インターネット通信のサービス別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

インターネット通信時系列データにおいて冪則に従う揺らぎがあることが報告され, その揺らぎの起源について研究が行われている. 本研究では, インターネットサービス別に DFA 解析し, それぞれの揺らぎのスケール則と共分散を解析し, 揺らぎの構造を分析する。

Detrended fluctuation analysis of the Internet services

Akihiro Shibata, Tadashi Murakami

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

Abstract

We apply the detrended fluctuation analysis (DFA) to the mixed signal the internet services. In order to analyze the structure of long range fluctuation in the internet traffic, we investigate the scaling law of the fluctuation in each service and correlations among them.

1 はじめに

インターネット（IP ネットワーク）は集中制御の仕組みを持たない自律的な成長と調整の機能を持ったネットワークで, その通信の交通流と同様に自己相似性などの動的な性質が報告され, 関心が高まっている. 高速道路の定点観測による交通量の時系列データ [1] やインターネットのルータを通過したデータ [2] において, トレンド除去法による解析（DFA, Detrended Fluctuation Analysis）がなされ, 冪則に従う揺らぎのあることが報告されている. これらデータは, 社会活動の影響を受けており日周期や週周期の周期性が観測され, その周期的なトレンドを除去したデータに対しても冪則が現れることが報告されている. また, インターネット通信における長時間スケールの揺らぎの構造の解析として, 電子メイルの通信ログ解析 [4] やファイアウォール (FW) のログ解析 [5] の報告がある.

本研究では, ルータやファイアウォール（FW）を通過するインターネットサービス（プロトコル）別の通信の DFA 解析を行う. 観測される通信には, 様々なサービスの信号やノイズ起源のことなる信号が混在しているため, 時系列データをサービス毎に分解し対比することで, インターネット通信のべき則揺らぎの起源について検討する.

2 DFA による解析

混合信号に対する揺らぎ成分の分解とその DFA 解析について考察する. $u(t) = \sum_{k=1}^{K} u^{(k)}(t)$ を K 種類の信号 $u^{(k)}(t)$ の ($k = 1,2,\cdots,K$, $t = 0,1,\cdots,T$) の混合とする. 単一信号における DFA 解析 [5] と同様の手続きで, それぞれのデータ $u^{(k)}(t)$ に対応する $y^{(k)}(t)$ 関数を定義する.

$$y(t) := \int_0^t (u(s) - \langle u \rangle)\, ds = \sum\nolimits_{k=1}^{K} y^{(k)}(t), \tag{1}$$

$$y^{(k)}(t) := \int_0^t \left(u^{(k)}(s) - \left\langle u^{(k)} \right\rangle\right) ds, \tag{2}$$

ここで $\left\langle u^{(k)} \right\rangle$ は $u^{(k)}(t)$ の時間 T における平均を表す：$\left\langle u^{(k)} \right\rangle = \frac{1}{T}\int_0^T u^{(k)}(s)ds$. また, 時間 T を M 等分 (区間長 $n = T/M$) し, m 番目の区間において $y^{(k)}(t)$ を p 次元多項式 (ここでは $p = 1$) で χ^2 フィットする. フィットするトレンド関数を $\tilde{y}^{(k)}_{n,m}(t) := \sum_{l=0}^{p} a^{(k)}_{n,m}(l)(t - nm)^l$ とすると, 信号 $u^{(k)}$ に対する F 関数が次のように与えられる.

$$\left(F^{(k)}_m(n)\right)^2 := \frac{1}{n}\int_{(m-1)n}^{mn} \left(y^{(k)}(t) - \tilde{y}^{(k)}_{n,m}(t)\right)^2 dt, \tag{3a}$$

$$\left(F^{(k)}(n)\right)^2 := \frac{1}{M}\sum_{m=1}^{M} \left(F^{(k)}_m(n)\right)^2. \tag{3b}$$

従って, $F^{(k)}(n) \propto n^{\alpha(k)}$ を満たすパラメータとしてスケーリングの指数 $\alpha(k)$ が定義される. また, 混合信号 $u(t)$ に

対する F 関数は次のように計算される.

$$F(n)^2 = \sum_{k=1}^{K} \left(F^{(k)}(n)\right)^2 + 2\sum_{k<k'=1}^{K} R^{(k,k')}(n) \quad (4)$$

ここで, $R^{(k,k')}(n)$ は次で定義される共相関関数 (共分散) である.

$$R_m^{(k,k')}(n) := \frac{1}{n}\int_{(m-1)n}^{mn} dt \left(y^{(k)}(t) - \tilde{y}_{n,m}^{(k)}(t)\right) \times \left(y^{(k')}(t) - \tilde{y}_{n,m}^{(k')}(t)\right) \quad (5a)$$

$$R^{(k,k')}(n) = \frac{1}{M}\sum_{m=1}^{M} R_m^{(k,k')}(n). \quad (5b)$$

ここで, (k,k') の組それぞれの揺らぎの起源が独立であれば,$R^{(k,k')}(n) = 0$ となる. $F(n) \propto n^{\alpha}$ となる α で混合信号に対するスケーリングの指数が計算される. また, $y(t)$ を単一のデータとして DFA 解析を行った $F(n)$ 関数と分解を加味して求められた $F(n)$ は一致する[1].

通信に含まれる周期的トレンドは、文献 [4][5] に従って除去を行う. データ $u(t)$ から周期 T_Q で平均化した関数

$$\tilde{u}_Q(\tau) = \frac{1}{N_Q}\sum_{k=0}^{N_Q-1} u(\tau + kT_Q), (\tau = t \bmod T_Q) \quad (6)$$

によって周期的トレンドが除去された関数を定義する.

$$u_Q(t) = u(t) - \tilde{u}_Q(t \bmod T_Q). \quad (7)$$

3 インターネット通信の時系列分析

インターネット通信のデータとして,KEK のファイアウォール (FW) の通信のログを利用する. KEK のネットワークは, 複数の VLAN に分割された LAN と DMZ 及びインターネットが FW を経由して接続されている (文献 [5] の図 1 参照). それぞれのセグメント間の通信が FW のログとして記録される. 通信のログには, 接続元及び接続先の IP 番号, port 番号 , 通信種別, 送信/受信のデータサイズがセッションの開始時刻と接続時間とともに記録される.

測定対象としたデータを表 1 に示す. 祝祭日を含まない 49 日間のログについて, 接続件数とデータ要求量を DFA 解析の対象とした. TCP/IP の接続要求を行う向きで WAN から LAN (WL), LAN から WAN(LW), LAN 内のセグメント間通信 (LL) の 3 種類のグループに分けた. 脆弱性診断装置のネットワーク管理用通信[2]を除いた, Mail(SMTP), Web(HTTP+HTTPS), DNS の各サービスと全通信 (ALL) について解析する.

4 サービス別揺らぎ解析

図 1 は, 解析で使用したデータ（ALL）に対するデータ要求量に対する $y(t)$ 関数を示す. ここでは, ログデータの

[1] $y(t)$ 及び、それに対応する $\tilde{y}_{n,m}(t)$ は それぞれ $y^{(k)}(t)$ 及び $\tilde{y}_{n,m}^{(k)}(t)$ の線形結合であり、χ^2 の係数も $a_{n,m} = \sum_k a_{n,m}^{(k)}$ と線形の関係にまとめることができる。

[2] DMZ 上の全ホストの脆弱性を、週 1 度診断している

期間	2008/5/12〜2008/6/29 (49 日間)
測定量	接続件数: 対象ログの行数 データ要求量: 送信/受信バイト数の合計 ※脆弱性診断装置は対象外.
ネットワークゾーン	WAN: Internet LAN: DMZ, KEK-LAN(複数の VLAN)
通信方向	WAN→LAN (WL と略す) LAN→WAN (LW と略す) LAN→LAN (LL と略す) ※ LL は、VLAN をまたがる通信のみが対象
サービス	Mail(SMTP), Web(HTTP+HTTPS), DNS ALL(全通信:上記サービス以外も含む)

表 1: データ収集の条件とデータの概要

サンプリングに伴う揺らぎの影響を抑えるため, $y^{(k)}(t)$ 関数にあらわれる積分を台形則による平滑化を行ってサンプル間隔以下の揺らぎを除去した.

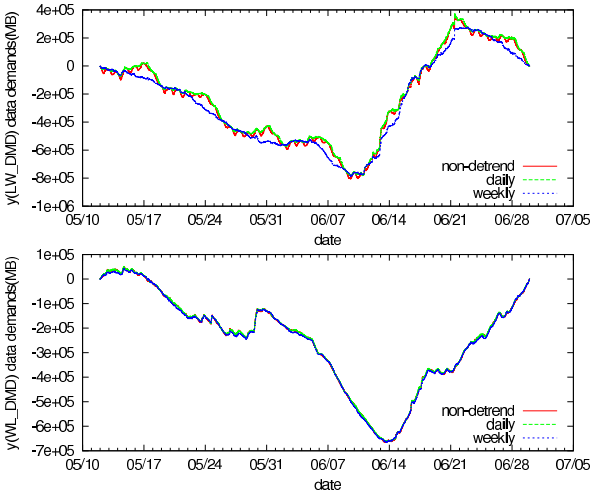

図 1: インターネットと KEK-LAN におけるデータ要求量に対する $y(t)$. raw データから計算されるもの, 平均の 1 日をディトレンドしたもの, 平均一週間をディトレンドしたものを合わせてプロット:(上段) LAN から WAN へのアクセス（LW）.（下段）インターネットから KEK へのアクセス (WL).

Web, Mail, DNS の 3 種類のインターネット・サービスに着目し, それぞれを DFA 解析する. 式（4）に示されるように, 各サービス間には共相関関数 $R^{(i,j)}$ が一般に存在するが, ここでは, サービス間の相関は無視することができるとして, サービスそれぞれのスケール則に着目する. また, Web サーバのアクセスに関しては, HTTP と HTTPS に通信の分解を行いその詳細について検討する.

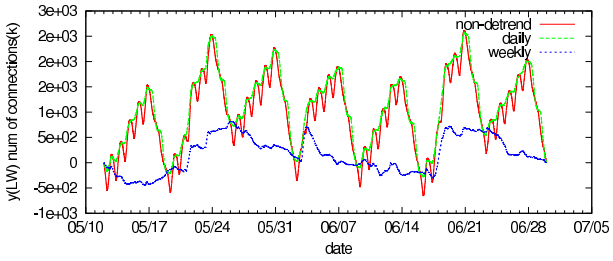

図 2: KEK-LAN からインターネット上の Web サーバへのアクセス数に対する $y(t)$ 関数

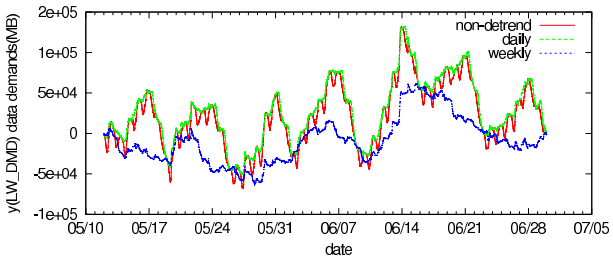

図 3: KEK-LAN からインターネット上の Web サーバへのデータ要求数に対する $y(t)$ 関数。

4.1 Web サーバのアクセスの解析

図 2 は KEK-LAN からインターネット上の Web サーバへの接続数のデータに対する y 関数 (式 2) である．y 関数には平均の 1 日, 平均の 1 週間の周期的トレンドが存在し, それを除去した y 関数を合わせてプロットしている. 図 4(左) は, それぞれの y 関数から計算される F 関数 (式 3b) を示す. 周期的なディトレンドを行わない場合においては, 長時間レンジの領域で折れ曲がりを示すが, 平均の 1 週間によるディトレンドによって F 関数の折れ曲がりが除去された. 接続数に対しては, $\log n \in [1.0, 3.5]$ において $\alpha = 0.77$, $\log n \in [4.0, 5.5]$ で $\alpha = 1.03$ を得た.

図 3 は KEK-LAN からインターネット上の Web サーバに対するデータ転送要求量についての DFA 分析の結果を示す．取得データの y 関数と, 平均の 1 日, 平均の 1 週間の周期的トレンドを除去した y 関数を合わせてプロットした．図 4(右) は, それぞれの計算される F 関数を表している. 平均の 1 週間によるディトレンドで, スケール則の指数は $\alpha = 0.58$ であった.

4.2 Web サーバのアクセスの共分散解析

Web サービスは, HTTP と その暗号化版の HTTPS の 2 種類のプロトコルが使用されており, この混合分布データの共分散解析を行う. 図 5 は,LW の接続数の平均の一週間の周期性を取り除いたデータについての結果を示す．HTTP と HTTPS, 及び Web 全体（HTTP+HTTPS）について

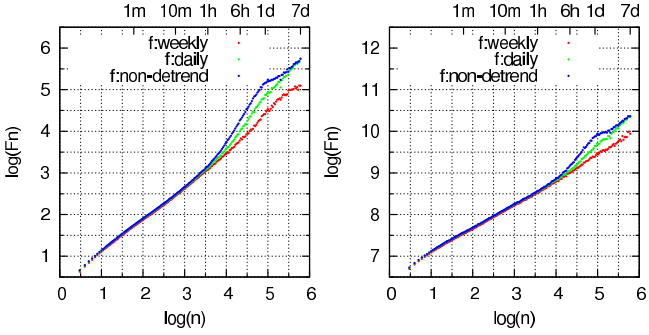

図 4: Web アクセス (LW) に関する F 関数： (左) 接続数（右）データ要求量

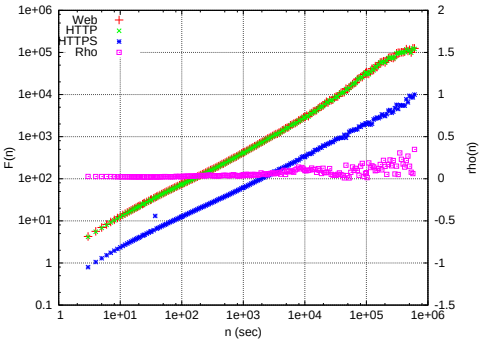

図 5: Web 通信 (LW) の平均の一週間でディトレンドを行ったデータに対する F 関数と共相関係数.

の F 関数と共相関係数 $R^{(i,j)}(n)$ を規格化した

$$\rho^{(i,j)}(n) := \frac{R^{(i,j)}(n)}{F^{(i)}(n)F^{(j)}(n)}.$$

を示している．HTTPS の通信量は HTTP に比して 1/10 であるため, Web と HTTP の F 関数は重なっている．Web の揺らぎは HTTP によって決定されている．スケール指数を求めると, $\log n \in [2.0, 4.0]$ において $\alpha_{\rm http} = 0.77$, $\alpha_{\rm https} = 0.7$ を得た．共分散は $\rho = 0.1$ と小さく, 独立性が高いと思われる．長時間スケールでは ρ は大きくなるが, 統計量が少なく測定誤差を含むことを示している, .

4.3 サービス別揺らぎのスケール指数

各サービスの DFA 解析の結果を, 表 2, 表 3 にまとめる.ALL, Web, Mail, DNS のサービス（プロトコル）について, WL, LW, LL の 3 つの通信グループに分解して解析した. 表 2 は, 接続数に対する解析結果を示す．F 関数はそれぞれスケール則を示した.（Web のデータは図 4 参照）平均の 1 日, 平均の 1 週間を使った周期性を除くと, グラフの折れ曲がりがとれ, スケール則を示した. 周期性除去後もグラフの折れ曲がりを示すものについては, 中時間スケール, 長時間スケールに分けて直線でフィットし, スケール係数 α_f を求めた. 接続数に対するスケールは,$\alpha = 0.65 \sim 1.0$ の弱べき則である. また, 表 3 には, データ要求量に対する

に対する α_f をまとめる. $\alpha_f = 0.5 \sim 0.74$ のランダムの結果を得た.

		中/長レンジ (log)		長レンジ (log)	
サーバ	向き	α_f	区間 (sec)	α_f	区間 (sec)
ALL	WL	1.3	[1.5, 3.0]	0.65	[3.0, 5.5]
	LW	0.6	[2, 4]	—	—
	LL	0.7	[2, 5]	—	—
Web	WL	0.71	[1.5, 3.5]	—	—
	LW	0.77	[1.0, 3.5]	1.03	[4.0, 5.5]
	LL	0.6	[1.5, 3.0]	1.02	[3.5, 5.0]
mail	WL	0.85	[1.0, 5.5]	-	-
	LW	0.75	[1.0, 3.5]	1.06	[3.5, 5.5]
	LL	0.63	[1.0, 4.0]	—	—
DNS	WL	0.73	[1.5, 4.0]	1.02	[4.0, 5.5]
	LW	0.82	[2.5, 4.0]	—	—
	LL	0.56	[1.0, 5.0]	—	—

表 2: 接続数に対するサービス別スケーリング係数

		中/長レンジ (log)		長レンジ (log)	
サーバ	向き	α_f	区間 (sec)	α_f	区間 (sec)
ALL	WL	0.5	[1.5, 5.0]	—	—
	LW	0.6	[1.5, 5.0]	—	—
	LL	0.58	[2.5, 5.0]	—	—
Web	WL	0.5	[1.0, 3.5]	0.76	[3.5, 5.5]
	LW	0.58	[1.0, 4.0]	—	—
	LL	0.54	[1.0, 5.5]	—	—
mail	WL	0.5	[1.5, 5.0]	-	-
	LW	0.6	[2.0, 4.5]	—	—
	LL	0.58	[2.0, 5.0]	—	—
DNS	WL	0.59	[1.0, 3.5]	0.91	[3.5, 5.5]
	LW	0.74	[2.5, 5.0]	—	—
	LL	0.56	[1.0, 5.0]	—	—

表 3: データ要求量に対するサービス別スケーリング係数

混合信号 (ALL) や分解したサービスのスケール係数はサービスごとにばらつきがある. 例えば接続数の ALL[WL] の係数は, 抽出した 3 種類のサービスの係数よりも大きな値を示しており, ALL に含まれる 3 種のサービス以外の影響や大きな共相関の存在を示唆する.

5 まとめと討論

インターネット通信のサービス別 DFA 解析を行った. 揺らぎのスケール指数はサービス間でバラツキを示した. しかしながら, 長時間スケールの結果は統計的誤差が大きく, さらに長期間のログデータの解析が必要となる.

混合信号のデータは自己相関関数 F の自乗と共相関の和であらわされるため, 各サービスのスケール則を独立成分として見なせるかという視点での解析が必要がある. 混合データと分離したデータのスケール則の大きな差が見つかっており共分散解析が必要である. また, 大きな共相関係数が得られた場合には, サービス間に何らかの強い依存関係が存在することを示唆している.

また, 測定結果の時期及び期間依存性や観測対象（場所）依存性などのデータの普遍性について検討が重要である. 時期や期間依存性に関しては, 現在解析を進めている.

対象（場所）依存性に関しては, 文献 [4] と比較することができる. Mail のスケール指数は異なる結果が得られたが, その相違の要因について検討が必要である. 一つは, サンプリング法に伴う雑音の除去の問題がある. 秒刻みの台形則積分による平滑化と,Mail のデータに対して 5 秒, 10 秒区間における平滑化（区間平均による粗視化）を行いスケール指数を比較した. 短時間スケールでの F 関数の振る舞いは, 台形則積分の平滑化では, 平滑化の前にあったの F 関数の折れ曲がりが改善した. 5 秒, 10 秒の区間平滑化では短時間のスケールでの指数がガウス的 ($\alpha \approx 1.5$) となった。 一方で, 長時間スケールでのスケール係数はあまり変化が見られなかった. もう一つは, 観測対象依存性の問題がある. 文献 [4] は単一サーバのログを直接解析したのに対し, 本研究では FW を通過した複数サーバの混合信号を対象としている. このため, 相違を明らかにするためには, 共相関係数を含めたデータの解析と検討が必要である.

謝辞

高エネルギー加速器研究機構で採取された FW の通信のログを解析に使用させていただいた.

参考文献

[1] S.Tadaki, M. Kikuchi, A. Nakayama, K. Nishinari, A. Shibata, Y. Sugiyama and S. Yukawa, J. Phys. Soc. Jpn.75 034002 (2006)

[2] Shin-ichi Tadaki, J. Phys. Soc. Jpn.76 044001(2007)

[3] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, Phys. Rev. E49, 1685-1689 (1994)

[4] 松原義継, 日永田泰啓, 只木進一, 第 14 回交通流のシンポジウム論文集 73-76 (2008)

[5] 柴田章博, 村上直, 第 14 回交通流のシンポジウム論文集 77-80 (2008)

[6] 只木進一, 第 14 回交通流のシンポジウム論文集 69-72 (2008)