サンプルフロー統計から元の トラヒックパターンを推定する方法

CQ/NS/TM 研究会
2006.11.16

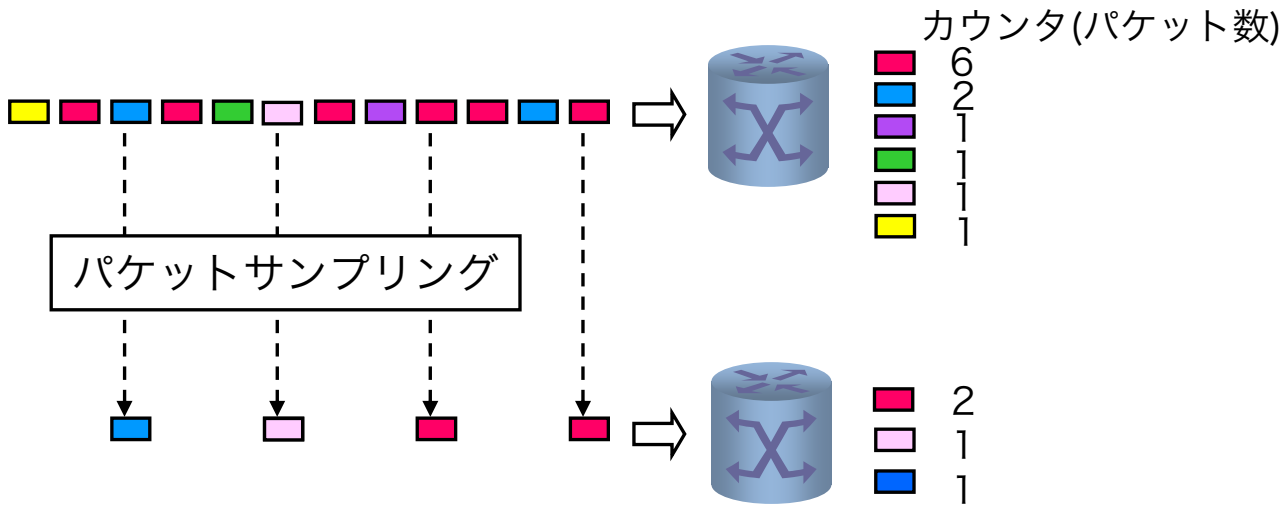
森達哉(1)　川原亮一(1)　上山憲昭(1)
石橋圭介(2)　原田薫明(1)

(1) NTTサービスインテグレーション基盤研究所
(2) NTT 情報流通プラットフォーム研究所

NTT

背景(1)

- **ネットワークの超高速化・大規模化**
  - スケーラブルなネットワーク監視技術
  - パケットサンプリング+フロー計測
    - 多くの実装・運用 (e.g., sampled NetFlow)



背景(2)

- **異常トラヒックの多様化**
  - DDoS : volume → heavy query
  - Worm : scanning → bot : scheduled scanning
- **ボリュームだけではなく、パターンに着目した異常検出が必須**
  - パターンを捕らえる尺度の一例：情報エントロピー [Anukool, sigcomm 2005]

本研究の課題と主な結果

- **課題**
  - 異常検出に必要となる統計（分布・エントロピー）を、サンプルデータから推定することは可能か？
- **主な結果**
  - EMアルゴリズムによる推定では元のトラヒックのサイズ分布を正しく推定することはできなかった
  - 追加の情報として、非サンプルフロー数の情報を使うと、サイズ分布の推定精度が著しく向上した

異常トラヒックとフロー統計

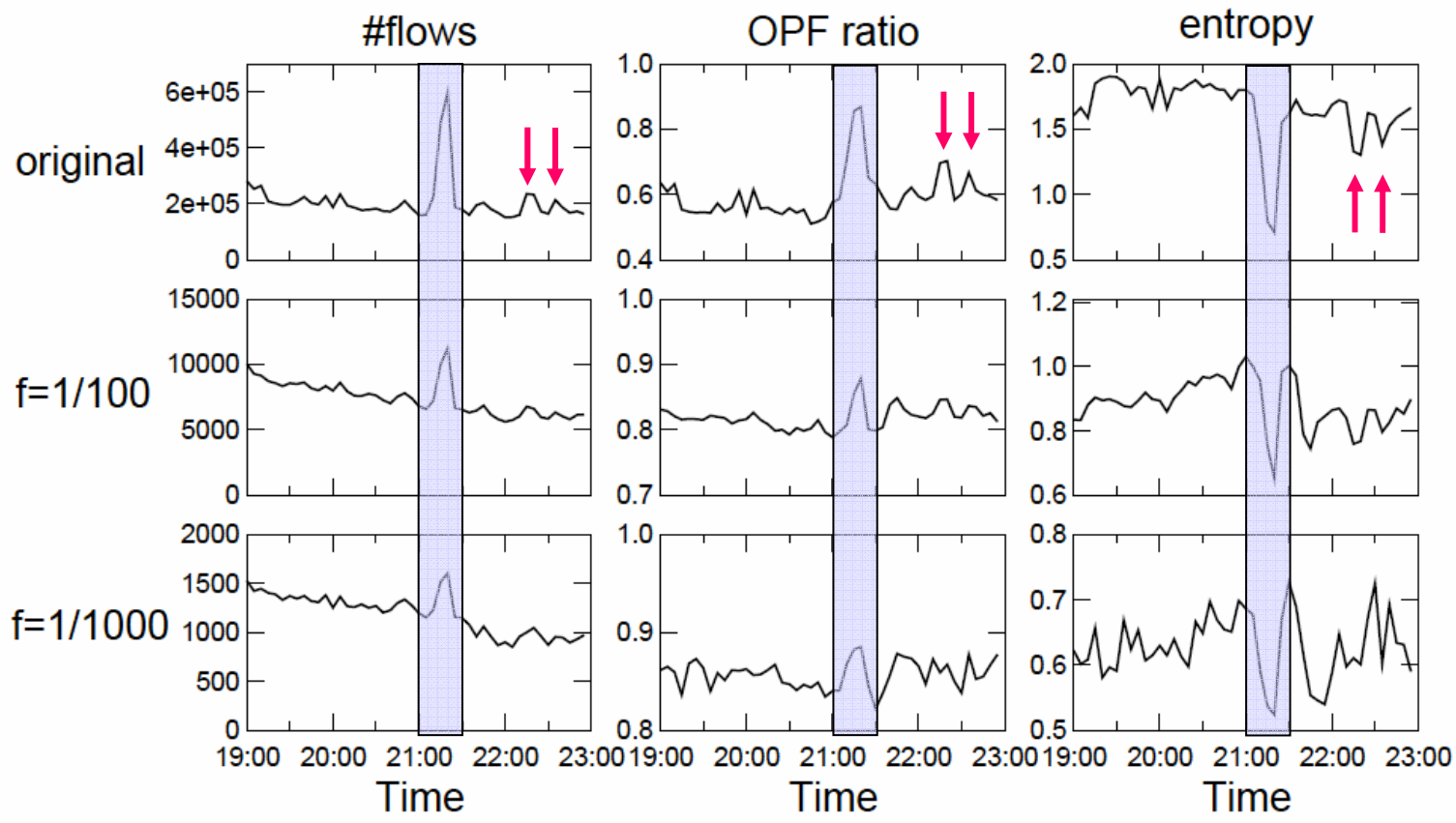
- **OPF (One-Packet Flow)**
  - 1パケットから構成されるフロー
    - network/port scan
    - DNS, NTP の一部
  - OPF 率の急激な増加 → スキャンの急激な増加

- **情報エントロピー**
  - 観測データの乱雑さをあらわす尺度
  - あるフローがi パケットからなる確率 P(i)
    - エントロピー = $-\sum_i P(i)\log P(i)$
  - エントロピー減少 → あるパケット数のフローに集中
  - エントロピー増加 → 様々なパケット数のフローに分散
  - OPF 率よりも表限力が高い

フロー統計間の相関

- **フロー数とOPF**
  - やや正の相関あり（相関係数=0.795）
  - 「フロー数の増加」⇔ 「OPF増加」は必ずしも成立しない
    - フロー数を観測しているだけでは不十分
- **OPFと情報エントロピー**
  - 強い負の相関あり (相関係数=-0.958)

サンプリングとフロー統計



サンプリングとフロー統計

- **フロー数、OPF、エントロピーのいずれもサンプリングによってパターンの変化が見えなくなる**

- **フロー分布をモデル化し、サンプルデータから分布のパラメタを推定する問題を考える**
  - 分布のモデル化（パラメトリックなアプローチ）
  - 不完全なデータに基づく最尤推定 → EMアルゴリズムを適用

フローサイズ分布のモデル

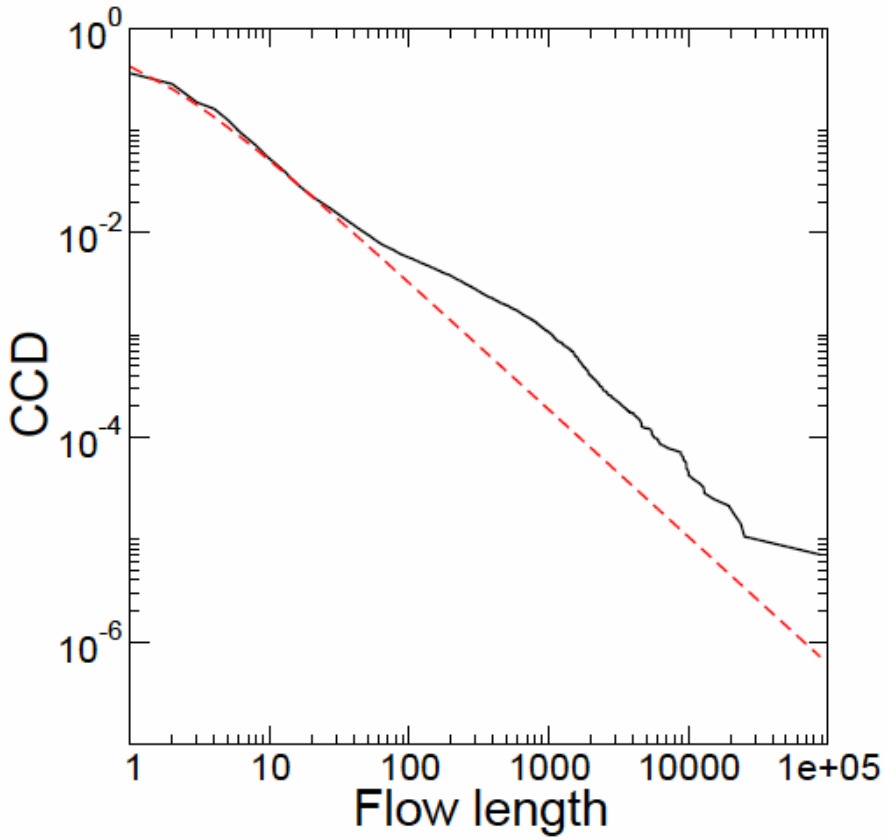
- あるフローのパケット数を離散確率変数 X とする
- X を離散パレート分布によってモデル化

$$f(x) = \frac{\theta a^{\theta}}{x^{\theta+1}}$$

$$p(k;\theta) = \Pr[X = k] = \int_{k}^{k+1} f(x)dx = k^{-\theta} - (k+1)^{-\theta}$$

- 1 パラメタのシンプルなモデル

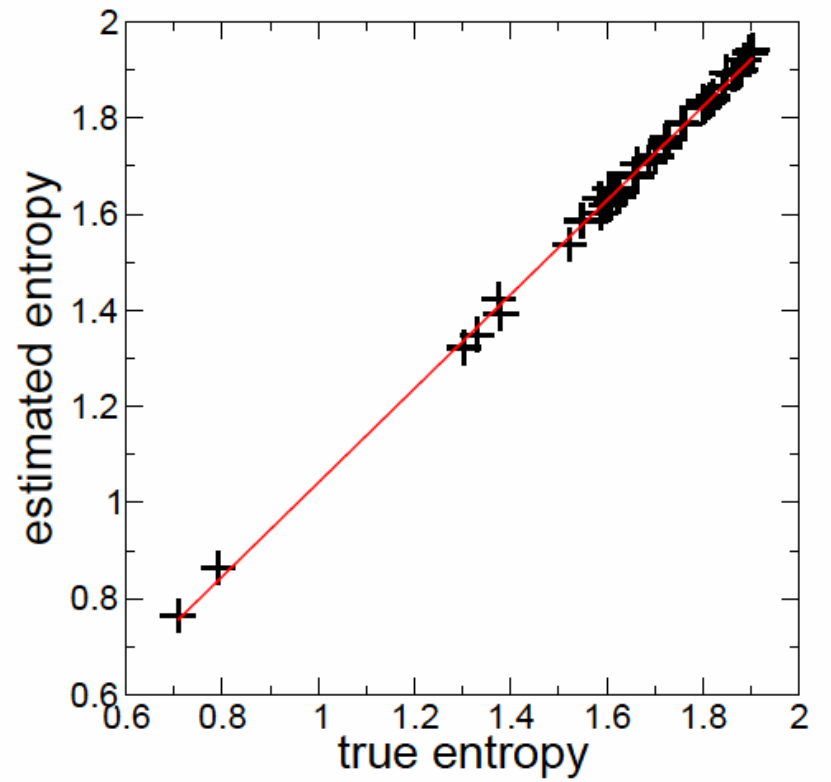
モデルの有効性(1)



サンプル前のデータより、離散パレート分布のパラメタを最尤推定した結果

特にフロー長の小さなところ（頻度の高いところ）でよくフィットする

モデルの有効性(2)



いくつかのデータについて実際のエントロピーとモデルから計算したエントロピーを比較した結果
（相関係数=0.999)

離散パレート分布モデルはエントロピーの性質をよくとらえている

EMアルゴリズムによるフロー分布推定(1)

- サンプル後にフローが i パケットを持つ確率

$$r(i;\theta) = \sum_{k=i}^{N} q\left(i \mid k\right) p\left(k;\theta\right)$$

- ただし q(i|k) は二項分布, f はサンプリング確率

$$q(i \mid k) = \binom{i}{k} f^{i} \left(1-f\right)^{k-i}$$

EMアルゴリズムによるフロー分布推定(2)

- 完全対数尤度 (ni: i パケットサンプルされたフロー数)

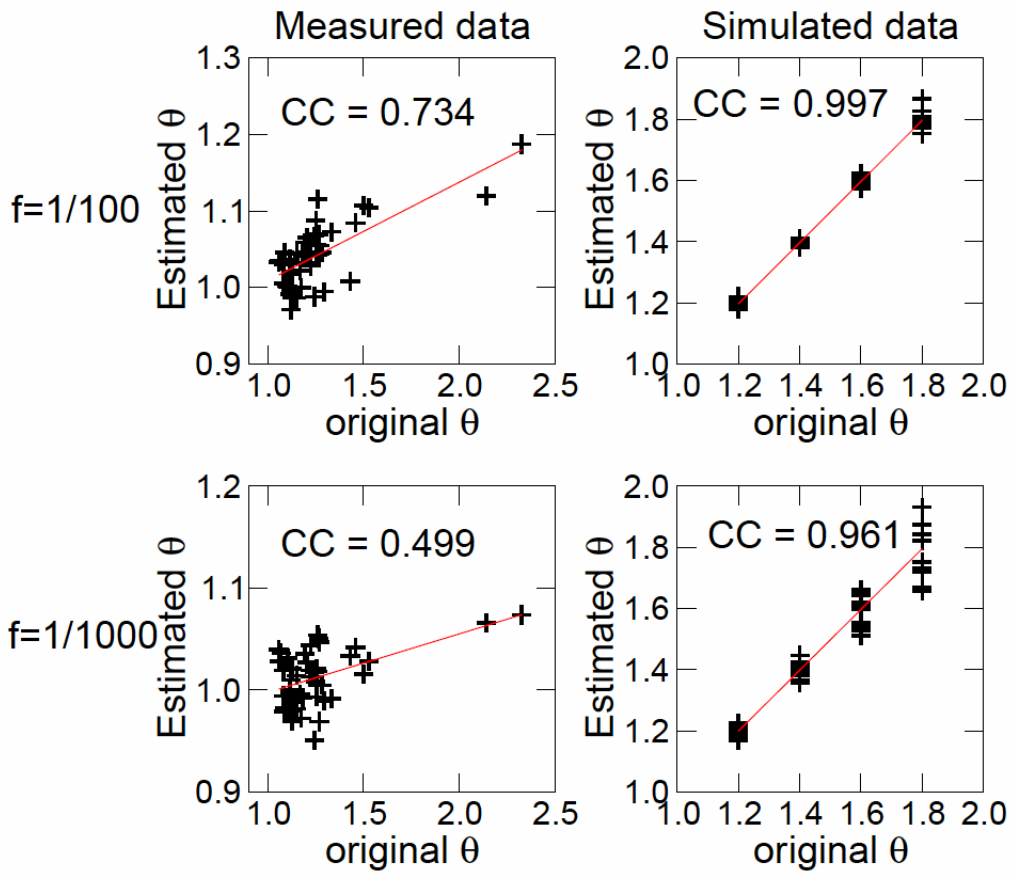
$$\log L_c(\theta) = \log \prod_{i=0}^{y_{\max}} r(i;\theta)^{n_i} = \sum_{i=0}^{y_{\max}} n_i \log r(i;\theta)$$

- パラメタの初期値を設定
- E-step: 観測データ(=サンプルデータ)と現在のパラメタを用い、完全対数尤度の条件付期待値を計算

$$Q(\theta;\theta^{(0)}) = E_{\theta(0)}\left\{\log L_c(\theta) \mid n_1, \ldots, n_m\right\}$$

- M-step: 上記を最大化するパラメタを計算
- E-step に戻る（推定値が収束するまで繰り返す）

EMアルゴリズムによるフロー分布推定(3)



実計測データ（WIDEプロジェクト国際線）および人工データ(Pareto分布)に対し、EMアルゴリズムによる分布パラメタ推定を行った結果

実データの方は推定精度がよくない

人工データの方は比較的良いが、サンプリングレートが低くかつ形状母数が大きい場合は精度がよくない

EMアルゴリズムによるフロー分布推定(4)

- **推定がうまくいかない理由**
  - パレート分布モデル: 大多数が小フロー
  - OPF のほとんどがサンプルされない
    - → 小フローの情報が消える
- **考えうる改良案**
  - 未観測データ(非サンプルフロー数)の情報を使うことによって推定精度をあげる

フロー数のカウント方法

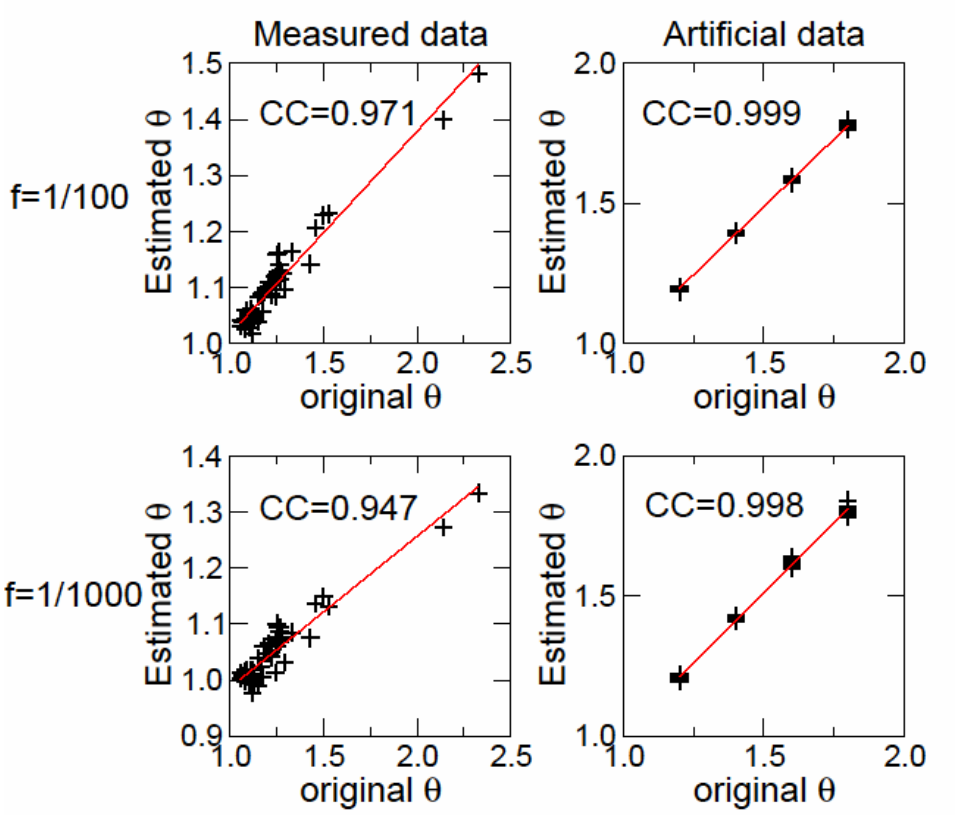
- **直接計測（推定）**
  - 専用ハードウェアで計測 [Estan, sigcomm 2004]
  - Bloom filter, probabilistic counting などの方法によって、超高速・小メモリ空間で到着フロー数を確率的にカウント [Keys, sigmetrics 2005]
  - ☹ルータの機能拡張が必要
- **間接的に推定**
  - サンプルしたTCP SYN フラグの数よりTCPフロー数を推定 [Duffiled, sigcomm 2003]
  - ☹TCP のみ適用可能

フロー数が与えられた元での最尤推定

- **完全対数尤度を最大にするパラメタが求めるパラメタとなる**
  - n_0：全フロー数 – サンプルフロー数

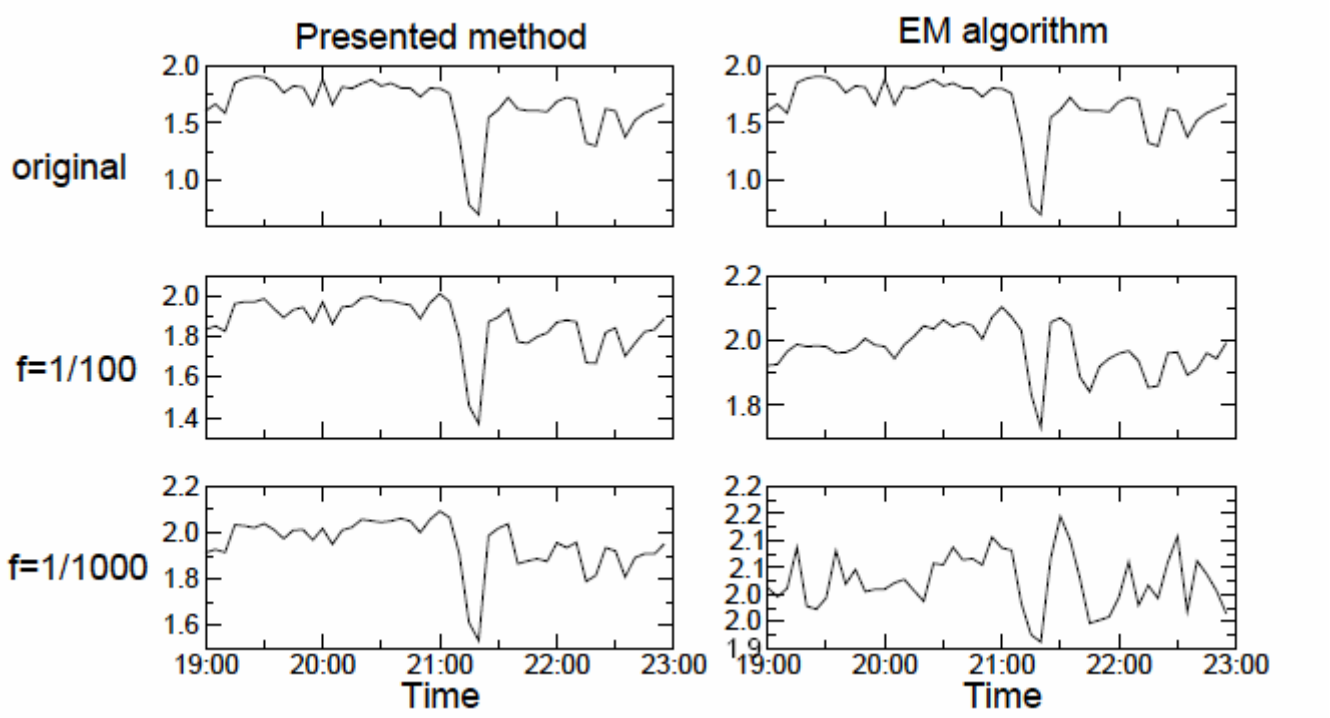
$$\log L_c(\theta) = \log \prod_{i=0}^{y_{\max}} r(i;\theta)^{n_i} = \sum_{i=0}^{y_{\max}} n_i \log r(i;\theta)$$

推定結果(1)



EMアルゴリズムによる推定と比較して、精度が著しく向上

推定結果(2)



エントロピーの変化（急激な減少）
をとらえることができた

まとめ

- **パケットサンプリング + 異常検出**
  - OPFやエントロピーの変化がパケットサンプリングによってみえなくなる問題がある
  - フロー分布をサンプルデータから最尤推定するアプローチを提案
- **フロー分布推定**
  - EMアルゴリズムでは正しく推定できない例を示した
  - 非観測情報であるフロー数情報を取り入れることで推定精度が向上

今後の課題

- **離散パレートモデルの妥当性**
  - 他のデータでの検証
  - 他の分布モデル・ノンパラメトリックな方法 [Duffield sigcomm 2003]との比較
    - 精度・計算量
- **フロー数カウント**
  - Light weight & practical なカウント法