

インターネット接続の揺らぎの DFA 解析

柴田章博, 村上直

高エネルギー加速器研究機構 計算科学センター

概要

交通流の流量揺らぎと同様に、インターネット通信の流量においてもべき則に従う揺らぎが、時間レンジの長短関わらずあることが報告されている。しかしながら、揺らぎの起源については理解されていない。インターネット機構に起因する揺らぎと人間や社会の活動などの外的要因に起因する揺らぎとを区別する試みとして、インターネット上のサーバーへの接続要求の時系列データに対する DFA 解析を行った。接続数の時系列データにも流量揺らぎと同様のべき則の揺らぎが見つかった。

Detrend analysis of fluctuation of the Internet connection

Akihiro Shibata, Tadashi Murakami

Computing research Center, High Energy Accelerator Research Organization (KEK)

Abstract

In the flow of the Internet traffic as well as that of the traffic flow of vehicles, the power-low scaling in the traffic fluctuation is observed in both short and long time regions. However, the origin of the fluctuation has not been understood yet. In order to distinguish the fluctuation between the dynamical origin in the Internet traffic and the external one in the social activity, we apply the detrend fluctuation analysis (DFA) to the time series data of the Internet connections, and find the power-low scaling of the fluctuation as same as in the Internet traffic flow.

1 はじめに

インターネット (IP ネットワーク) は現代社会の基本的な手段の通信手段のひとつとなり、大学及び企業などの業務では不可欠のものとなった。従来の主要な通信手段であった電話はその構造や流量の調整を行う集中制御の仕組みの上に築かれたネットワークである。一方、インターネットの通信は集中制御の仕組みを持たない中継機 (ルータ) の接続による自律的な成長と調整機能を持ったネットワークである。IP 通信はパケットと呼ばれる単位分割され、パケットにつけられた送信元・送信先のアドレスに基づいてネットワーク上に配置されたルータを中継して行われる。IP 通信のパケットの流れは、道路網に流入した自動車を単位とする交通流との類似した性質が知られている。両者ともに人工的に作られた経路に形成される点状粒子の自律的な流れであり、また人や社会活動の影響をうけて 1 日及び 1 週間を単位とする周期性を示す。

交通流については、武者・樋口 [1] のパワースペクトラムの解析や只木ら [2] による高速道路の時系列データにおけるトレンド除去法による解析 (Detrend Fluctuation Analysis) において べき則の揺らぎが報告されている。IP 通信については、ルータを通過するパケットの数は待ち行列にあらわれるポアソン分布が信じられてきたが、実際の観測によるとべき則であることが報告された [3]。DFA の解析においてべき則に従う揺らぎがあることが、日周期の影響を避けた短い時間スケールに対しても [4]、数時間に及ぶ長い時間スケールに対しても [5] 発見されている。しかしながら、その起源については一貫した説明が与えられていない。

本研究では、インターネット通信における揺らぎの起源を探るため、従来の研究で観測されているルータにおける流量揺らぎとは異なる量に着目し、IP 通信の接続数の時系列データを DFA 解析する。IP 通信は、電子メールの送受信や Web ページの取得などのように「セッション」を単位として数えることができる。ルー

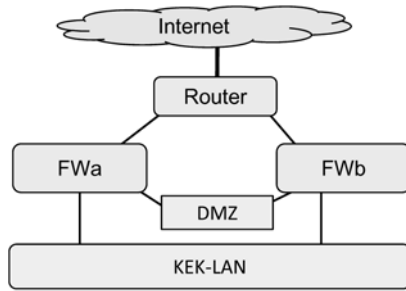


図 1: データ計測のネットワーク接続

タで観測している流量は、セッションの通信データを小分けにしたパケットの時系列データを対象としており、短時間のスケールではインターネット機構の力学的な揺らぎが反映される。一方で、セッション数は通信の最初のパケットの交換を捉えるものであり、インターネット機構の力学的な影響を受けることなく社会の活動に起因する揺らぎを捉えることができると期待される。

2 インターネット通信の時系列分析

本研究で使用する解析のデータは、2008年4月27日(日)～6月28日(土)の9週間間に高エネルギー加速器研究機構(KEK)で採取されたファイアウォールの通信ログを使用する。KEKとインターネットとの通信は、図1に示すようにファイアウォールFWaまたはFWbのいずれかを經由する。通信ログには、接続元及び接続先の(IP番号、ポート番号)、通信種別、送信・受信のデータサイズなどの情報がセッションの開始時刻とその継続時間とともに記録されており、通信を再構成できる。ここでは、揺らぎの相違を比較できるように通信の発信元と発信先の組によって、(a)インターネットからLAN(DMZ)¹への接続(内向きの接続)、(b)LAN(KEK-LAN&DMZ)からインターネットへの接続(外向きの接続)、(c)LAN(KEK-LAN&DMZ)内のセグメントを跨ぐ接続(LAN内通信)、の3種類に区分したデータセットを用意する。

図2は、6月1日(日)～7日(土)の一週間の接続数を秒単位でカウントした時系列データを示す。上段は(a)内向きの接続に対応し下段は(b)外向きの接続に対応する。通信種別の内訳は、Webや電子メールなど情報発信や情報交換関連ではほぼ半数を占めている。

¹KEKのセキュリティ規則によって、インターネットからLANへの直接のアクセスすることは禁止されている。

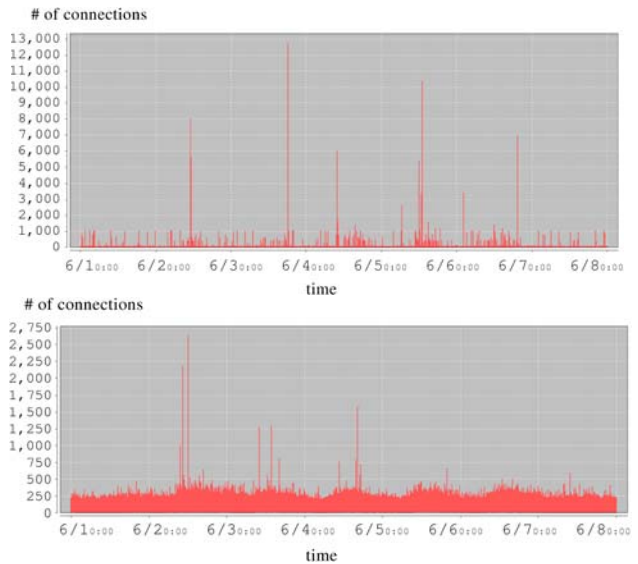


図 2: IP 接続数 (1 秒単位に集計) の時刻推移: 上段はインターネットから KEK-LAN へのセッション数、下段は、LAN からからインターネット への接続数の推移を表す。

2.1 DFA による解析

非定常の時系列解析の手法として流量解析と同様に、DFA (Detrend Fluctuation Analysis) を適用する [6][7]。接続数時系列 $u(t)$ に対して、プロファイル $y(t)$ を定義する。

$$y(t) = \sum_{t=0}^{T-1} [u(t) - \langle u \rangle] \quad (1)$$

ここで、 $\langle u \rangle$ は計測時間 T に対する平均 $\langle u \rangle = (1/T) \sum_{t=0}^{T-1} u(t)$ を表す。区間 T を長さ n の短い区間に分割し、区間 l ごとの $y(t)$ のトレンド $\tilde{y}_l(t)$ を定義する。ここでは $\tilde{y}_l(t)$ は最小二乗法による直線としてフィットする。トレンドを除去したプロファイル (detrend profile) $y_l(t)$ をプロファイルのトレンドからのずれとして定義する。

$$y_n(t) := y(t) - \tilde{y}_l(t), \quad nl \leq t < n(l+1). \quad (2)$$

トレンドを除去したプロファイルの自乗平均

$$F(n)^2 := \frac{1}{T} \sum_{t=0}^{T-1} y_n(t)^2 \quad (3)$$

の区間の長さ l への依存性を調べることで、長時間相関を求めることができる。また、 $F(n)$ が

$$F(n) \sim n^\alpha \quad (4)$$

とべきで振る舞う場合には、パワースペクトラムにべき則揺らぎがあることに対応する [5]:

$$P(k) \sim k^{-\beta}, \quad \beta = 2\alpha - 1.$$

プロファイルに周期性がある場合には、その周期を取り除いた $u'(t) = u(t) - \tilde{u}_{day/week}(t)$ に対して DFA 解析を行う。日周期のトレンドは同時刻のデータを平均したものと定義する ($\tau \bmod T_{day} = 24h$)。

$$\tilde{u}_{day}(\tau) = \frac{1}{D} \sum_{d=1}^D u(d \times T_{day} + \tau). \quad (5)$$

同様に、週周期のトレンドは各週の同時刻のデータを平均したものと定義する ($\tau \bmod T_{week} = 168h$)。

$$\tilde{u}_{week}(\tau) = \frac{1}{W} \sum_{w=1}^W u(w \times T_{week} + \tau). \quad (6)$$

2.2 IP 接続数の揺らぎ解析

図 3 及び図 4 はそれぞれ (a) 内向きの接続と (b) 外向きのインターネット接続に対応するプロファイルを示す。図の赤線は 4/27 0:00 ~ 6/28 24:00 の 63 日間の接続数より作成されたプロファイルを示す。両者の時系列データには日常的な社会活動を反映した一日及び一週間を単位とする周期性が示されている。平均的な 1 日 (式 (5)) 及び平均的な 1 週間 (式 (6)) の接続数を用いて、周期性を取り除いたプロファイルを作成した。図の上段及び下段の青線はそれぞれ平均的な 1 日、及び平均的な 1 週間を引いたプロファイルを示す。

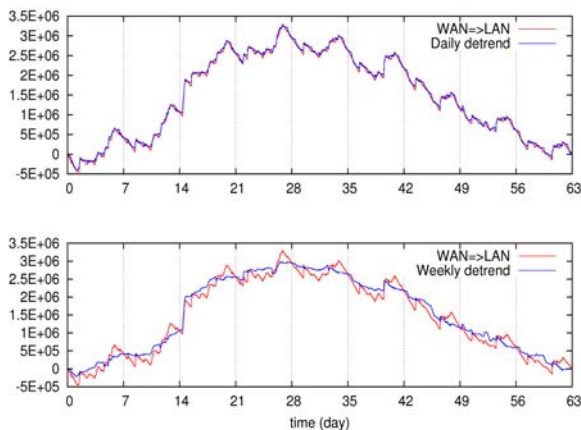


図 3: インターネットから LAN への接続数のプロファイル (4/27 0:00 ~ 6/28 24:00) (赤線)。上段の青線は、平均した 1 日の周期を引いたプロファイル。下段の青線は平均した週の周期を引いたプロファイル。

図 5、図 6 はそれぞれ (a) 内向き及び (b) 外向きの接続数に対応した $F(n)$ の log-log プロットである。赤線は、1 日または 1 週間の周期トレンドを処理する前の $F(n)$ を示しており、三つの区間においてそれぞれよい直線の関係にあり、べき則に従っている。それぞれの上段及び下段は、文献 [2] 及び [5] と同様に周期性のト

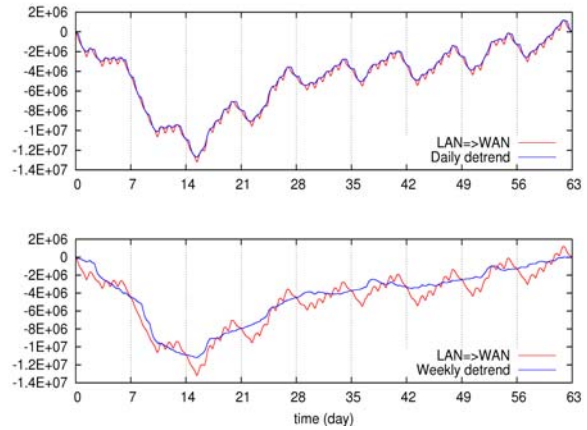


図 4: LAN からインターネットへの接続数のプロファイル (4/27 0:00 ~ 6/28 24:00) (赤線)。上段の青線は平均した 1 日の周期を引いたプロファイル。下段の青線は平均した週の周期を引いたプロファイル。

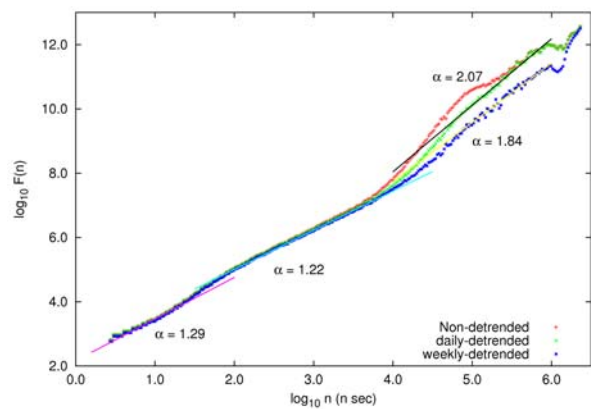


図 5: (a) インターネットから LAN への接続数に対する $F(n)$ の log-log プロット (赤線)。日周期のデイトレンド (青線)、週周期のデイトレンド (緑線)。

レンドを除去した結果を平均的な 1 日のデータを引いた場合 (緑線)、及び平均的な 1 週間のデータを引いた場合 (青線) を示す。

通信区分 (c) LAN 内通信の場合を図 7 に示す。(c) は RTT が無視できる高速なスイッチネットワークの通信のみである。(b) とは発信元や、主要なプロトコルが Web 及びメール関連であり共通であるが、POP/IMAP や DNS の割合が多いなどの差異がある。周期性を取り除く前の $F(n)$ は直線の領域が (a),(b) とほぼ共通するべき則の揺らぎを示す。

短い期間レンジ (~ 100 秒) では、(c) は共通の母体である (b) と異なり、むしろインターネット側からの接続 (a) とほぼ同じ指数を占めず。長時間レンジの揺らぎにおいては、共通の母体である (b),(c) で、日周期を引いた場合において同じ揺らぎのべき則の指数 $\alpha \simeq 1.5$ (ガウスノイズ) を示し、週周期を引いた場合でも (b) $\alpha \simeq 1.3$, (c) $\alpha \simeq 1.5$ でほぼ同じ指数を示す。一方、(a) は (b) とは大きく異なる $\alpha = 1.8 \sim 2.0$ の結果

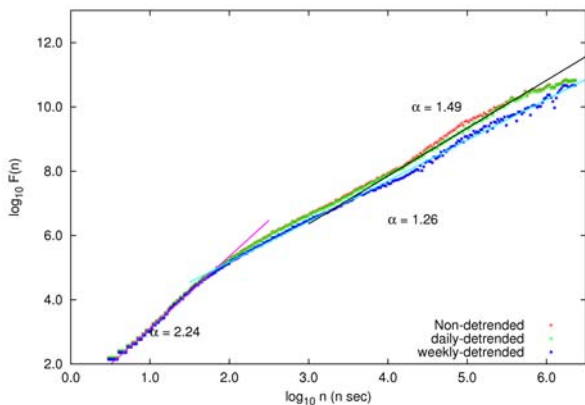


図 6: (b) LAN からインターネットへの接続数に対する $F(n)$ の log-log プロット (赤線)。日周期のディトレンド (青線)、週周期のディトレンド (緑線)。

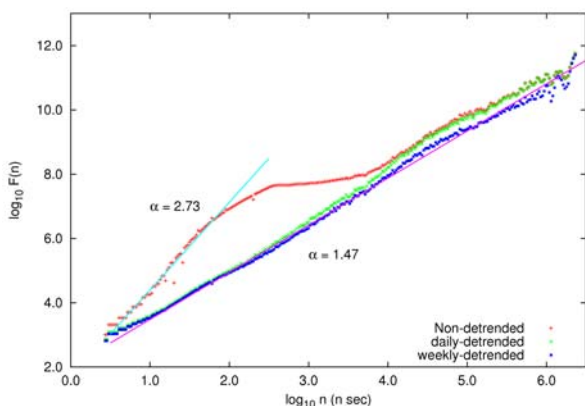


図 7: (c) LAN のサブネット間の通信の $\log F(n)$ v.s $\log n$ プロット。

を得た。しかしながら、図 3 及び図 4 の下段のグラフにあるように、1 週間を単位とする周期性があるが各週の振る舞い大きくばらついており、週周期性の取り除き方の妥当性などさらなる検討が必要である。

3 議論：IP 通信の揺らぎの起源

インターネット通信はその通信の特性から (i) 通信制御に掛かる時間：RTT(Round Trip Time) の数百ミリ秒～数秒、(ii) セッションの通信時間：数秒～数分、(iii) 1 日の活動周期、(iv) 1 週間の活動周期などの時間スケールに分けて考えることができる。(i) は通信の揺らぎが直接反映され、(ii) は通信速度や全パケット送信にかかる時間などとして流量揺らぎに反映される。一方、セッションの開始時刻は、(i) における RTT の揺らぎの影響を受けるが (ii) の時間スケールにおいては独立な事象と考えられる。(iii) 及び (iv) の時間スケールは、流量、セッションともに揺らぎの起源を通信の特性に求めることが難しく、むしろ外的要因であると考えられる。これらの性質は、実際の観測した $F(n)$

のグラフの折れ具合と時間スケールの領域 (ii)～(iv) とが対応していることが示している。

長時間スケールにおいて、流量のべき則揺らぎは、社会的活動などの外的要因に求めることができる。流量揺らぎは、接続数の揺らぎと同じべき則を示しているが、指数は $\alpha = 0.92$ [5] は異なる結果を示しており、その起源については通信の詳細についての分析が必要である。接続数の揺らぎは (a) と (b) 及び (c) において数分 ($n = 10^2$) から数日 ($n = 5 \times 10^4$) の間で一致しており、グループ間の依存がないと思われる。従って、その他の要因、例えば、データ要求量にあるのか、インターネットサービス (プロトコル) による揺らぎの違いはあるのか、などについての検討が必要である。

4 まとめ

IP 通信の接続数の時系列データの DFA 解析によって、べき則の揺らぎを発見した。インターネット機構の影響を受けないセッション数に流量揺らぎと同じべき則の揺らぎが見つかったことは、その揺らぎが社会的な活動起因することを示唆する。しかし、揺らぎの起源の詳細についてはデータ要求量やプロトコル別の解析が必要である。

最後に、インターネット通信の時系列データとして高エネルギー加速器研究機構における通信ログを使用させていただいたことを感謝する。

参考文献

- [1] T.Musha and H.Higuchi, Jpn. J. Appl. Phys. 15 (1976) 1271-1275
- [2] S.Tadaki, M. Kikuchi, A. Nakayama, K. Nishinari, A. Shibata, Y. Sugiyama and S. Yukawa, J. Phys. Soc. Jpn. 75 (2006) 034002
- [3] M.Takayasu, H.Takayasu and T. Sato, Physica A233 (1996) 824-834
- [4] Masao Masugi, IEICE technical report. Information networks Vol.104, No.121 (20040611) 7-12
- [5] Shin-ichi Tadaki, J. Phys. Soc. Jpn. 76 (2007) 044001
- [6] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, Phys. Rev. E 49, 1685 - 1689 (1994)
- [7] C.-K. Peng, Shlomo Havlin, H. Eugene Stanley, Chaos 5, 82 (1995)