提案手法における推定に必要な計算量を導出することで、提案手法が少ない計算量で真の統計情報を推定可能であることを示す。最後に 6. でまとめと今後の課題について述べる。

2. 差分情報のモデル化

本章では、異なる間隔のサンプリングで得られた統計情報間の差分情報のモデル化を行う。

本稿では、異なる間隔のパケットサンプリングの結果を得るために、サブサンプリングという手法を利用する。サブサンプリングとは、ある間隔のパケットサンプリングによって得られた統計情報を、再度サンプリングすることにより、異なるサンプリング間隔の統計情報を得る手法である。サブサンプリングを利用することで、異なる間隔のサンプリングを同時に行う必要がなくなり、ルータの計測に必要な負荷を軽減することができる。

以下、2.1 で、異なる間隔のサンプリングによるフロー統計情報間の差分情報について述べる。そして、2.2 で、差分情報のモデル化を行う。

2.1 異なる間隔サンプリングによる統計情報間の差分情報

本節では、異なる間隔のサンプリングで得られた統計情報間の差分情報について述べる。統計情報間の差分情報は、[3] で述べられており、サンプリング間隔が変化した場合に、統計情報がどのように変化するかを示したものである。

異なる間隔のサンプリングで得られた n 個の統計情報を考え、t をサブサンプリング係数と定義し、$0 \leqq t \leqq n$ を満たすとする。また、サンプリング間隔を $s^{(t)}$ とし、$1 \leqq t \leqq n$ で $s^{(t-1)} < s^{(t)}$ を満たすとする。このとき、パケット数が i であるフローの数を $f_i^{(t)}$、バイト数を $b_i^{(t)}$ とし、統計情報に含まれるフローうちもっともパケット数が多いフローのパケット数を $M^{(t)}$ とすると、フローの総数 $F^{(t)}$ は、

$$F^{(t)} = \sum_{i=1}^{M^{(t)}} f_i^{(t)} \tag{1}$$

となり、フローの平均パケット数 $B^{(t)}$ は、

$$B^{(t)} = \sum_{i=1}^{M^{(t)}} b_i^{(t)} / f_i^{(t)} \tag{2}$$

とおける。ここで、フローの総数およびフロー平均パケット数の変化率は

$$\Delta F^{(t)} \equiv \frac{F^{(t)}}{F^{(t-1)}} \tag{3}$$

$$\Delta B^{(t)} \equiv \frac{B^{(t)}}{B^{(t-1)}} \tag{4}$$

で与えられる。ただし、$1 \leqq t \leqq n$ である。

さらに、2 つの変化率の積はサンプリング間隔の変化率と一致し、

$$\Delta F^{(t)} \times \Delta B^{(t)} = \Delta s^{(t)} \tag{5}$$

で与えられる。ただし、

$$\Delta s^{(t)} \equiv \frac{s^{(t)}}{s^{(t-1)}} \tag{6}$$

である。

表 1　利用したトレースデータ

Dataset		Date	Number of packets	Number of flows
Abilene III	KSCY to IPLS	2004/06/01	1,909,039	173,549
	IPLS to KSCY	2004/06/01	4,575,186	415,926
Auckland II	Outbound	2000/01/28	7,886,659	716,969
	Inbound	2000/01/28	12,065,999	1,096,909

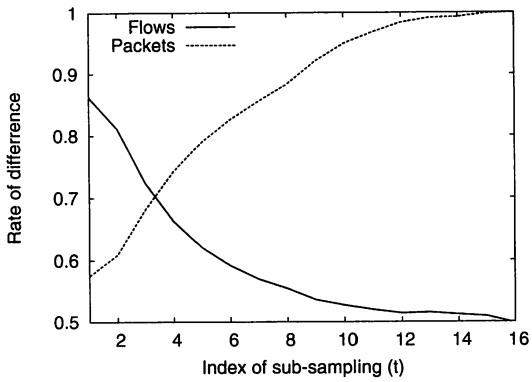

図 1　フローの総数と平均パケット数の変化率 (Auckland II Outbound)

図 4 に NLANR [10] で公開されているトレースデータをサブサンプリングすることによって得られた統計情報間の差分情報のグラフの例を示す。また、利用したトレースデータの名称、収集された日、トレースデータに含まれるパケットの総数、およびフロー総数を表 1 に示す。Abilene III は、Abilene ネットワークの OC192c バックボーンのうち Indianapolis と Kansas City 間のネットワーク、Auckland II は、オークランド大学と ISP 間の OC3c ATM ネットワークで収集されたものである。Abilene III のデータセットは、10 分間のトレースデータであり、Auckland II のデータセットは、24 時間のトレースデータである。ここでは、サンプリング間隔を 2 とし、得られた統計情報をさらにサブサンプリングすることを繰り返すことで異なる間隔のサンプリングによる統計情報を取得している。すなわち、$s^{(t)} = 2^t$ である。図の横軸はサブサンプリング係数、縦軸はフローの総数の変化率 ($\Delta F^{(t)}$) とフローの平均パケット数の変化率 ($\Delta B^{(t)}$) を示している。 図より、サンプリング間隔が長くなるにつれて、それぞれの変化率が変化することが分かる。つまり、サンプリング間隔が短い区間ではフローの総数の変化率が大きく減少するが、サンプリング間隔が長くなるにつれてその減少割合は小さくなり、フローの平均パケット数の変化率が大きく増加することが見られる。

2.2　差分情報のモデル化

本節では、異なる間隔のサンプリングにより得られた統計情報を分析することで、パケットサンプリングによる統計情報の変化のモデル化を行う。

ここで、サンプリング間隔を n とすると、サンプリング確率 q は、$q = 1/n$ となる。パケット数が i であるフローが、サンプリングによって、パケット数が j に変化する確率 $p_{i,j}$ は、独立試行において、確率 q で、i 個のうち、j 個選択される確率と考えることができ、

$$p_{i,j} = {}_iC_j q^j (1-q)^{i-j} \tag{7}$$

とおけ、$0 \leqq j \leqq i$ である。ここで、$j = 0$ の場合は、フローの全パケットがサンプリングされない確率、つまりフロー自体がサンプリングされない確率を示している。ここで、式 (7) より、サンプリングによって、フローのパケット数が j に変化したフローの数 $f_j^{(t+1)}$ は、サンプリングを行う前の各フローの数 $f_i^{(t)}$ より、

$$f_j^{(t+1)} = \sum_{i=j}^{M^{(t)}} p_{i,j} \times f_i^{(t)} = \sum_{i=j}^{M^{(t)}} {}_iC_j q^j (1-q)^{i-j} \times f_i^{(t)} \tag{8}$$

となり、$0 \leqq j \leqq M^{(t)}$ である。式 (8) より、サンプリングによるフローのパケット数の変化を数式でモデル化することができる。3. では、本節で導出したモデルに基づいた真の統計情報を推定する手法について述べる。

3.　モデル化に基づいた推定手法

本章では、異なる間隔のサンプリングによって得られた統計情報間の差分情報を分析することで得られたパケットサンプリングによるフローのパケット数の変化のモデルに基づいたオリジナルフローの統計情報を推定する新たな手法を提案する。

すでに述べたように、式 (8) より、パケットサンプリングによってフローのパケット数が j に変化したフローの分布を得ることができる。そこで、本稿では、この分布にしたがって、パケット数が j であるフローのパケット数を変化させることでサンプリングを行う前のパケット数ごとのフローの数を推定する。

ここでは、推定に利用する統計情報をサンプリング間隔 $s^{(t)}$ でネットワークから取得した場合を考え、サブサンプリング係数 t を、$t = \log_q s^{(t)}$ と定義する。そして、サンプリング間隔 $s^{(t)}$ の統計情報から サンプリング間隔 $s^{(t-1)}$ の統計情報の推定を行う。

まず、式 (8) において、

$$g_{i,j}^{(t)} = p_{i,j} \times f_i^{(t)} \tag{9}$$

とおくと、式 (8) は、

$$f_j^{(t+1)} = \sum_{i=j}^{M^{(t)}} g_{i,j}^{(t)} \tag{10}$$

となり、$0 \leqq j \leqq M^{(t)}$ である。ここで、パケットサンプリングによって、パケット数が j に変化したフローの総数に対するパケット数ごとのフローの数の割合を得るために、$g_{i,j}^{(t)}$ の確率密度関数 (PDF) $d_{i,j}^{(t)}$ を、

$$d_{i,j}^{(t)} = \frac{g_{i,j}^{(t)}}{\sum_{i=j}^{M^{(t)}} g_{i,j}^{(t)}} \tag{11}$$

と定義する。$d_{i,j}^{(t)}$ を パケット数が j であるフローの数 $f_j^{(t+1)}$ に掛けることで、サンプリングによって、パケット数が j になったフローの数からサンプリングを行う前にパケット数が i で

あったフローの数 $f_i^{(t)}$ を得ることができる。以上より、パケット数が j であるフローの数からパケットサンプリングによって変化する前のパケット数 i のフローの数を推定する式は、

$$f_i^{(t-1)} = d_{i,j}^{(t-1)} \times f_j^{(t)} \tag{12}$$

となり、$j \leqq i \leqq M^{(t-1)}, 0 \leqq j \leqq M^{(t)}$ である。

しかし、式 (12) を利用して推定を行うには、式 (11) に含まれる $f_i^{(t-1)}$ の値が必要となる。しかしながら、$f_i^{(t-1)}$ は、推定する値であるため、得ることができない。そこで、パケットサンプリングを行う前のパケット数を k とおき、k を 2.1 で述べたフローごとの平均パケット数の変化率 $\Delta B^{(t)}$ を用いて、サンプリングによってパケット数が j に変化したフローのパケット数をスケーリングしたものとし、

$$k = j/\Delta B^{(t)} \tag{13}$$

とする。ここで、$\Delta B^{(t)}$ は、サブサンプリングを行うことで得られるフローの総数 $F^{(t)}$ から、式 (5) を利用することで、

$$\Delta B^{(t)} = \frac{1}{\Delta F^{(t)}} \times \frac{1}{q} = \frac{F^{(t-1)}}{F^{(t)}} \times \frac{1}{q} \tag{14}$$

と導出することができる。よって、サンプリングを行う前にパケット数が k であったフローの数 $f_k^{(t-1)}$ は、

$$f_k^{(t-1)} = f_j^{(t)} \tag{15}$$

と近似できる。また、フローの最大パケット数 $M^{(t-1)}$ も同様にスケーリングすることで、

$$M^{(t-1)} = M^{(t)}/\Delta B^{(t)} \tag{16}$$

で近似できる。

しかし、この近似ではパケット数が $k = j/\Delta B^{(t)}$ であるフローの数しか得ることができず、これら以外のフローの情報が欠落してしまう。これまでの研究で、フローの分布はパレート分布に従うことが示されている [11]。よって、近似できない部分のフローの数をパレート分布 $e(x)$ を利用して補完する。$e(x)$ は、

$$e(x) = a\left(\frac{1}{x}\right)^b \tag{17}$$

で与えられる。また、式 (14) より、$\Delta B^{(t)}$ を得るには、$F^{(t-1)}$ が必要となる。サブサンプリングを行うことで $F^{(t+1)}$ は得ることができるが、$F^{(t-1)}$ を得ることができない。このため、サブサンプリングにより、$F^{(t+1)}$ を取得することで計算できる $\Delta B^{(t+1)}$ で $\Delta B^{(t)}$ を近似する。

また、パケットサンプリングを行うことでサンプリングされないフローが発生する。このため、推定する際にもこれらのサンプリングされなかったフローを考慮する必要がある。しかし、サンプリングされなかったフローから推定を行うには、サンプリングによって、サンプリングされなかったフローの数 $f_0^{(t)}$ が必要となる。サンプリングを行った後のフローの総数を $F^{(t)}$ とし、サンプリングを行う前のフローの総数を $F^{(t-1)}$ とすると、$f_0^{(t)}$ は、$F^{(t-1)} - F^{(t)}$ となる。ここでも同様に $F^{(t-1)}$ の値を取得することができないので、サンプリングによってサンプリングされなかったフローの数 $f_0^{(t)}$ をサブサンプリングによって得られた $F^{(t+1)}$ を利用して、$F^{(t)} - F^{(t+1)}$ と近似する。

以上より、式 (12) を利用した推定を $t = 0$ となるまで繰り返すことで、オリジナルのフローの分布の推定を行う。提案手法は、簡単な数式を計算することで推定を行うので、容易でかつ高速にオリジナルの分布を推定することができる。

次に提案手法の具体的な手順を示す。

（1） $t \leftarrow \log_q s^{(t)}$ を計算。
（2） $F^{(t+1)}$ をサブサンプリングで取得。
（3） $f_0^{(t)} \leftarrow F^{(t)} - F^{(t+1)}$ を計算。
（4） $\Delta B^{(t+1)} \leftarrow 1/\Delta F^{(t+1)} \times 1/q$ を計算。
（5） $f_s^{(t-1)} \leftarrow f_j^{(t)}, (s = j/\Delta B^{(t+1)})$ を計算。
（6） $e(x)$ のパラメータ a, b を $f_s^{(t-1)}, (s = j/\Delta B^{(t+1)})$ から計算。
（7） $f_s^{(t-1)} \leftarrow e(s), (s \neq j/\Delta B^{(t+1)})$ を計算。
（8） $M^{(t-1)} \leftarrow M^{(t)}/\Delta B^{(t+1)}$ を計算。
（9） $d_{i,j}^{(t-1)} \leftarrow \frac{g_{i,j}^{(t-1)}}{\sum_{i=j}^{M^{(t-1)}} g_{i,j}^{(t-1)}}$ を計算。
（10） $f_i^{(t-1)} \leftarrow d_{i,j}^{(t-1)} \times f_j^{(t)}, \left(0 \leqq j \leqq M^{(t)}, j \leqq i \leqq M^{(t-1)}\right)$ を計算。
（11） $t \leftarrow t - 1$ を計算。
（12） $t = 0$ になるまで、手順 2 から 11 を繰り返す。

4. 推定結果

本章では、実際のトレースデータに対して提案手法を適用し、真の統計情報の推定を行う。そして数値結果により、提案手法が真の統計情報を精度よく推定することが可能であることを示す。ここで、フローの識別は、すでに述べた 5 つの情報が等しいパケットを同一フローとする。また、タイムアウトを 30 秒とし、5 つの情報が等しい場合であっても、ひとつ前のパケットを受信してから 30 秒以内に次のパケットを受信しなかった場合は別フローとみなす。評価は、真の統計情報に対する重み付き平均相対誤差 (WMRD) で行う。ここで、フロー数の確率密度関数 (PDF) を、

$$c_j^{(t)} = f_j^{(t)}/F^{(t)} \tag{18}$$

と定義する。$c_0(m)$ をオリジナルのフロー数の PDF、$c_0'(m)$ をサンプリングにより得られた統計情報から推定した真の統計情報におけるフロー数の PDF、m をフローのパケット数としたとき、WMRD を

$$WMRD \equiv \frac{\sum_i |c_0'(m) - c_0(m)|}{\sum_i (c_0'(m) + c_0(m))/2} \tag{19}$$

と定義する。WMRD は、値が小さいほどより高精度で真の統計情報を推定可能であることを示す。

図 2 から 図 5 に、Abilene III と Auckland II の各トレースデータをサンプリング間隔 8 および 128 でサンプリングしたデータにサンプリング確率 q を $q = 1/2$ とした提案手法を適用し推定した結果を示す。比較のために EM アルゴリズムを利用して推定を行った場合とオリジナルのフローの分布を示す。図の横軸

表2　推定したフローの分布とオリジナルフローの分布の WMRD

Datasets		Sampling Interval	Proposed method	EM
Abilene III	IPLS to KSCY	8	0.222763	0.302987
		128	0.646861	0.620598
	KSCY to IPLS	8	0.261485	0.412306
		128	0.561418	0.406428
Auckland II	Outbound	8	0.493364	0.430878
		128	0.498417	0.620273
	Inbound	8	0.540956	0.710972
		128	0.661831	0.953552

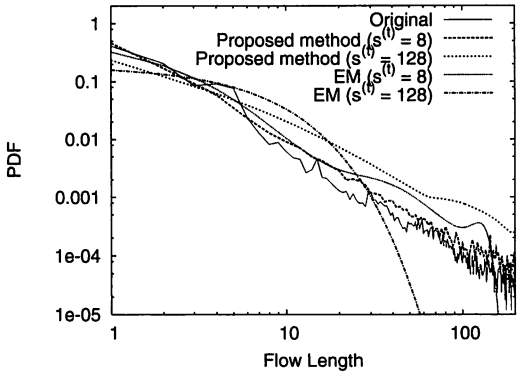

図2　パケット数別のフロー分布 (Abilene III IPLS to KSCY)

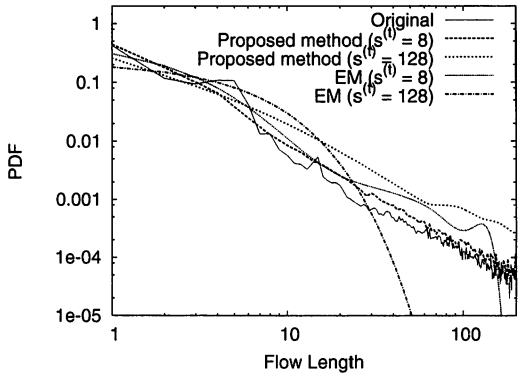

図3　パケット数別のフロー分布 (Abilene III KSCY to IPLS)

は、フローのパケット数を示し、縦軸は、フロー数の確率密度関数を示している。各トレースデータの推定結果の WMRD の値を表 2 に示す。

提案手法は、サンプリング確率が 1/2 であるので、サンプリング間隔が 8 でサンプリングされたデータから真の統計情報は、3 回推定を繰り返すことで得ることができ、サンプリング間隔が 128 のデータからは、7 回推定を繰り返すだけで真の統計情報を推定することが可能である。

図および WMRD の値より、提案手法は、サンプリング間隔に依存せず、正確にオリジナルのフロー分布を推定できていることがわかる。また、WMRD の値より、提案手法は、EM アルゴリズムと同程度の精度で推定できていることがわかる。図より、Auckland II のデータセットでは、パケット数が 8 以下の区間でサンプリング間隔が 8 のデータから推定された分布とオリジナルの分布との間に、大きな誤差が生じているが、これは、サンプリング間隔 8 でサンプリングが行われているために、パケット数が 8 以下の区間の特徴を得ることができていないためである。

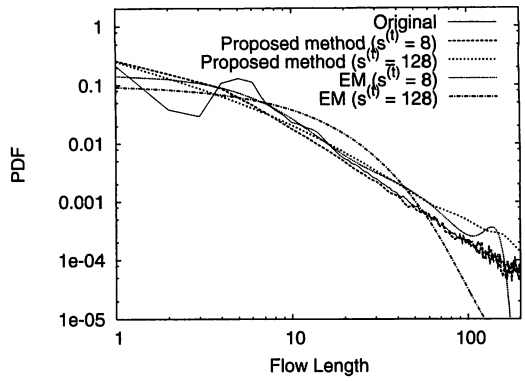

図4　パケット数別のフロー分布 (Auckland II Outbound)

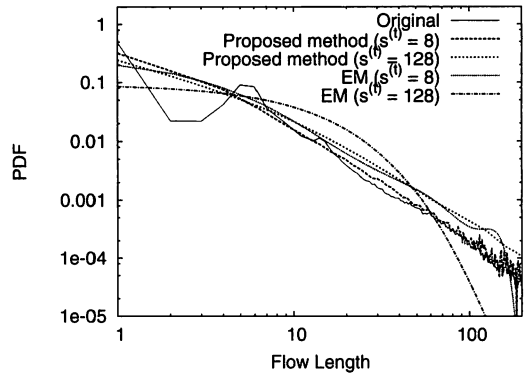

図5　パケット数別のフロー分布 (Auckland II Inbound)

また、EM アルゴリズムにおいて、サンプリング間隔が 128 の場合に誤差が大きくなっているのは、サンプリングが間隔 128 でサンプリングされたデータを利用しているために、パケット数が 128 以下の分布の特徴を得ることができていないためである。さらに、提案手法において、サンプリング間隔が 128 の場合、推定結果の分布の各値がオリジナルの分布の値に比べ、大きくなっているのは、$f_i^{(t-1)}$ を近似するのに必要となる $\Delta B^{(t)}$ を $\Delta B^{(t+1)}$ で近似しているために、スケーリングした際に誤差が生じるためである。このことにより、本来ならば、よりパケット数が多いフローが推定されるはずであるのに、適切なスケーリングが行われないことで、パケット数が少ないフローについてより多くフロー数が推定されるためである。この誤差の影響を軽減することができれば、提案手法において推定精度が向上すると考えられるので、今後、より正確に $\Delta B^{(t)}$ を近似できる手法を検討する必要がある。

5. 推定に必要な計算量

本章では、EMアルゴリズムと提案手法において、推定に必要となる計算量の評価を行う。ここで、推定するフローの最大パケット数を m、計測されたフローの最大パケット数を n とし、$m \leq n$ を満たすものとする。

EMアルゴリズムにおいて、Eステップでは、まず、推定するフローごとに計測された各フローのうち、推定するフローのパケット数以下のフローのパケット数を利用して期待値の計算を行うので、計算量は、$O(m)$ である。そして計算した各フローの期待値の合計を計算するため、必要な計算量は、$O(m^2)$ となる。また、Mステップでは、推定するフローごとのパラメータ更新に推定するフローのパケット数以下の計測されたフローの数と他の推定するフローのパラメータの値を利用するので、計算量は、$O(m^2)$ である。そして、フローごとのパラメータ更新を推定する全フローについて行うので、必要となる計算量は、$O(m^3)$ となる。よって、Eステップと Mステップを行うのに必要となる計算量は、$O(m^2 + m^3) = O(m^3)$ となる。

一方、提案手法では、式 (12) を利用して繰り返し計算を行う際に、まず、計測されたフローごとに推定する全フローについて計算を行うので、計測されたフローごとの計算量は、$O(m)$ である。そして、フローごとの計算を計測された全フローについて行うので、推定必要な計算量は、$O(mn)$ である。

以上より、提案手法は、EMアルゴリズムに比べ、少ない計算量で、EMアルゴリズムと同程度の精度の推定を行うことができる。

6. ま と め

本稿では、異なる間隔のサンプリングにより得られたフローの統計情報間の差分情報の分析を行い、パケットサンプリングによるフローのパケット数の変化を数式を用いて解析を行った。解析結果にもとづた差分情報のモデルを利用して、真の統計情報を容易に推定可能な手法を提案した。また、実際のトレースデータを分析することにより、提案手法が真の統計情報を精度良く推定することが可能であることを示した。

今後の課題としては、推定精度を向上させるために、フローのパケット数のスケーリングに利用している平均パケット数の変化率とサンプリングされなかったフローの総数をより高精度で近似することができる手法の検討、フローのパケット数のスケーリング後に欠落している情報をより実際の分布に近いものにするための補完式の再検討などがある。

謝　　辞

本研究の一部は、文部科学省科学研究費補助金若手研究 (A) (課題番号: 19680004) によって行われた。ここに記して謝意を表する。

文　　献

[1] IETF PSAMP Working Group, "Packet Sampling (psamp) Charter." http://www.ietf.org/html.charters/psamp-charter.html.

[2] N. Duffield, "Sampling for Passive Internet Measurement: A Review," *Statistical Science*, vol. 19, no. 3, pp. 472–498, 2004.

[3] H. Isozaki, S. Ata, and I. Oka, "Estimation of original flow distribution by using difference information from sampled flows," *IEICE Technical Report (IN2006-174)*, vol. 106, pp. 71–76, February 2007.

[4] N. Duffield, C. Lund, and M. Thorup, "Properties and prediction of flow statistics from sampled packet streams," in *Proceedings of the 2nd ACM SIGCOMM Internet Measurement Workshop (IMW 2002)*, pp. 159–171, November 2002.

[5] N. Hohn and D. Veitch, "Inverting sampled traffic," in *Proceedings of ACM SIGCOMM Internet Measurement Conference (IMC 2003)*, pp. 222–233, October 2003.

[6] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, pp. 1–38, May 1977.

[7] N. Duffield, C. Lund, and M. Thorup, "Estimating flow distributions from sampled flow statistics," in *Proceedings of ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM 2003)*, pp. 325–336, August 2003.

[8] Q. G. Zhao, A. Kummar, and J. J. Xu, "Data streaming algorithms for accurate and efficient measurement of traffic and flow matrices," in *Proceedings of the International Conference on Measurements and Modeling of Computer Systems (SIGMETRICS 2004)*, pp. 177–188, June 2005.

[9] T. Mori, R. Kawahara, N. Kamiyama, K. Ishibashi, and S. Harada, "Inferring original traffic pattern from sampled flow statistics," *IEICE Technical Report (NS2006-125)*, vol. 106, pp. 13–18, November 2006.

[10] NLANR Measuament and Network Analysis Group, "NLANR PMA: Special Traces Archive." http://pma.nlanr.net/Special/.

[11] S. Ata, M. Murata, and H. Miyahara, "Analysis of Network Traffic and its Application to Design of High-Speed Routers," *IEICE Transactions on Information and Systems*, vol. E83-D, pp. 998–995, May 2000.