実インターネットトラフィックデータへのCPP手法の適用とその改善に関する研究

A Method to Apply the Cluster Point Process for Short-Time Internet Traffic Traces

by

本舘 拓也

Takuya Motodate

48-106435

February 8, 2012

A Master's Thesis Submitted to
Department of the Information and Communication Engineering
Graduate School of Information Science and Engineering
the University of Tokyo

指導教官:　江崎浩 教授

Abstract

インターネットトラフィックの特性把握や，異常検出などのアプリケーションのため，これまでさまざまなトラフィックのモデル化が提案されてきた．本研究では，マルチスケールガンマモデルというモデルに着目し，研究を行った．一方，広帯域のバックボーンリンクでは，さまざまな技術的な制限のため，パケットサンプリングの手法を用いてトラフィックの特徴把握を行うことが一般的となっている．本研究ではまず，MAWI Traffic Trace という日米間の実トラフィックデータを用いてパケットサンプリングの影響がマルチスケールガンマモデルに対してどのように及ぶのかの調査を行い，そのパラメータの動きが大きく二つの状態に分かれることを示した．しかし，実トラフィックには統計的な異常が含まれており，結果の一般化において困難が伴うことが明らかとなった．そこで，トラフィックの再生成の手法として有効であるとされている Cluster Point Process を取り上たが,，短時間のトラフィックに対して Cluster Point Process をそのまま適用することができないことを発見することができた．そこで，この原因を調査し，問題が起こる原因の特定と，それを回避するためのパラメータチューニング手法を提案し，その評価を行った．その結果，提案した 3 つのチューニング手法のうち最もよい手法を同定し，元のチューニング手法よりもよい特性を持つことを示した．

Wide variety of Internet traffic model has been proposed to understand characteristics of traffic and is applied to various applications(e.g., anomaly detection methods). This paper focuses on multi-scale gamma model as a traffic model. Since packet sampling is common in current high speed backbone network, the evaluation on the effects of the each sampling method is important so as to adopt the particular traffic model for the given sampling method. This paper challenges this issue by using actual academic traffic captured between Japan and the United States and finds that parameter dynamics can be divided to two parts, depending on each sampling methods. To generalize this result, we employ the Cluster Point Process to generate synthetic traffic, and obtain clearer view of the parameter dependency of the model. However synthetic traffic generated by the Cluster Point Process fails to have same characteristics of the original traffic. We point out reasons of the failure and propose a solution that is a new method of parameter tuning of the Cluster Point Process. We evaluate this method by using the Logscale Diagram, and choose the best method to fit it.

Contents

List of Figures

List of Tables

Chapter 1

はじめに

1.1 バックボーン回線と異常検出

今日，“ネットワークのネットワーク” として世界中のコンピュータネットワークをつないでいるインターネットは，既に社会の新たなインフラとして十分な認知を得られている．インターネットにおいて，ネットワークの相互接続や長距離の伝送を行うような，インターネットの中心部分のことをバックボーンと呼んでいる．バックボーンはインターネットを支える広帯域の幹線である．バックボーンの例として，日本の学術機関のネットワークを相互接続している SINET が挙げられる．

バックボーンに適用されるアプリケーションとして，トラフィック流量推定や異常検出があるが，本稿では特に異常検出に着目する．異常検出 (Anomaly Detection) とは，一般にはデータマイニングによってデータ列に含まれる特異なケースを抽出することを指す．インターネットトラフィックでの異常検出では，この統計的に特異なケースを抽出することによって何らかの攻撃や不具合を検出することを目的としている．この異常検出では，評価に当たってそのバックボーンを通るトラフィックのデータが必要となる．

1.2 ネットワークでのセキュリティ

インターネットが普及し，そこでのウィルスやワームなどが与える被害が大きくなっていくにつれ，ネットワークにおけるセキュリティの問題が顕在化してきた．この対策として考えられる手段として，インターネットに接続するホストでの対策がある．市販のウイルス対策ソフトなどがこれに該当する．一方，各ホストでの対策とは別に，ネットワークの側で悪質な通信を検知するようなシステムが考案されてきた．

ネットワークでのセキュリティ対策は，大別してインターネットの末端側で行われるものと，バックボーン側で行われるものに大別できる．末端側で一般的に使用されているのは IDS(Intrusion Detection System:侵入検知システム) という機構である．これは，監視対象のリンクを流れるパケットに対してパターンマッチングを行い，登録されたパターンに該当するような通信パターンが観測された場合，ネットワーク管理者に通知を行う機構である．代表的なソフトウェアとして Snort が挙げられる．設置場所としては，企業や大学などの組織内の LAN からインターネットへ接続するゲートウェイ部分に置くのが一般的である．一

方，ネットワークのバックボーンでは，統計手法を用いた異常検出の手法が考案されている．ネットワークでのセキュリティ機構をバックボーンネットワークの段階で行うメリットとして，幾つかの点が挙げられる．

インターネット全体での傾向把握が可能 特定ウィルスやワームの流行り廃りの可視化が可能

攻撃トラフィックの通信パターンが分かる Scan は典型的なトラフィックパータン

フィードバックが可能 攻撃元 IP アドレスや攻撃に使用されるポートのフィルタリング

以上のような特性から，これまでバックボーンを対象とした異常検出手法が提案されてきた．

1.3 パケットサンプリングと影響評価

しかし，バックボーンは広帯域であるため，そこを通るパケットを全て保存することは技術的に難しい．そこで，バックボーンを流れるトラフィックはパケットサンプリングを通してその特徴を把握するということが一般的に行われている．パケットサンプリングとは，あるリンクを通るパケットを特定のアルゴリズムに従って拾い上げるものである．パケットサンプリングのうち，基本的なサンプリング手法については RFC5476 にて標準化が行われている [6]．標準化が行われているサンプリング手法には，パケットのデータに対して独立なもの (Systematic Sampling, Uniform pseudorandom Sampling)，パケットのデータを参照して行うもの，ハッシュ関数を用いて行うもの (Hash-based Selection) がある．

パケットサンプリングが適用されたトラフィックに対して，異常検知を適用した際の影響を評価した研究はいくつか行われている．バックボーントラフィックにサンプリングを適用した際の影響評価のさきがけとして Claffy らの評価がある [15]．この評価では，サンプリング手法として簡単な Packet-based と Time-based の手法 5 つをバックボーンのトラフィックに対して適用し，評価を行った．また，Choi らは，Sampled NetFlow に関して，(i) サンプリングレートと理論的な性能限界の考察，(ii) 実装のオーバーヘッドおよび性能検証，(iii) サンプリング手法の性能評価を，総通信量の推定値の分散の大きさを基準として行った [5]．Duffield らは，パケットサンプリングされた状態から得られた値を元に，元の値をどの位復元できるかの詳細な調査を行った [8]．この調査の問題意識として，(i) サンプリングによるフロータイムアウトの発生，フロー数の推定の正しさ，(ii) サンプリング後に記録されたフローに含まれるパケット数からオリジナルの推定の正しさ，(iii) サンプリング後に記録されたフローのバイト数からオリジナルの推定の正しさなどがある．Brauckhoff らは，パケットサンプリングが異常検出手法で用いるパラメータについて与える影響について調査を行った [3]．

筆者は卒論において，パケットサンプリング手法がマルチスケールガンマモデルを用いた異常検知手法に与える影響評価を行い，そこからさらにサンプリング手法がマルチスケールガンマモデル自体に与える影響とその記述パラメータの変化モデルの構築を試みた．この過程において，実トラフィックにおいて統計的な異常が含まれていることが逆に問題となった．異常が含まれているトラフィックでパケットサンプリングの影響評価を行うと，パラメータ変化のパターンに影響を及ぼす可能性があるためである．そこで，異常の影響を無くすために，適切なトラフィックモデルを用いてトラフィックの再構成を行うことにより，異常を含まない統計的定常性が保たれたトラフィックを用いて評価を行うことが考えられる．

1.4 インターネットトラフィックのモデル化

トラフィックのモデル化は，過去に様々な研究が行われてきた．モデル化を行う対象として，パケットの到着過程，フローの到着過程，トラフィックの周辺分布などがあるが，本稿ではパケットの到着過程に着目する．なぜならば，パケット到着はインターネットトラフィックを考える上で最も基本的な過程であるためである．バックボーンにおけるパケットの到着過程に関しては，Long-Range Dependence と Self-Similarity という性質があることが広く知られている．電話の呼の到着過程として広く用いられている Poisson 過程は，この性質を満たさないことからパケットの到着過程としてはふさわしくないことが指摘されており [14]，この性質を満たすようなパケット到着過程のモデルが長年議論されてきた．本稿では，それらのモデルのうち，Cluster Point Process について着目する．このモデルは，Logscale Diagram と呼ばれる離散ウェーブレット変換を用いた手法を使い，そこでの挙動を替えないようなトラフィック再生成を目指したモデルとなっている．このモデルの優れた点として，Long-Range Dependece が成立した理由をフローに含まれるパケット数 (フローサイズ) の広がり方に結びつけて考え，(i) フローの到着過程 (ii) フローサイズの分布 (iii) フロー内のパケット到着過程の 3 つの独立な過程を考えることで，Logscale-Diagra 上で等価なトラフィックの再生成に成功している点である．

1.5 本研究の目的と貢献

この Cluster Point Process を，実トラフィックデータである MAWI Traffic Trace に適用にすることを試みた．このトラフィックデータは，毎日 14 時から 15 分間の日米間のバックボーン回線のトレースデータである．この，900 秒間のトレースへの適用に関して，トラフィックの再構成に問題が起こることが分かった．本研究では，その問題が起こる発生の原因を指摘し，それを回避するためのパラメータチューニング手法を提案した．また，既存のチューニング手法と提案手法において，5 年分のトラフィックデータに対してトラフィックの再構成を行い，再生成したトラフィックの精度評価を行った．結果として，提案したチューニング手法がもとの手法よりも元のトラフィックデータの特性をよく保存していることを示した．

本研究の貢献は，Cluster Point Process のツールの実装及び公開，Cluster Point Process による MAWI Traffic Dataset の評価データセットの公開，Cluster Point Process を短いダンプデータに適用する際の評価方法の提案である．

1.6 本論文の構成

第二章では，インターネットトラフィックの性質とモデル化についての関連研究について紹介を行う．第三章では，これまでに行ったパケットサンプリングがトラフィックのモデルに与える影響評価について述べる．第四章では，本研究で用いる Cluster Point Process の詳細な紹介を行う．第五章では，Cluster Point Process を MAWI Traffic Trace におけるデータに適用する際の問題点と，その解決のための提案を行う．第六章で，提案手法の評価を行い，第七章で，その結果の議論を行う．

Chapter 2

関連研究

2.1 インターネットトラフィックの性質

2.1.1 Long Range Dependence

今，一定時間内に到着するパケット数について考える．これを確率変数として考える．この確率変数 X を離散的に扱う．つまり $X_n, n = 1, 2, \cdots$ とする．一定時間 Δ 内に到着したパケット数が図 2.1 のようだった時，$X_1 = 1, X_2 = 3, X_3 = 2, \cdots$ と表せる．もし X_n が独立であるなら，X_n それぞれについての分布を定式化することが可能だが，X_n が独立でない場合はそうすることはできない．

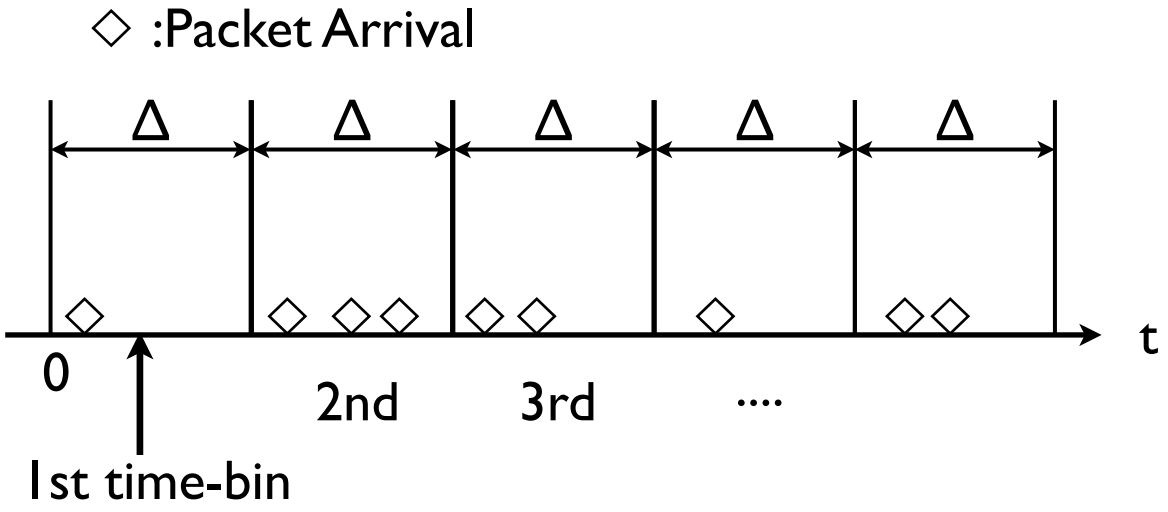

Figure 2.1: パケット到着数のモデル化

各確率変数 X_i の従属度を測る指標として便利なものに，共分散がある．確率変数 X_i, X_j の共分散は $\sigma_{X_i,X_j} = Cov(X_i, X_j) = E[(X_i - \mu_{X_i})(X_j - \mu_{X_j})]$ のように定義される．ここで，

X_i の描写を行うにあたって，

$$p_{X_1}(\cdot), p_{X_2}(\cdot), \cdots \tag{2.1}$$

を first order properties,

$$p_{X_1,X_2}(\cdot), p_{X_1,X_3}, \cdots, p_{X_2,X_3}(\cdot), p_{X_2,X_4}, \cdots \tag{2.2}$$

を second order properties と呼ぶ．ある統計過程が厳密に定常であるとは，

$$p_{X_n,X_{n+1},\cdots,X_{n+N-1}}(\cdot) = p_{X_{n+k},X_{n+k-1},\cdots,X_{n+k+N-1}}(\cdot) \tag{2.3}$$

を任意の n, k について満たすことを言うが，これは厳しすぎる過程であるため，一般的には用いられない．もっと単純な仮定として，ある統計過程が二次定常であるという仮定がある．これは

$$p_{X_i,X_j}(\cdot) = p_{X_{i+k},X_{j+k}}(\cdot) \tag{2.4}$$

が任意の i, j について成り立つ，というものである．また，さらにシンプルな仮定として，広義定常性

$$E[X_n] = E[X_1], Cov(X_n, X_{n+k}) = Cov(X_1, X_{k+1}) \tag{2.5}$$

が任意の n, k について成立する，というものがある．また，一番シンプルなものとして，first order stationarity があり，これは

$$p_{X_i}(\cdot) = p_{X_{i+k}}(\cdot) \tag{2.6}$$

が任意の i, k について成立する，というものがある．これが整理するのは，先述した確率変数 X_i が独立である必要がある．

広義定常性が成り立つ過程においては，自己共分散 (autocovariance) は確率変数 X_i と X_{i+k} の距離 k のみに依存する．これをラグが k であると呼ぶ．ラグが k の時の自己共分散の値を $\gamma(k)$ と表記すると，自己相関 (autocorrelation) $r(k)$ は自己共分散によって正規化され

$$r(k) \equiv \gamma(k)/\gamma(0) = \gamma(k)/\sigma_X^2 \tag{2.7}$$

と定義される．自己相関は-1 から 1 までの値をとり，0 以外の値を取るとき，ラグ k において確率変数が独立でないことを表している．

2.1.2 Self Similarity

平均値が 0 で，離散時間の定常過程 X_n は，もし任意の m に対して，集約過程 $X^{(m)}$ が $m^H X_n$ と同じ分布を持っていれば，Hurst Parameter H をもった self-similar であると呼ばれる．つまり，

$$m^H X_n =^d X^{(m)} m > 0, \frac{1}{2} \leq H < 1 \tag{2.8}$$

これはつまり，集約過程は統計的に見ると元々の過程を m^H でスケールしたものに見える，ということである．今，$\{X_n\}$ の平均値が 0 である場合を考えたが，$\{X_n\}$ が平均値 μ を持つ場合は $\{X_n - \mu\}$ が self-similar であると考える．

もし，$\{X_n\}$ が有限な分散をもつ独立な乱数であった場合，中心極限定理により，標準偏差は $\sqrt{m}$ に$\propto$する．$\{X_n\}$ が 0 でない平均をもつ場合，平均値は m に比例するので，変動係数 σ/μ は $1/\sqrt{m}$ に比例して急速に縮んでいくことになる．このとき，$H = 1/2$ となる．一方，$X > 1/2$ のときは，集約過程は上記の時ほど急速に収束しない．変動係数は m^{H-1} に比例する．また，$H = 1$ の時は，変動係数は一定値となる．

2.1.3 フローサイズの分布

フローのことを書きたい．ON-OFF モデル．セッション，各セッションの中で通信が行われている区間 (train)，train の中で実際にパケットを処理している時間の 3 階層のモデル．フロー．パケットのペイロードを見ないと，その通信を行なっているアプリケーションを正しく把握することができない．しかし，トラフィックのサマリやモデリングにおいては，中身の情報は不要である．そこで，通信の単位としてフローという単位が一般的に用いられている．IP フロー：IP ヘッダの情報をもとに分別されるフロー．いくつかの集約レベルが存在する．RFC でのフローの定義 [16] ．Network Defined Flow：IP ヘッダの情報だけでなく，Ingress や Egress のルータ情報を含めて分別を行なっているもの．

分布関数 $F(x)$ を持つような変数 X について，大きな x を考える時は $1-F(x) = P[X > x]$ の分布について考える．分布の裾が指数的に減衰するとは，$\lambda > 0$ を用いて

$$1 - F(x) \sim e^{-\lambda x} \tag{2.9}$$

となるような分布のことを指す．これが意味するところは，x が大きくなるにつれて，x が観測される確率がそれ自体の値よりも小さくなっていくことである．このような特徴を持つ分布を short-tailed もしくは light-tailed と呼ぶ．一般的な分布に関しては，分布の裾は指数関数かそれより早く減衰する．例えば，一様分布と正規分布はそれに該当する．このような分布では，非常に大きな x が現れる確率は十分に小さいものとして無視することができる．

対照的に，分布の裾が指数的なものよりもゆっくり減衰するものを，subexponential 分布と呼ぶ．形式的には以下のように表すことができる．

$$(1 - F(x))e^{\lambda x} \to \infty \text{ as } x \to \infty \text{ for all } \lambda > 0 \tag{2.10}$$

このような分布は長い裾 (long tail) を持つと言い，非常に大きな，もしくは無限大の分散を持つ．また，非常に大きな観測値 x が無視できない頻度で現れる．この subexponential な分布の中で特別な場合を heavy-tail distribution と呼ぶ．このような分布は以下のような式によって表される．

$$1 - F(x) \sim x^{-\alpha} \tag{2.11}$$

また，確率密度分布関数は以下のように書ける．

$$p(x) \sim x^{-\alpha-1} \tag{2.12}$$

このような分布において，x の分散は無限大になる．また，$0 < \alpha \leq 1$ においては平均も無限大になる．

2.2 ツール

2.2.1 Logscale Diagram

トラフィックにおける Hurst Parameter を推定する方法として，Logscale Diagram という手法が提案されている [1]．これは，時間スケールを表す j と，下記に定義される S_j をプロットすることで得られる．

$$S_j = \frac{1}{n_j}\sum_{n_j}^{k=1}|d_X(j,k)|^2 \sim C2^{j(2H-1)}(2^j \to \infty) \tag{2.13}$$

ただし，$d_X(j,k)$ は離散ウェーブレット変換の X_{Δ_0} の $2^{j\Delta_0}$ における係数で，時間軸上の位置は $k2^{j\Delta_0}$ である．実際に作成した Logscale Diagra の図を図 2.2 に示す．

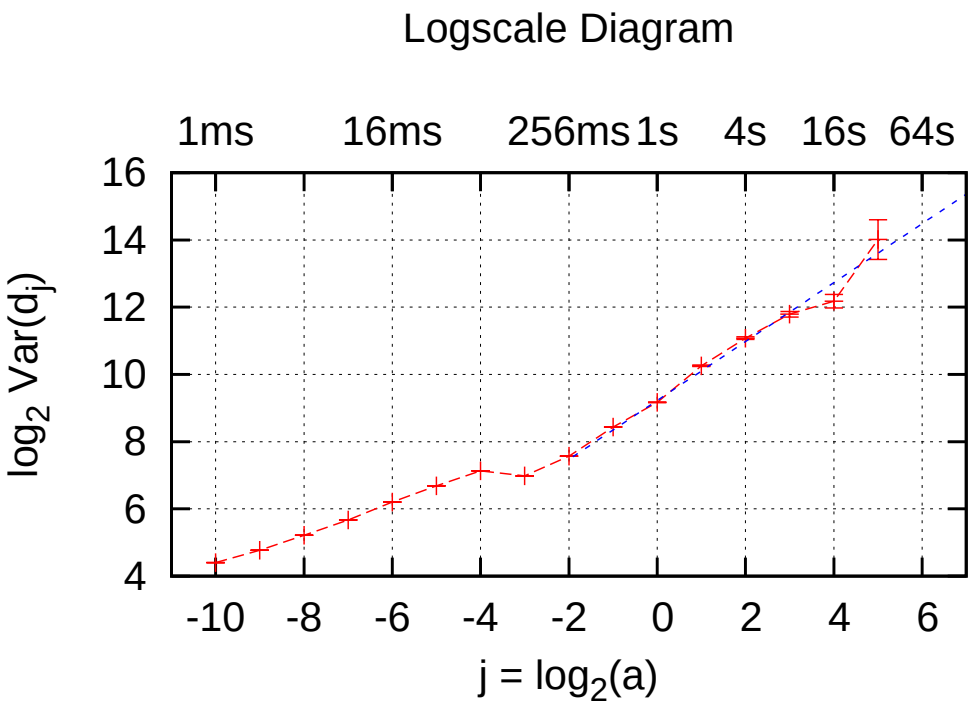

Figure 2.2: トラフィックの時系列から生成した Logscale Diagram

図 2.2 の右側に関して，各成分が一直線上に並んでいることが観察できる．この傾きから Hurst Parameter を推定できる．この直線部分は一般的に 1s のあたりから成立することが知られている．図 2.2 の場合は 256ms の付近から直線が現れていることが分かる．この傾きを α とすると，$\alpha = 2H - 1$ が成立することが知られている．図 2.2 の場合，$H = 0.932$ となっている．

2.3 インターネットトラフィックのモデル化

2.3.1 マルチスケールガンマモデル

コンピュータネットワークのトラフィックは IP パケットの到着過程によって構成されている．これを $\{(t_l, A_l), l = 0, 1, 2, \cdots\}$ と数式で表すことができる．ただし，t_l は l 番目のパケットの到着時間を表し，A_l はパケットの何らかの属性 (例えばペイロード，送信元または送信先 IP アドレス等) を表すものとする．このパケット到着過程は，単純なポアソン過程や

再生過程とは違っていることが指摘されている [13]．パケットの到着間隔に着目したモデル化の提案もある [11], [2] が，様々なネットワークとつながるような，多数のパケットを含むトラフィックを扱うと，その結果も巨大なデータセットになってしまう欠点がある．したがって，扱うデータを少なくしたいとき，これらを集約した到着パケット数やバイト数という形でトラフィックを扱うことがある．ある単位時間 $\Delta > 0$ (これをウインドウと呼ぶ) を定めた時，k 個目のウインドウは $k\Delta \leq t_l < (k+1)\Delta$ となるが，このウインドウに含まれる到着パケット数及びバイト数をそれぞれ，$X_\Delta(k), W_\Delta(k)$ と表すことにする．この $X_\Delta(k), W_\Delta(k)$ について，様々なモデル化が行われてきた [9] [7] [20] [17] [19] [21]．近年の評価では $X_\Delta(k)$ と $W_\Delta(k)$ は同様の統計的な特性を持っているため，特に $X_\Delta(k)$ について絞って評価が行われている．

$X_\Delta(k)$ は，定義より正の確率変数として見ることができる．従って，この分布を片側指数分布やワイブル分布，ガンマ分布などで記述する試みが行われてきた [12]．$X_\Delta(k)$ は経験的に，大きい Δ に対してはガウス分布により近似ができ，小さい Δ に対してはポアソンまたは指数分布による近似が期待できることが明らかになっている [18]．このような傾向のため，$X_\Delta(k)$ の分布をガンマ分布を用いたモデル化が行われており，これをマルチスケールガンマモデルと呼ぶ．ガンマ分布を用いるメリットとして，ガンマ分布は指数分布からガウス分布へ連続した分布を表現することができる点と，経験的に広いレンジの Δ に対して X_Δ をよく近似できることが明らかになっている [18]．

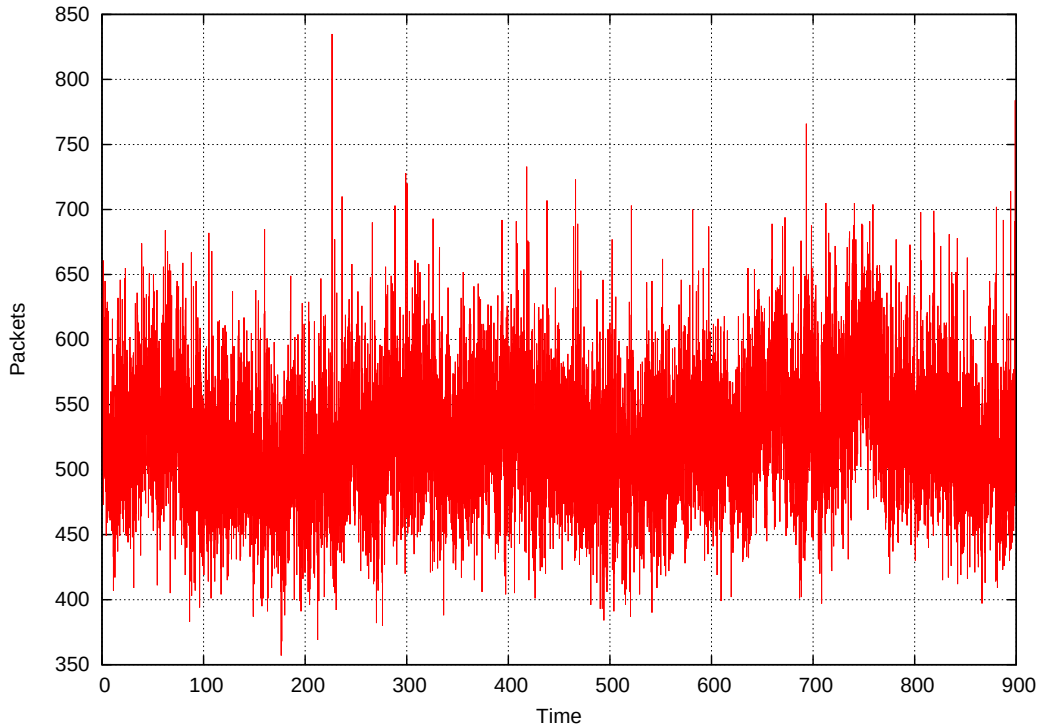

Figure 2.3: パケット到着数の時系列変化

ガンマ分布はパラメータ α, β を用いて次のように表される．ただし，$\Gamma(u)$ は標準ガンマ

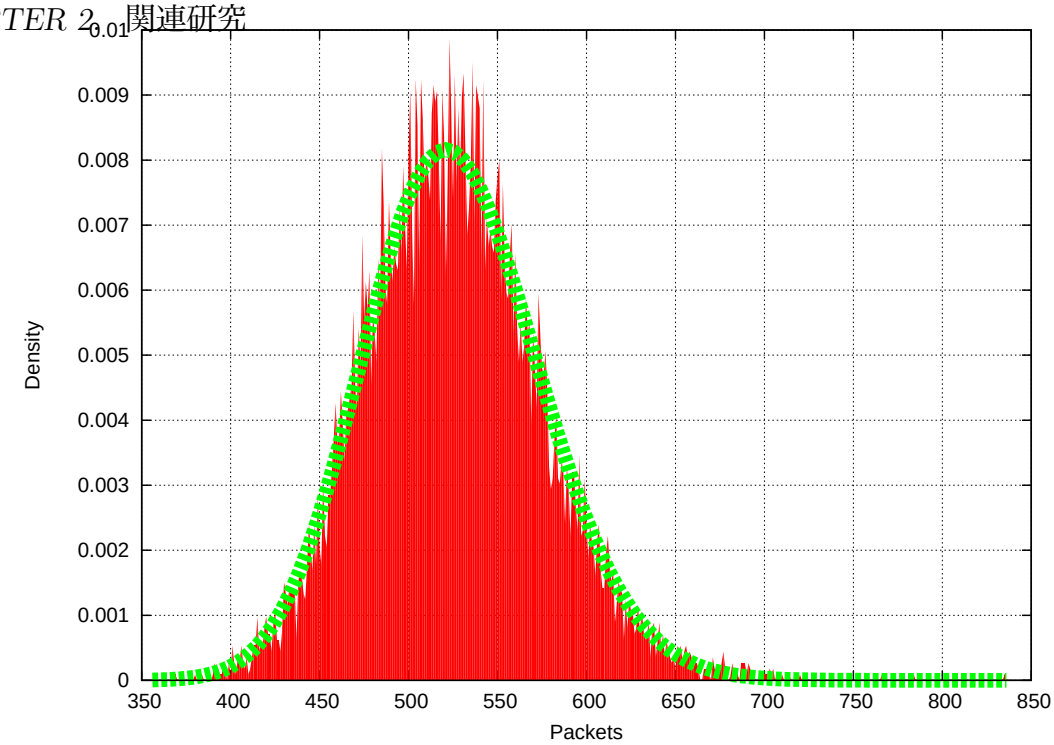

Figure 2.4: パケット到着数の分布

分布を表すものとする．

$$\Gamma_{\alpha,\beta}(x) = \frac{1}{\beta\Gamma(\alpha)}(\frac{x}{\beta})^{\alpha-1}\exp(-\frac{x}{\beta}) \tag{2.14}$$

図 2.3 でのパケット到着数分布を確率密度分布にし，このモデルとの比較を行ったものを図 2.4 に示す．α はこの分布を決定するパラメータでありまた β は分布大きさを決定するパラメータである．このため，前者を shape parameter，後者を scale parameter と呼ぶ．この二つのパラメータは，モーメント法により確率変数の平均 μ と 分散 σ^2 を用いて，

$$\alpha = \frac{\mu^2}{\sigma^2}, \beta = \frac{\sigma^2}{\mu} \tag{2.15}$$

と簡便に求めることができる．

Chapter 3

パケットサンプリングによるマルチスケールガンマモデルの影響評価

3.1 パケットサンプリング手法

本評価では 3 つのパケットサンプリング手法を用いている．用いたサンプリング手法についてそれぞれ説明する．

Random Packet Sampling Random Packet Sampling は，各パケットをランダムに確率 $p(0 < p < 1)$ でサンプリングする．

Stratified Random Sampling Stratified Random Sampling は，N 個毎の到着パケットの中からランダムに一個を取り出す．

Random Host Sampling Random Host Sampling は，ある特定の送信元 (または宛先)IP アドレスを持つパケットを取り出す．この方法は，Random Projection(Sketch) と呼ばれ，異常検出手法でトラフィックを分割するのに使われている．この手法では，IP アドレスを Q 個に分けてそのうちの一つを用いている．

以下，サンプリングレートは S で表すものとし，ここで紹介したサンプリング手法はそれぞれ， $S = p = \frac{1}{N} = \frac{1}{Q}$ という関係に対応するものとする．

3.2 トラフィックデータ

これら 3 つのパケットサンプリング手法を適用したトラフィックデータを用いて，このモデルの記述パラメータ α, β の変化の評価を行った．用いたトラフィックデータは日米間の基幹回線のトラフィックデータである MAWI Working Group Traffic Archive [4] を用いた．このうち，Sampling-point F では，14 時から 14 時 15 分の間の 15 分間のトレースを記録しており，今回の実験では特に 2006 年 3 月 1 日から 7 日のデータを用いて評価を行った．

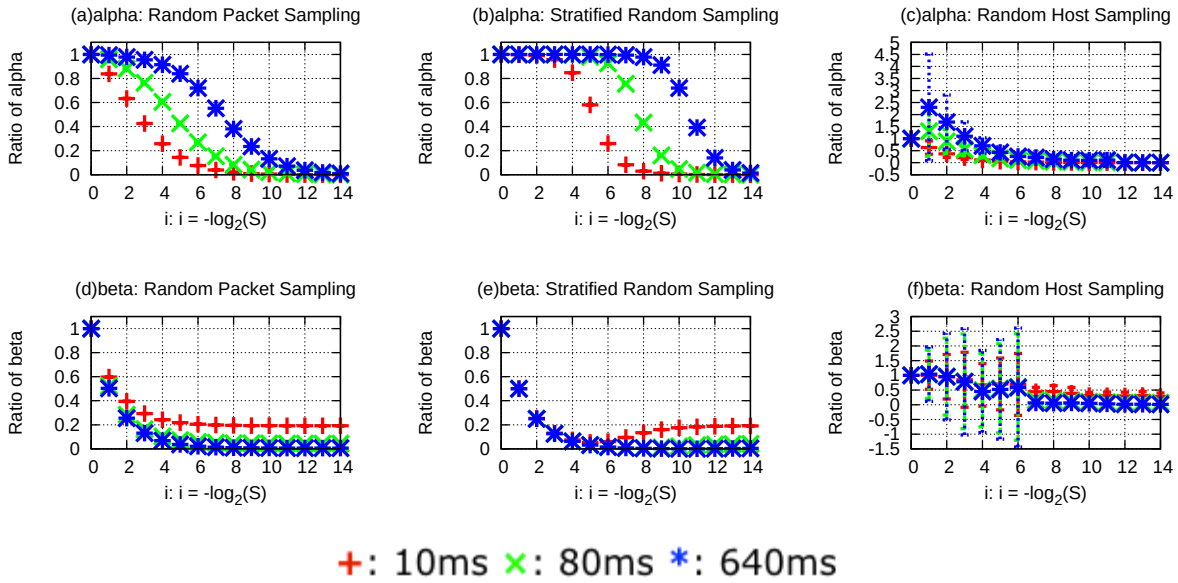

Figure 3.1: α and β change versus sampling ratio

3.3 初期評価結果

3.3.1 α と β の評価

α, β をサンプリングレートの関数 $\alpha_{\Delta_j}(S_i)$ と $\beta_{\Delta_j}(S_i)$ として評価を行った．また， Δ_j は集約時間を表し， $\Delta_j = 5 \times 2^j$ms とする．これをサンプリングが適用されていないトラフィックと比較を行うために，オリジナルのパラメータを基準として評価を行った．つまり，$\alpha_{\Delta_j}(S_i)/\alpha_{\Delta_j}(1), \beta_{\Delta_j}(S_i)/\beta_{\Delta_j}(1)$ をプロットした．図 3.1 にその結果を示す．プロットした集約時間は 10, 80, 640ms であり，これらは $j = 0, 3, 6$ に相当する．(a)(b)(c) は α の結果であり，(d)(e)(f) は β の結果である．(a)(d) は Random Packet Sampling の結果であり，(b)(e) は Stratified Random Sampling，(c)(f) は Random Host Sampling の結果である．

図 3.1(b) は α の値は図の左側では維持されているが，反対側では 0 に収束している．この減少が始まるサンプリングレートは，集約時間ごとに異なっている．この値が 0.8 となるのは $j = 0, 3, 6$ のときそれぞれ $i = 5, 7, 10$ である．Stratified Random Sampling と比べると，Random Packet Sampling では少し違った結果が現れている．図 3.1(a) では，サンプリングレートが 1/2 の時から，どの集約時間の場合でも減少を始めているのが観察できる．つまり， α はフラットな部分，つまり定数を維持する部分が存在しない．α の比が 0.8 を下回るのは $j = 0, 3, 6$ の時それぞれ $i = 2, 3, 6$ の時である．これは，Random Packet Sampling の時に α の値が急速に小さくなっていることを示している．Random Host Sampling の場合も，Random Packet Sampling と同じような傾向を示している．つまり，α の値が定常的

になっている部分がなく，0 に収束している．また，高い分散を示している．

図 3.1(d) は，β がある定数に収束してそこからは定数となっていることが分かる．収束するサンプリングレート及び収束する値は Δ ごとに異なる．図 3.1(e) は β が極小値を取ることを示している．この極小値は 図 3.1(b) で α が減少し始めたサンプリングレートに等しい．また，この β が収束する値は，(d) で β が収束した値と等しい．図 3.1(f) は Random Host Sampling での β の値を示している．これは，大きい分散を取りながら振動しているが，サンプリングレートが 1/128 より小さい時 (i が 7 より大きい時) に収束する．

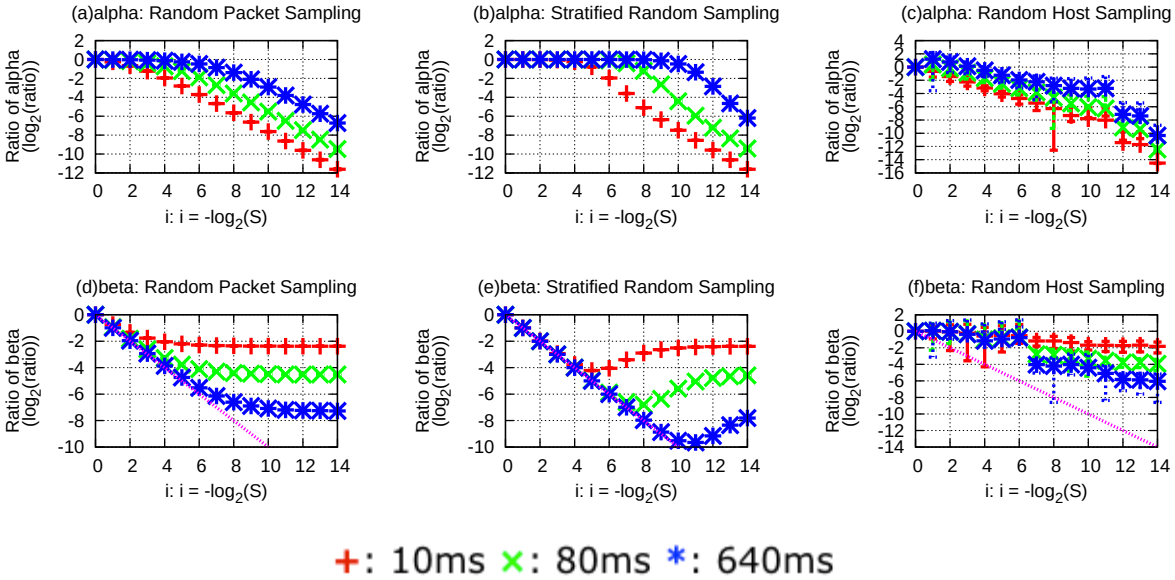

Figure 3.2: logscale α and β change versus sampling ratio

3.4 α と β の評価 (logscale)

図 3.2 は サンプリングレートに対する α と β の変化をプロットしたものだが，Y 軸が対数軸になっている．つまり， $\log_2(\alpha_{\Delta_j}^{(S_i)}/\alpha_{\Delta_j}^{(1)})$ と $\log_2(\beta_{\Delta_j}^{(S_i)}/\beta_{\Delta_j}^{(1)})$ を表している．図 3.2(d)(e)(f) に描かれている直線は $\beta_{\Delta_j}(S_i) = S_i\beta_{\Delta_j}(1)$ である．図 3.2(b) は Stratified Random Sampling での α の値が定数部分と線形変化部分に分かれることを示している．これと比べると，図 3.2(a) は値の変化が 図の左端と右端では同じような動きをしているが，定数部分と線形変化部分が連続的に変化していることが分かる．図 3.2(c) は Random Host Sampling での α の動きを示している．これは，$i = 11$ まである値に収束していき，その後階段状に値が減少している．

図 3.2(d) では，Random Packet Sampling での β の値にも定数部分と線形部分があることが分かった．図 3.1 (e) は違った傾向の曲線を描いている．最初は $\beta_{\Delta_j}(S_i) = S_i\beta_{\Delta_j}(1)$ に従い，最終的には定数に収束している．図 3.2(f) は Random Host Sampling の β の値を示している．

3.5 集約時間による正規化

図 3.3 は それぞれの集約時間ごとの値の変化のパターンが同じであることを示した．この仮定の検証のため，集約時間でサンプリングレートを正規化した図を 図 3.3 に示した．Random Packet Sampling と Stratified Random Sampling における β の値をプロットした．X 軸は $i-j=-\log_2(S_i\frac{\Delta_j}{\Delta_0})$ を表し，Y 軸は $\log_2\frac{\beta_{\Delta_j}(S_i)}{\beta_{\Delta_j}(1)}+j=\log_2\frac{\beta_{\Delta_j}(S_i)}{\beta_{\Delta_j}(1)}\frac{\Delta_j}{\Delta_0}$ を表わしている．この結果，集約時間で正規化した場合 β の値はひと通りに描けることが明らかになった．ただし図の右側では，それぞれの曲線について隙間があることがわかる．これは，$\beta_{\Delta_j}(1)$ による影響であり， $\beta_{\Delta_j}(S_i)$ のせいではない．

$\beta_{\Delta_j}(S)$ の収束値を調べると，全ての Δ について $\lim_{S\to 0}\beta_{\Delta_j}(S)=1$ となっている．従って $\frac{\beta_{\Delta_j}(S_i)}{\beta_{\Delta_j}(1)}\frac{\Delta_j}{\Delta_0}=\frac{1}{\beta_{\Delta_j}}\frac{\Delta_j}{\Delta_0}(S\to 0)$．もし，それぞれの曲線がぴったり重なるなら，$\beta_{\Delta_j}(1)\propto\Delta_j$ である．しかし，図 3.3 はそうでないことを示している．つまり， $\beta_{\Delta_j}(1)\not\propto\Delta_j$ である．

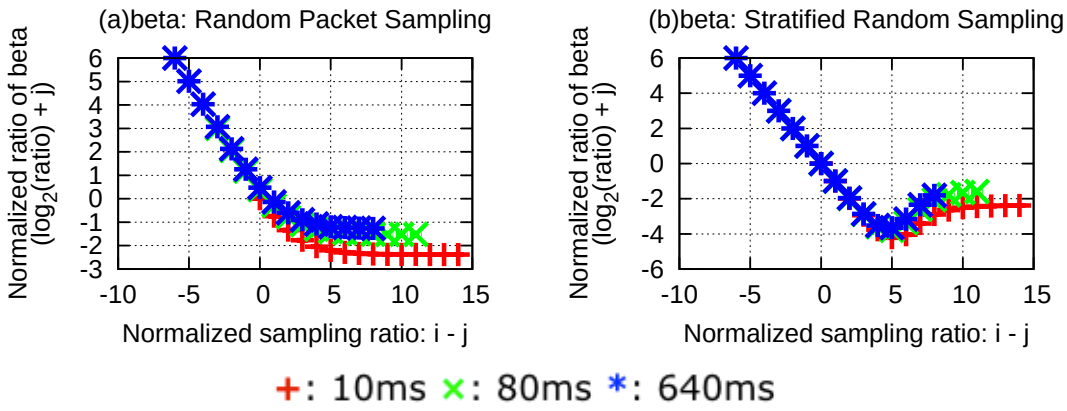

Figure 3.3: logscale α and β change versus normalized sampling ratio

3.6 初期評価のまとめと課題

本研究では，パケットサンプリングが異常検知手法に及ぼす影響に着目し，異常検出手法が用いるトラフィックのメトリックについて，パケットサンプリングがどのような影響を与えるかの評価を行った．評価に用いたサンプリング手法は Random Packet Sampling, Stratified Random Sampling, Random Projection(Sketch) の 3 つの手法を用い，結果を得た．この結果として，Random Packet Sampling 及び Stratified Random Sampling について，あるサンプリングレートを境にしてパラメータの挙動が変わることが明らかになった．また，このパラメータ挙動の変化が観測できる境界付近において，パラメータの振る舞いが異なっているということが分かった．

今後の課題として以下の点を挙げる．

- 初期評価で得られた結果の一般性の確認

- Random Projection(Sketch) におけるパラメータの挙動についての考察
- それぞれのサンプリングごとのパラメータの振る舞いの違いの原因調査
- パラメータ変化の詳細なモデル構築

このうち，現在初期評価で得られた結果の一般性の確認について取り組んでいる．次節ではこの取組について現在の進捗をまとめる．

3.7 初期評価結果の一般性の確認

初期評価結果の一般性を確認するためには，任意のトラフィックから得られる時系列 $X_\Delta(k)$ に対して同様の評価を行う必要がある．しかし，サンプリングが行われていない広帯域のトラフィックデータの数は限られている．そこで，実際のトラフィックをモデル化した時系列 $X_\Delta(k)$ を生成するために Cluster Point Process(CPP) [10] を用いて評価を行うこととする．

Chapter 4

Cluster Point Process

4.1 Cluster Point Processの概要

ここではCPPモデルについての解説を行う．このモデルは，[10] で提案されている統計的モデルであり，インターネットバックボーンのような広帯域リンクでのTCPフローのモデルである．このモデルを提案する前提として，[10] では現実に計測されたトラフィックの以下の特徴を指摘している．本章では，まずCluster Point Processの概要について説明し，次に既存のパラメータチューニング手法について説明する．最後に，Cluster Point Processの実装が公開されていないため，行った実装の詳細について説明を行う．

- IPレベルでのスケーリングはフローレベルでのスケーリングを説明しない．
- 違うフロー同士でのパケット到着過程の独立性は高い．
- 小さいタイムスケールでの振る舞いは各フローのパケット到着パターンに原因がある．
- LRDはフローに含まれるパケット数がheavy tailであることに原因があり，フローの中のパケット到着分布にはその原因はない．

以上の観察により，CPPは各フローの到着過程，フローの大きさ，各フローでのパケット到着過程を独立とみなし，次のように定義した．

- フローの到着過程はポワソン過程に従う
- フローに含まれるパケットの数はベキ分布に従う
- フロー内でのパケット到着の到着間隔はガンマ分布に従い，異なるフロー間のパケット到着過程は独立である．
- それぞれの確率過程は独立である

ここで，それぞれの確率過程に関して詳しく見ていく．

確率過程を記述する前に，ここで用いられる記号の意味について表4.1で説明する．

パラメータ	説明
$t_F(i)$	i 番目のフローの到着時刻
N	トレース中に含まれるフローの総数
$P(i)$	i 番目のフローに含まれるパケット数 (=フローサイズ)
$D(i)$	i 番目のフローの持続時間
$A_i(l)$	i 番目のフローの，l 番目のパケット到着間隔

Table 4.1: 記号の表記

[10] での実験結果より，フローの到着過程はそれぞれが独立したポワソン過程と見えることが分かっている．これにより，パケット到着過程 $X(t)$ を以下のように表すことができる．

$$X(t) = \sum_i g_i(t - t_F(i)) \tag{4.1}$$

ただし，i はフローのインデックス，$t_F(i)$ はフロー到着時刻，g_i は i 番目のフロー内でのパケット到着過程を表わしている．$t_F(i)$ は母数 λ_F のポアソン過程に従う．また，g_i は次のように表せる．

$$g_i(t) = \sum_{j=1}^{P(i)} \delta(t - \sum_{l=1}^{j-1} A_i(l)) \tag{4.2}$$

ここで，$P(i)$ は i 番目のフローに含まれるパケットの数，$A_i(l)$ は i 番目のフローにおける l 番目と $l+1$ 番目のパケットの間の到着間隔を表す．P の確率分布関数は，P が heavy tail distribution であることにより次のように表せる．F_P を P の確率分布関数とすると，

$$1 - F_P(j) \sim Lj^{-\beta}(j \to \infty, \beta \in (1,2)) \tag{4.3}$$

ただし，$E[P] \equiv \mu_P$ である．また，パケットの到着間隔である A はガンマ分布に従うのが妥当であると [?] より結論づけられている．ガンマ分布は二つのパラメータがあり，ここでは shape parameter c と scale parameter b を用いるとする．これら 5 つのパラメータ $\lambda_F, \beta, \mu_\beta, b, c$ がこのモデルに必要なパラメータである．

これら，チューニングを行う必要があるパラメータを表 4.2 にまとめる．

パラメータ	説明
λ_F	トレース中のフローの到着密度 [flow/s]
β	フローサイズの分布を決定するパラメータ
L	フローサイズの分布を決定するパラメータ
λ_A	個々のフロー中のパケット到着密度 [pkt/(flow・s)]
c	パケットの到着間隔の分布形状を決定するパラメータ

Table 4.2: Cluster Point Procee でチューニングが必要なパラメータセット

4.2 CPP におけるパラメータチューニング手法

ここでは CPP モデルに必要なパラメータの計算方法について述べる．

λ_F　フローの到着分布はポアソン仮定に従うため，記述パラメータが一つ必要である．ポアソン仮定の一般式は下記のようになる．ただし，N_t は時刻 t までに発生した事象の数，λ は単位時間あたりの平均発生回数である．

$$P(N_t = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!} \tag{4.4}$$

必要な記述パラメータは単位時間あたりの平均発生回数で，これを λ_F とする．λ_F は，対象トラフィックのフロー到着間隔の平均の逆数により計算される．つまり，

$$\lambda_F = \frac{1}{\frac{\sum_{i=0}^{N-1} t_F(i+1)-t_F(i)}{N-1}} = \frac{N-1}{\sum_{i=0}^{N-1} t_F(i+1) - t_F(i)} \tag{4.5}$$

ただし N はフロー数を表す．

(λ_A, c)　フロー内のパケット到着分布はガンマ分布に従うため，記述パラメータは二つ必要である．これを λ_A, c とする．λ_A は フローの単位時間あたりパケット数 $R(i)$ によって計算される．$R(i)$ は，フローに含まれるパケット数 $P(i)$ と フロー持続時間 $D(i)$ によって，$R(i) = \frac{P(i)}{D(i)}$ と書ける．λ_A は各フローの $R(i)$ をフローに含まれるパケット到着間隔の個数 $P(i) - 1$ で重み付けした平均を計算することで得られる．これは次のような式で書ける．

$$\lambda_A = \sum_{i=1}^{N} \frac{R(i)}{N}(P(i) - 1) = \sum_{i=1}^{N} \frac{P(i)(P(i) - 1)}{ND(i)} \tag{4.6}$$

ただし，$P(i) \geq 2$ の場合のみに限る．また，c も同様に

$$c = \sum_{i=1}^{N} \frac{c(i)}{N}(P(i) - 1) = \sum_{i=1}^{N} \frac{(P(i) - 1)\mu_A^2(i)}{N\sigma_A^2(i)} \tag{4.7}$$

として求められる．ただし，$\mu_A(i)$ は i 番目フローの平均パケット到着間隔を表し，$\sigma_A^2(i)$ はその分散を表している．また b は，$b = \frac{1}{c\lambda_A}$ で与えられる．

(μ_P, L, β)　フローサイズの分布を決めるのに必要なパラメータを計算する．μ_P は平均フローサイズなので， $\mu_P = E[P]$ である．(L, β) に関しては，$1 - F_P(j)$ の log-log のプロットにおいて最小二乗法を用いることによって推定を行う．すなわち，$x = \log j, y = \log Lj^{-\beta}$ とする．このとき，$y = \log L - \beta x$ となるので，この直線と実際の観測点で推定を行う．ただし，推定を行う区間は $j \geq 6$ とする．

ここで，離散パレート分布のような変数 H は以下のような確率分布関数を持つ．

$$F_H(k; a, \beta) = 1 - (ak + 1)^{-\beta} \sim 1 - Lk^{-\beta} \tag{4.8}$$

ここで，$a = L^{-1/\beta}$ である．この分布は平均 $E[H] = a^{-\beta}\zeta(\beta, 1/a)$ を持つ (ただし $\beta > 1$)．ここでの問題は，先程の最小二乗法で求めた (L, β) では，$E[H] = \mu_P$ を満たさない場合があり得るという点である．そこで，この確率分布を以下の式のように拡張する．

$$F_P(k; p, a, \beta) = pF_H(k; a_2, \gamma) + (1-p)F_H(k; a, \beta) \tag{4.9}$$

ただし，(γ, a_2) はそれぞれ $\gamma > 2, a_2 > 0$ を満たす定数である．ここでは $\gamma = 3, a_2 = 0.005$ として定める．また，p は mixture parameter であり $p \in [0, 1]$ である．これで，$E[P] = \mu_P$ を満たすような p を求める．すなわち，

$$p = \frac{\mu_P - a^{-\beta}\zeta(\beta, \frac{1}{a})}{a_2^{-\gamma}\zeta(\gamma, \frac{1}{a_2}) - a^{-\beta}\zeta(\beta, \frac{1}{a})} \tag{4.10}$$

として求めることができる．

このチューニング手法を用いて，Cluster Point Process を 24 時間のトラフィックトレースに適用した結果が図 4.1 である．

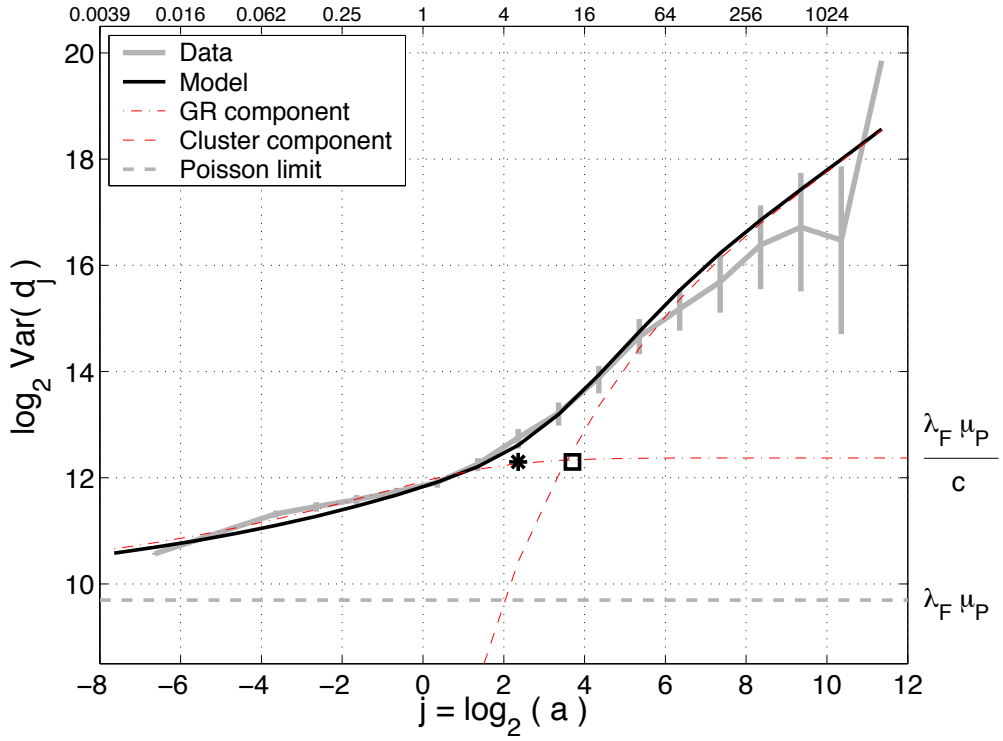

Figure 4.1: 再生成したトラフィックの比較 (Logscale Diagram)

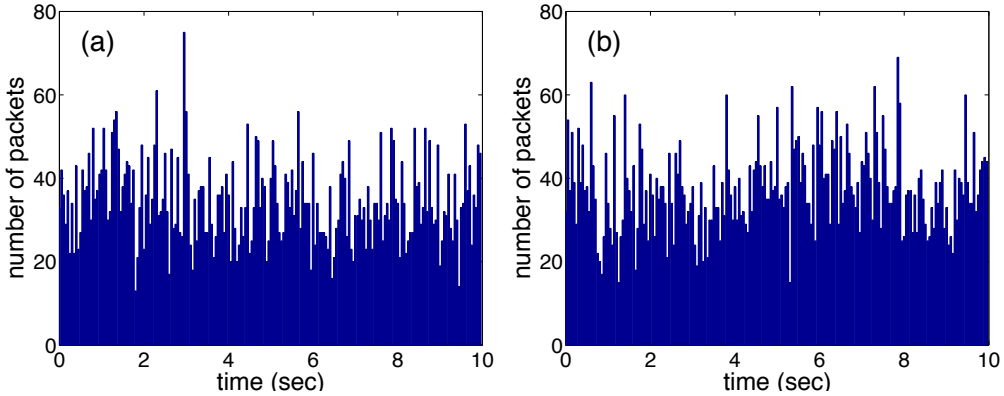

Figure 4.2: 再生成したトラフィックの比較 (50ms ごとのパケット到着数)

4.3 実装

Cluster Point Process の実装は公開されていないため，評価にあたって実装を行った．パケット到着の生成は，以下の 3 段階に分けて計算される．

1. フローの到着時間の生成
2. フローのサイズの生成
3. 各フローのパケットの到着時間の生成

この概要を図 4.3 及び図 4.4 に示す．図 4.3 では，フローの到着時間および各フローに含まれるサイズが生成される．実際生成された例を表 4.3 に示す．このように，実際のトラフィック再生成においては，まずフローの到着と，そのフローに含まれるパケットのサイズが与えられ，この情報がファイルに記録される．

次の段階として，各フローの到着時間とパケット数から，各フロー内のパケット到着を計算する．この様子を図 4.4 に表す．実装した処理では，トレース長よりも遅くに到着するパケットは全て切り捨てられるため，Step2 で生成したフローサイズ分布と，最終的に得られるフローサイズ分布は異なる．最終的に得られる出力の例を表 4.4 に示す．この，Step2 の段階で得られるフローサイズの分布を Optimistic Flowsize Distribution，最終的に得られるフローサイズの分布を Effective Flowsize Distribution と呼ぶこととする．

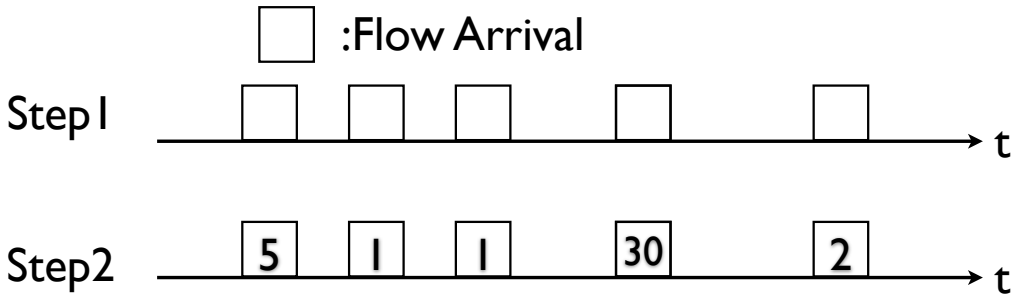

Figure 4.3: Cluster Point Process の実装概要：Step1,2

Flow ID	# of Pkts	Arrival Time
...	...	...
392744	5	120.002136
392745	4	120.002869
392746	5	120.008484
392747	22	120.009155
392748	17	120.010742
392749	20	120.016724
...	...	...

Table 4.3: フローの到着時刻とフローサイズの例

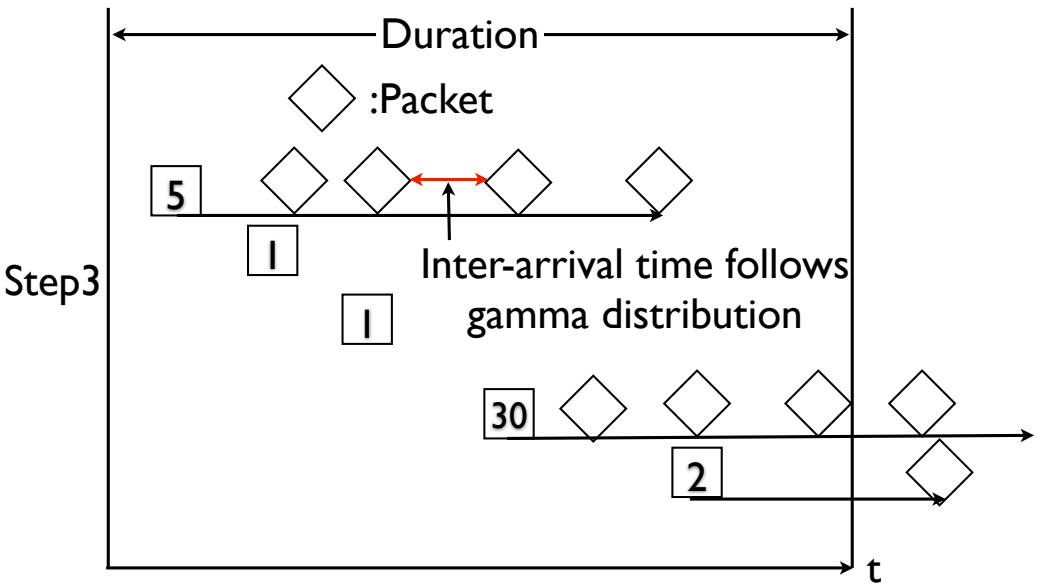

Figure 4.4: Cluster Point Process の実装概要：Step3

Packet Arrival Time	Flow ID
...	...
120.000058177	392742
120.000120227	373015
120.000147929	389340
120.000246994	323489
120.00027283	388863
120.000295668	392737
120.000369636	392737
120.000442273	389012
120.000500302	389012
120.000545772	387446
120.000657541	392716
120.00068182	392672
120.000702742	392712
...	...

Table 4.4: パケットの到着時刻とフロー ID

Chapter 5

提案手法

Cluster Point Process を 900 秒間のトレースに適用するにあたって，二つの問題があることを指摘する．この問題を解消するため，それぞれについてパラメータを正しくチューニングする方法を提案する．

5.1 短いダンプデータに対する適用の際の問題点

実際に適用した結果を図 5.1 に示す．この図より，Cluster Point Process によって再生成されたトラフィックの特性が，もとのトラフィックの特性と大きく異なっていることが分かる．

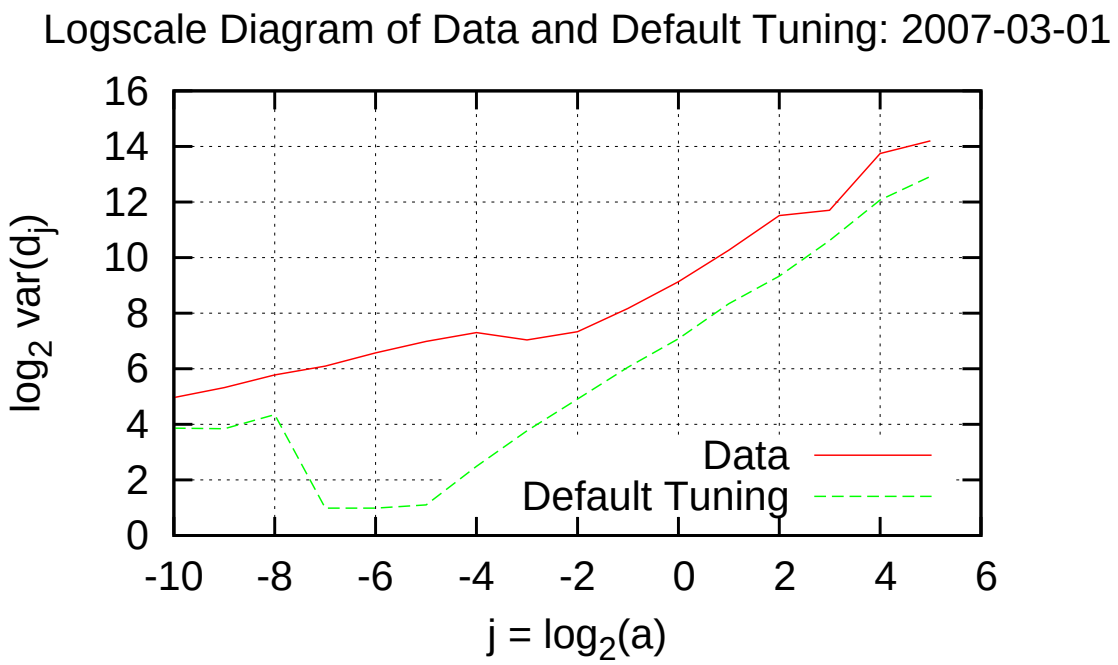

Figure 5.1: Cluster Point Process によるトラフィック再生成の失敗：Logscale Diagram

λ_F	λ_A	c	β	L
674.941490	91.530543	24414991	0.958169	0.778291

Table 5.1: 推定されたパラメータセット

図 に，この時のフローサイズの分布を示した．

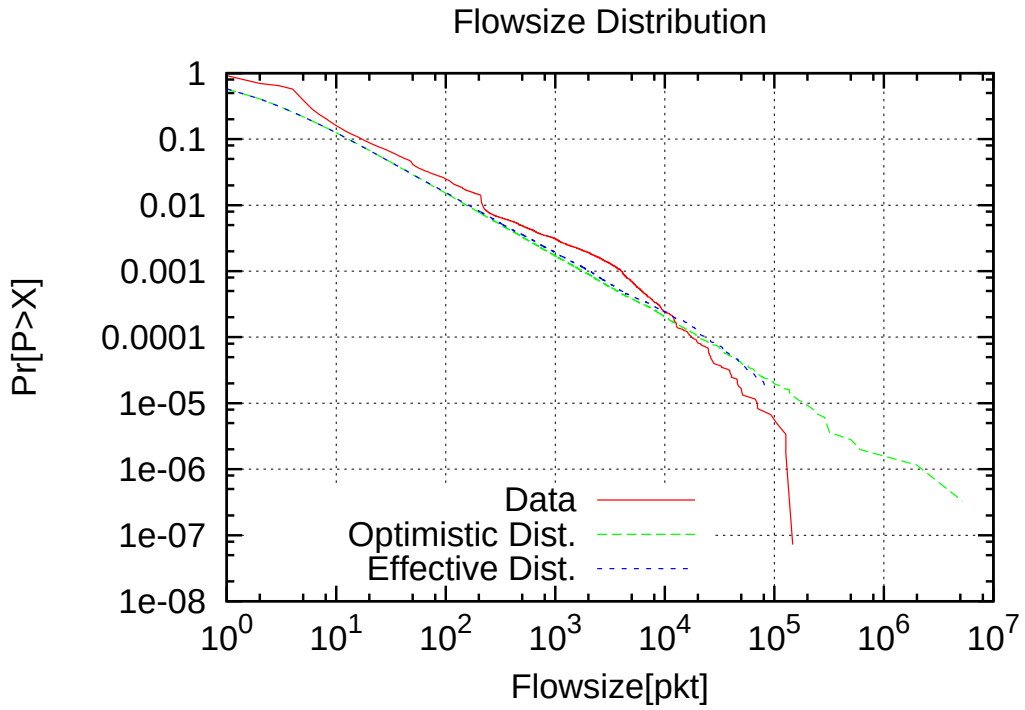

Figure 5.2: Cluster Point Process によるトラフィック再生成の失敗：フローサイズの分布

また，この時得られたパラメータセットを表 5.1 に示す．これを見ると，c の値が極端に大きくなっていることと，$1 < \beta < 2$ になっているべき β の値が 1 より小さくなっていることが分かる．

トラフィックの再生成がうまくいかない原因として，以下の二点を指摘した．

- β の値が小さくなりすぎる ($\beta < 1$)
- c の値が大きくなりすぎる

このようなパラメータ推定が起こってしまう原因として，

- フローサイズの分布が観測時間の制限により大きなフローサイズの部分で折れてしまうこと
- ping などのアプリケーションにより，一定間隔でパケットが送信されるフローが存在し，それによってフローごとの推定量 c_i が極端に大きな値を取ることがあり，それらの加重平均である c も影響を受ける．

ということを挙げる．

5.2 パケット到着間隔のパラメータ

c を 加重平均で求めるのではなく，中央値を推定値として用いる．この際，フローサイズによって重みをつけた中央値と，フローサイズを考慮しない一般的な中央値を用いることが考えられる．この二つのパラメータそれぞれについて評価を行う．加重平均による中央値の計算の仕方の概要を図 5.3 に示す．

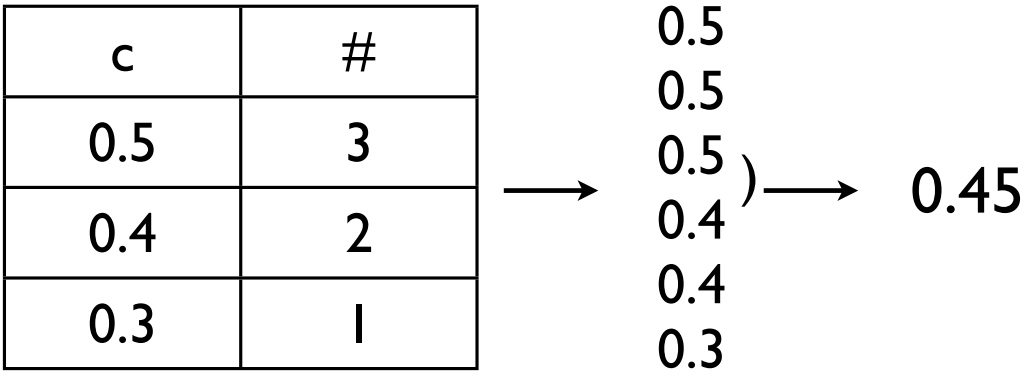

Figure 5.3: 重み付き中央値の計算の仕方

5.3 フローサイズ分布のパラメータ

900 秒間のトレースでは，long-lived なフローが観測時間に収まらないため，フローサイズの分布が元の分布と異なって見えてしまう．しかし，900 秒間のトレースでも LRD を確認することはできるので，Hurst Parameter から推定を試みる．

5.4 その他のヒューリスティクス

図 5.2 の Optimistic と Effictive Distribution において，カットオフされているフローサイズを観察することができる．このフローサイズは 81651 であり，$91.530543 \times 900 = 82377 \sim 81651$ と，λ_A により調整することができる．したがって，一番大きいフローを元データと合わせ

るためには，λ_A の推定値を最大フローサイズ / トレースの持続時間することで実現できると考えられる．

また，トラフィックの再生成を行う際に，0 秒からフローの生成を行なうと到着するパケット数が小さくなると考えられる．そこで，フローの到着を-900 秒から計算することで，0 から 900 秒の間のパケット到着の定常性を保つこととした．

5.5 考えられるパラメータチューニング手法

以上の議論より，適用可能と思われるチューニングの方法を表 5.5 にまとめた．これらのパラメータチューニングによって生成されたトラフィックの特性を評価する．

<table>
<tr><td></td><td>Default</td><td>Tuning1</td><td>Tuning2</td><td>Tuning3</td></tr>
<tr><td>λ_F</td><td colspan="4">総フロー数 / トレース長</td></tr>
<tr><td>β</td><td>フローサイズ分布の線形近似</td><td colspan="3">LD における線形近似</td></tr>
<tr><td>L</td><td>フローサイズ分布の線形近似</td><td colspan="3">β と μ_P に合わせる</td></tr>
<tr><td>λ_A</td><td>パケット到着間隔数重み付き平均</td><td colspan="2">最大フローサイズ / トレース長</td><td>Defualt に同じ</td></tr>
<tr><td>c</td><td>パケット到着間隔数重み付き平均</td><td>中央値</td><td colspan="2">パケット到着間隔数重み付き中央値</td></tr>
</table>

Table 5.2: チューニング方法のまとめ

Chapter 6

評価

6.1 トラフィックデータ

トラフィックデータとして，日米間の学術ネットワークのバックボーン回線である MAWI Traffic Trace を用いる．その中でも特に，Sampling Point F の 2007 年 1 月 1 日から 2007 年 12 月 31 日までの，347 日間のデータを評価に用いた．

6.2 Logscale Diagram での評価

2007 年 3 月 1 日のトレースを，既存のパラメータチューニング手法と提案したチューニング手法でトラフィックの再生成を行った結果を図 に示した．

この評価を，複数の日にまたがって行い，俯瞰的な評価を行うために，各 j に対する成分の二乗誤差を評価の対象パラメータとする．つまり，

$$e = \sum_j (\log_2 \mathrm{Var}(\hat{d}_j) - \log_2 \mathrm{Var}(d_j)) \tag{6.1}$$

のようなメトリックを評価に用いる．ただし，j が小さい部分はフローの内部構造が，大きい部分はフローサイズの分布が寄与しているので，それぞれを分けて評価を行う．すなわち，

$$e_1 = \sum_j (\log_2 \mathrm{Var}(\hat{d}_j) - \log_2 \mathrm{Var}(d_j))^2 \ (j < 0) \tag{6.2}$$

$$e_2 = \sum_j (\log_2 \mathrm{Var}(\hat{d}_j) - \log_2 \mathrm{Var}(d_j))^2 \ (j \geq 0) \tag{6.3}$$

$$e_3 = \sum_j (\log_2 \mathrm{Var}(\hat{d}_j) - \log_2 \mathrm{Var}(d_j))^2 \ (\text{for all } j) \tag{6.4}$$

として，それぞれの評価を行う．

これを，2007 年のトラフィックデータに適用した結果を図 6.2，6.3，6.4 に示す．また，それぞれのチューニングごとに，最小二乗誤差を最大にしたトレース数と最小にしたトレース数を表 7.1 に示す．

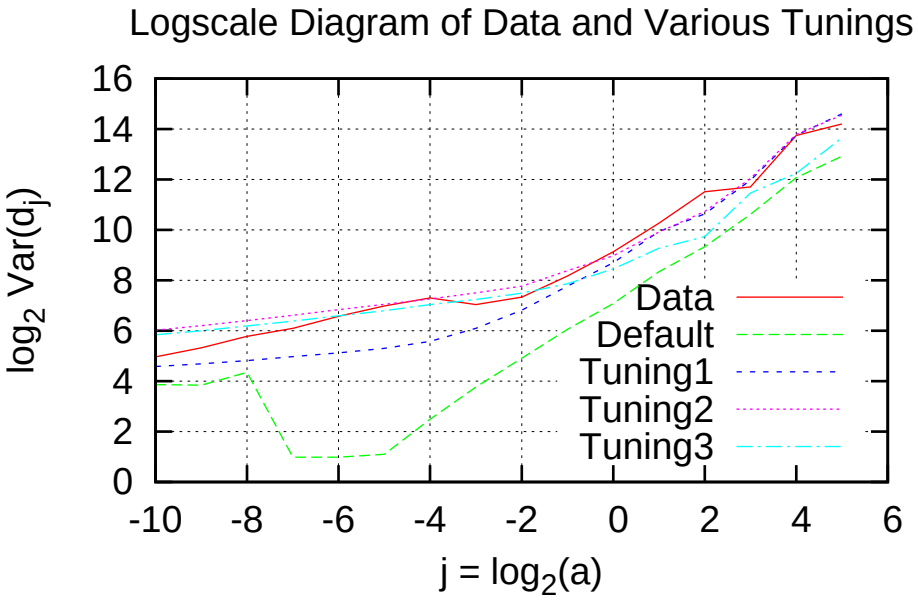

Figure 6.1: 提案チューニング手法で再生成したトラフィックの比較 (Logscale Diagram)

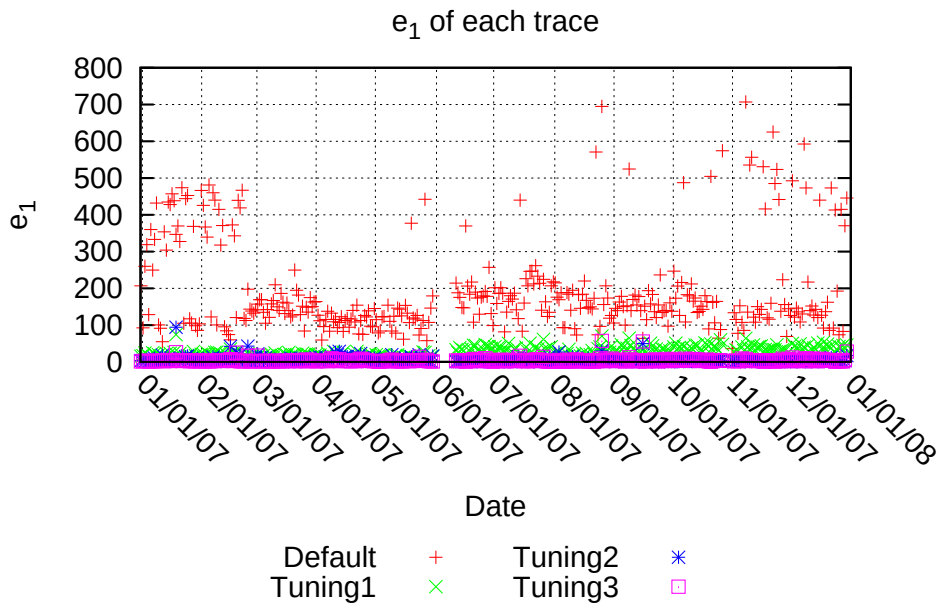

Figure 6.2: e_1 の結果

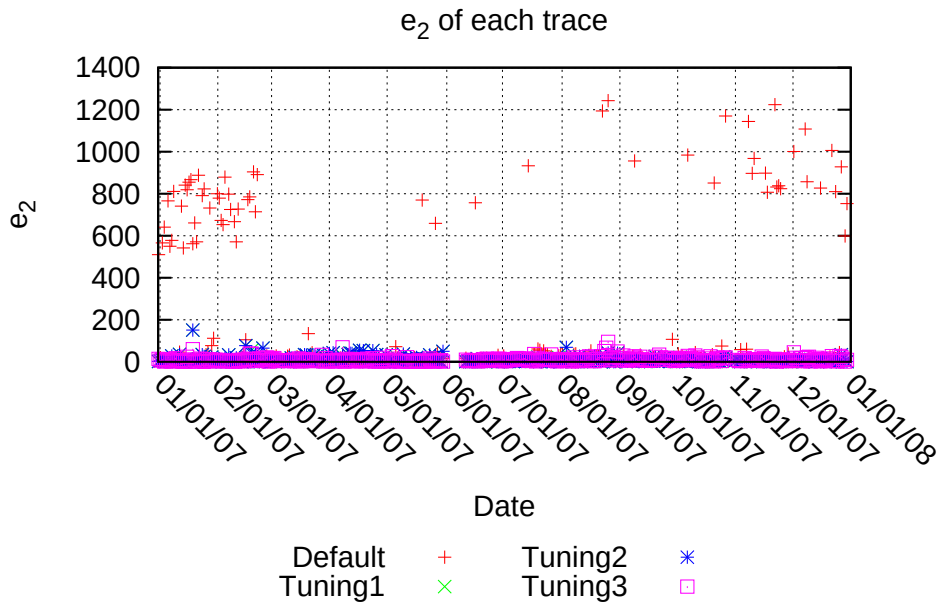

Figure 6.3: e_2 の結果

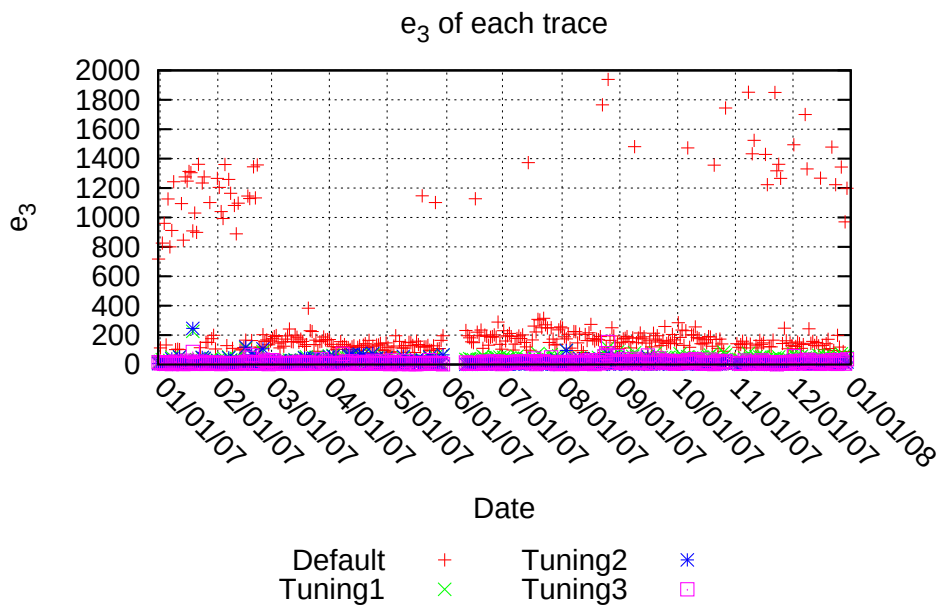

Figure 6.4: e_3 の結果

トレース数	Default	Tuning1	Tuning2	Tuning3
最小 e_1	0	4	187	156
最大 e_1	347	0	0	0
最小 e_2	59	63	132	93
最大 e_2	182	26	42	92
最小 e_3	0	2	222	123
最大 e_3	347	0	0	0

Table 6.1: チューニング方法ごとの，LD の各要素の二乗誤差を最小にしたトレース数

Chapter 7

議論

7.1 最適なチューニング手法

図 6.2 を見ると，フロー内のパケット到着の構造が問題になるような小さいタイムスケールでは，提案された手法全てが元の手法に比べて非常によい成果を上げていることが分かる．さらに，表 の e_1 の比較を見ると，最も良い精度を出したトレース数が最も多いのが Tuning2, その次が Tuning3 となっており，この二つのチューニングのどちらかが最も適したチューニング手法であるということができる．

次に，6.3 に関して見ると，小さいタイムスケールよりも元の手法の誤差が小さい日があることが分かる．また，表 から，e_2 の誤差が最小になるような日は Tuning2 が最も多く，その次が Tuning3 であることは変わらない．しかし，Tuning3 は，e_2 が最大になるようなトレースが 92 個もあることがこの表から分かる．しかし，6.3 より，元のチューニング手法以外の場合は e_2 が顕著に大きい場合は見受けれらないため，Tuning3 を用いても問題ないと結論づける．

これらより，MAWI Traffic Trace の 900 秒間のダンプデータには，チューニング手法 2 が最も適していると結論づけることができる．

7.2 今後の課題

本研究の Future Work として，以下の項目を挙げる．

- 長時間トレースによるチューニング結果比較
- パケットサンプリングを適用した場合の比較

長時間トレースにおけるチューニング結果比較は，チューニング方法 2 が 900 秒以外に適用可能かどうかの議論のために必要である．チューニング手法 2 では，λ_A のチューニングが最大フローサイズ / トレースの長さとなっており，トレースの長さに依存するチューニング方法となっている．このため，トレースの長さが変化したとき，必ずこのチューニング手法が最適かどうかは，更なる議論が必要となる．また，その場合，チューニング手法 3 を用いることができるかどうかも同様に議論の必要がある．

また，パケットサンプリングを適用した場合の比較は，元々Cluster Point Process をトラフィックに適用する理由として，Multi Scale Gamma Model へのパケットサンプリングの影響を評価したいということを挙げていたためである．これを行うことによって，第二章の初期的評価が再生成されたトラフィックでも同様に観測できるか，議論の必要がある．

Chapter 8

まとめ

本研究では，マルチスケールガンマモデルというモデルに着目し，日米間の実トラフィックデータを用いてパケットサンプリングの影響がどのように及ぶのかの調査を行った．その結果，パラメータの動きが大きく二つの状態に分かれることを示した．しかし，実トラフィックには統計的な異常が含まれており，結果の一般化において困難が伴うため，そこで，本研究では Cluster Point Process を用いてトラフィックの再生成を行おうとした．その際，短時間のトラフィックに対して Cluster Point Process をそのまま適用することができないことを発見し，問題が起こる原因の特定と，それを回避するためのパラメータチューニング手法を提案した．評価の結果，提案したパラメータチューニング手法は元の手法より全体的に良い結果を収め，さらにその中で最も手法がどれであるか，結論付けることができた．

謝辞

本論文を執筆するにあたり，大変多くの方々のご支援，ご指導を頂きました．そこで，この場をお借りして謝辞を申し上げたいと思います．

まず，私を研究室に受け入れて頂き，3 年間，研究から生活のことまで面倒を頂いた東京大学大学院 江崎浩 博士に感謝致します．トラフィック研究に携わる先達として，私の研究を，厳しくも真摯に指導して頂いた国立情報学研究所 福田健介 博士に感謝致します．私の一ヶ月間のフランス滞在を支え，及び研究に対する強力な助言を与えてくださった E cole Normale Sup erieure de Lyon の Patrice Abry 博士,Pierre Borgnat 博士に感謝致します．私の研究室生活を支えてくださった江崎研秘書の高橋富美さん，田坂佳苗さん，岩井愛映子さんに感謝致します．トラフィックを行う仲間として議論を行なって頂いた Romain Fontugne 博士，肥村洋輔さん，神田良輝さん, 美嶋　勇太朗くんに感謝致します．研究室の先輩として，研究における指導だけでなく，ネットワークの設定や様々な仕事の進め方を教えて頂いた，土本康生博士, 山本成一博士, 阪本裕介さん藤田祥博士, 金海好彦さん, 土井裕介さん, 落合秀也博士 , 白井俊宏さん，Thomas Silverston 博士, 浅井大史さん, 肥村洋輔さん，下忠健一さん，杉田毅博 さん, 川上雄也 さん Sathita Kaveevivitchai さん,Leela-amornsin Lertluck くん,Luciano Aparicio さんに感謝致します．同期として 3 年間という長い時間を共に過ごした呉和賢くん, 川口紘典くんに感謝致します．研究生として研究室での時間を共に過ごし活動した David Jageberg くん,Jonas Johansson くんに感謝致します．後輩として，一緒に研究室の活動を共に行い苦楽を共にした石橋尚武くん，石田渉くん，園田大剛くん，李聖年くん，林東權くん，正原竜太くん，小坂良太くん，木下僚くん，朴　成軍くん，美嶋　勇太朗くん，東浦　成良くんに感謝致します．最後に，これまでの学生生活を支えてくれた友人，家族に感謝致します．

Bibliography

[1] Abry, P. and Veitch, D.: Wavelet Analysis of Long-Range-Dependent Traffic (1995).

[2] Andersen, A. and Nielsen, B.: A Markovian approach for modeling packet traffic with long-range dependence, *IEEE Journal on Selected Areas in Communications*, Vol. 16, No. 5, pp. 719–732 (1998).

[3] Brauckhoff, D., Tellenbach, B., Wagner, A. and May, M.: Impact of Packet Sampling on Anomaly Detection Metrics, *Time*, pp. 159–164 (2006).

[4] Cho, K., Mitsuya, K. and Kato, A.: Traffic Data Repository at the WIDE Project, *ATEC '00 Proceedings of the annual conference on USENIX Annual Technical Conference* (2000).

[5] Choi, B. and Bhattacharyya, S.: Observations on Cisco Sampled Netflow, *ACM SIGMETRICS Performance Evaluation Review*, Vol. 33, No. 3, pp. 18–23 (2005).

[6] Claise, B., Johnson, E. A. and Quittek, J.: RFC5476: Packet Sampling (PSAMP) Protocol Specifications (2008).

[7] Desaulniers-Soucy, N. and Iuoras, A.: Traffic modeling with universal multifractals, *IEEE Global Telecommunications Conference, GLOBECOM'99*, Vol. 1b, pp. 1058–1065 (1999).

[8] Duffield, N., Lund, C., Thorup, M., Avenue, P. and Park, F.: Properties and Prediction of Flow Statistics from Sampled Packet Streams, pp. 159–171 (2002).

[9] Feldmann, A., Gilbert, C. and Willinger, W.: Data networks as cascades: investigating the multifractal nature of Internet WAN traffic, *ACM SIGCOMM '98 Proceedings of the ACM SIGCOMM '98 conference on Applications, technologies, architectures, and protocols for computer communication* (1998).

[10] Hohn, N., Veitch, D. and Abry, P.: Cluster processes: a natural language for network traffic, *IEEE Transactions on Signal Processing*, Vol. 51, No. 8, pp. 2229–2244 (2003).

[11] Karagiannis, T., Molle, M., Faloutsos, M. and Broido, A.: A Nonstationary Poisson View of Internet Traffic, *IEEE INFOCOM*, Vol. 3, pp. 1558–1569 (2004).

[12] Melamed, B.: An Overview of Tes Processes and Modeling Methodology, *Performance Evaluation of Computer and Communication Systems, Joint Tutorial Papers of Performance '93 and Sigmetrics '93* (1993).

[13] Paxon, V. and Floyd, S.: Wide-area traffic: The failure of Poisson modeling, *IEEE transactions on Networking*, Vol. 3, No. 3, pp. 226–244 (1995).

[14] Paxson, V. and Floyd, S.: Wide-Area Trafc: The Failure of Poisson Modeling (1995).

[15] Polyzos, C.: Application Network of Sampling Traffic Methodologies to Characterization, *SIGCOMM*, pp. 194–203 (1993).

[16] Quittek, J., Zseby, T., Claise, B. and S.Zander: RFC3917: Requirements for IP Flow Information Export(IPFIX) (2004).

[17] Sarvotham, S., Riedi, R. and Baraniuk, R.: Connection-level analysis and modeling of network traffic, Technical report, ECE Dept., Rice Univ. (2001).

[18] Scherrer, A., Larrieu, N., Owezarski, P., Borgnat, P. and Abry, P.: Non-Gaussian and Long Memory Statistical Characterizations for Internet Traffic with Anomalies, *IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING*, Vol. 4, No. 1, pp. 56–70 (2007).

[19] Taqqu, M., Teverovsky, V. and Willinger, W.: Is Network Traffic Self-Similar Or Multifractal?, *Fractals*, Vol. 5, pp. 63–73 (1997).

[20] Veitch, D., Hohn, N. and Abry, P.: Multifractality in TCP/IP Traffic: the Case Against, *Computer Networks: The International Journal of Computer and Telecommunications Networking - Special issue: Long range dependent traffic*, Vol. 48, No. 3, pp. 293–313 (2005).

[21] Zhang, Z., Ribeiro, V., Moon, S. and Diot, C.: Small-time scaling behaviors of Internet backbone traffic: an empirical study, *IEEE INFOCOM*, Vol. 3, pp. 1826–1836 (2003).