インターネット通信のDFAに基づく共分散分析

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

混合信号に対して拡張したDFA解析をファイアウォールの通信ログに適用し，インターネット通信の力学的性質やネットワーク構造の分析を行った．異なる時期，異なる期間のデータに対してDFAを適用しユニバーサルな特性を抽出していることを確かめた．また，トレンドを除去した揺らぎの共分散分析よって，インターネットサーバの連携などの力学的特性を検知できることを示した．

Analysis of Covariance of the Internet traffic based on DFA

Akihiro Shibata, Tadashi Murakami

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

Abstract

We apply the extended DFA (Detrended Fluctuation Analysis) for log data of the firewall, and investigate the dynamical profile of the internet traffic and a structure of the network. By using data of the different epochs and different periods, we show the universal characteristic in the internet traffic. Then, we apply the analysis of covariance to the detrended data. We show that the analysis of covariance can detect the dynamical profile of the Internet such as the linkage of internet servers over the firewall.

1 はじめに

インターネットは，集中制御の仕組みを持たない自律的なネットワークで，ルータを通過するパケットと交通流との類似性で注目が集まっている．また，人やものや情報の行き交うネットワークとしての類似性がある．インターネットのサービスは，複数のサーバが連携・協調して提供されており，ルータを通過するトラフィックは，Web，Mailなどに代表されるend-to-endのインターネットアプリケーションの通信の他に，ネットワーク制御やアドレス解決などがある．前者は，人の社会的活動に起因され，日周期や週周期の周期的活動の影響を受ける．後者は，インターネットアプリケーションからの要求に起因し，サーバのネットワーク上の配置やその動的性質を反映する．

インターネットの動的性質を調べる方法として，インターネット上のルータ（ファイアウォール）を通過するトラフィックの定点観測に対して，トレンド除去法による解析（DFA， Detrended Fluctuation Analysis）がしばしば使われる．ルータを通過するトラフィックで（長時間スケールを含めて）冪則が現れることが報告されている [1][3]． また，インターネットプロトコルに着目したログの解析：電子メイルの通信ログの解析 [2] やファイアウォールのログ [5] が報告されている． しかしながら，着目するサービス（プロトコル）や計測量（接続数，データ量）によってそのスケーリングは異なる．

本研究では，ルータ（ファイアウォール）のログに対して，混合信号に対するDFA解析を行う．前回のシンポジウムの報告 [5] と異なる長期間のデータを用いて，プロトコル別に分解された信号にDFA解析を行い，ユニバーサルな特性があるかを確認する．また，トレンド除去された信号の共分散解析を行うことで，インターネット上のサーバ・サーバ間の連携などの動的な性質についての知見を得ることができるかを検討する．

2 DFA による解析

混合信号に対して拡張されたDFAについて考察する．K 種類の信号 $u^{(k)}(t)$ の混合信号 $u(t) = \sum_{k=1}^{K} u^{(k)}(t)$ ($k = 1,2, \cdots,K$， $t = 0,1,\cdots,T$) に対して，それぞ

れの信号 $u^{(k)}(t)$ に対応する $y^{(k)}(t)$ 関数を次のように定義する．

$$y^{(k)}(t) := \int_0^t \left(u^{(k)}(s) - \left\langle u^{(k)} \right\rangle \right) ds. \tag{1}$$

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：$\left\langle u^{(k)} \right\rangle = \frac{1}{T}\int_0^T u^{(k)}(s)ds$．単一信号の DFA 解析と同様の手続きによって，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 次多項式 (ここでは $p = 1$) で χ^2 フィットしたトレンド関数を定義する．それとの差でトレンドを除去した揺らぎを次で定義する．

$$\Delta y_m^{(k)}(t) := y^{(k)}(t) - \tilde{y}_{n,\,m}^{(k)}(t). \tag{2}$$

区間 m に於ける揺らぎ $\Delta y^{(k)}(t)$ に対する共分散を

$$R_m^{(k,k')}(n) := \frac{1}{n}\int_{(m-1)n}^{mn} \Delta y_m^{(k)}(t)\Delta y_m^{(k')}(t)dt \tag{3}$$

で定義すると，その M 個の区間における平均で（同時刻）共相関関数

$$R^{(k,k')}(n) := \frac{1}{M}\sum\nolimits_{m=1}^{M} R_m^{(k,k')}(n). \tag{4a}$$

が得られる．各成分 k の揺らぎ関数 F は，$F^{(k)}(n) := \sqrt{R^{(k,k)}(n)}$ で与えられる．

混合信号 $u(t)$ に対する F 関数は，対応する $y(t) := \sum_{k=1}^{K} y^{(k)}(t)$ を用いて次のように計算される．

$$F(n)^2 = \sum\nolimits_{k=1}^{K} \left(F^{(k)}(n) \right)^2 + 2\sum\nolimits_{k<k'=1}^{K} R^{(k,k')}(n). \tag{5}$$

ここで注意すべきは，混合信号の揺らぎ係数は各成分揺らぎ $F^{(k)}$ の二乗と共相関関数の和であらわされるため，最も大きな振幅を与える成分によって F 関数は支配される．

信号の各成分の力学的関係は，共相関係数に反映される．k_1, k_2 の揺らぎの起源が独立であれば (k_1, k_2) の組の共相関係数は，$R^{(k_1,k_2)}(n) = 0$ となる[1]．$R^{(ij)}(n)$ は，$F^{(i)}(n)$ の振幅に比例するため，規格化した $\rho^{(i,j)}(n)$ を導入する．

$$\rho^{(i,j)}(n) := \frac{R^{(i,j)}(n)}{F^{(i)}(n)F^{(j)}(n)} \in [-1,\,1]. \tag{6}$$

インターネット通信に含まれる周期的トレンドを 文献 [2][3] 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

$$\tilde{u}_Q(\tau) := \frac{1}{N_Q}\sum\nolimits_{k=0}^{N_Q-1} u(\tau + kT_Q),\, (\tau = t \bmod T_Q\,) \tag{7}$$

によって周期的トレンドが除去された関数を定義する．

$$u_Q(t) := u(t) - \tilde{u}_Q(t \bmod T_Q). \tag{8}$$

[1]同時刻共相関関数がゼロであることをもって，独立成分であるとすることはできないが，力学的性質を検知する良い指標である．

期間	(A)2008/5/12〜2008/6/29 (7 週間)[5] (B)2009/4/20〜2010/1/10 (38 週間) (C)2009/5/11〜2009/6/28 (7 週間)
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 ※脆弱性診断装置の通信は除外
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 Others(上記サービス以外の全通信)

表 1: データ収集の条件とデータの概要

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，インターネット，DMZ，LAN の三つのセグメントに分割され FW を経由して接続される．([3] の図 1 参照）LAN はさらに複数の VLAN に分割され，VLAN 間の通信は FW を経由する．FW のログは，FW を経由する全ての通信が記録の対象となっており，接続元及び接続先の IP 番号， port 番号 ，通信種別，通信の可否，送信/受信のデータサイズ，セッションの開始時刻と接続時間などの情報を保持する．

測定対象としたデータを表 1 にまとめる．期間 (A)[5] に加え，新たに 38 週間の期間 (B) のデータを解析の対象とし，長時間スケールのデータを得るとともにデータの期間依存性について検討を加える．データ解析は，脆弱性診断装置などのネットワーク管理用通信を除外した接続件数とデータ要求量について行う．また，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．さらに，プロトコルで分類して，Mail(SMTP)，Web(HTTP+HTTPS)， DNS，及びその他のカテゴリーについて解析する．

4 通信揺らぎ解析

本節では，インターネット通信におけるユニバーサルな特性の検討として，表 1 の条件に従ってプロトコル別 DFA 解析を行う．図 1，図 2 は期間 (A)，(B) それぞれの電子メイル送信数に対する $F^{(k)}(n)$ 関数を示している．期間 (B) における週周期トレンド除去後の $F^{(k)}(n) \sim n^{\alpha_f}$ について，それぞれの指数 α_f を表 2 と表 3 にまとめる． なお期間 (A) の解析結果は論文 [5] に報告済みである．$F^{(k)}(n)$ はどれもスケール則を示した．いくつかの $F^{(k)}(n)$ は、周

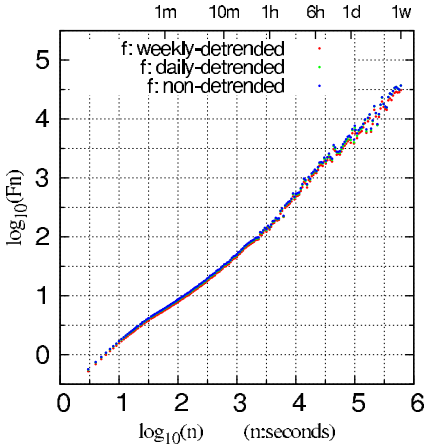

図 1: 期間 (A) の電子メイル送信数に対する $F^{(k)}(n)$ 関数

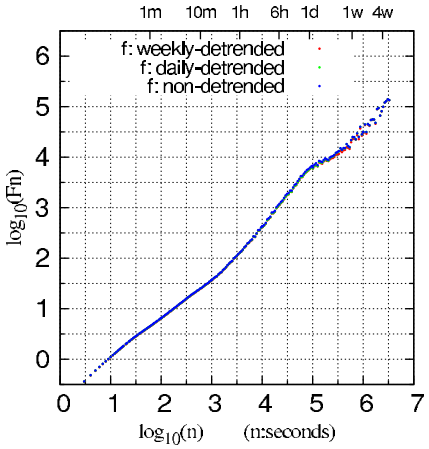

図 2: 期間 (B) の電子メイル送信数に対する $F^{(k)}(n)$ 関数

期的トレンドを除去後も数時間の時間スケールで折れ曲がりを示すものがあり，区間を分けてフィットを行った。表中の区間 (log) の欄は、α_f を求める際に log-log プロットした $F^{(k)}(n)$ 関数を直線でフィットした n の区間を log スケールで示す．

期間 (B) のスケーリング係数 α_f については，3 時間を越える長レンジ ($n > 10^4$) では Web や Mail の接続数についてべき則を示した．それ以外の通信，特に 3 時間未満の中レンジデータ要求量の係数 α_f については，概ねランダムであると解釈できる．

期間 (A)，(B) について，α_f および区間で比較したところ，スケーリング則が良い一致を見た． ここで，KEK のネットワーク運用や利用状況は時とともに変化しており，それに伴って通信のトレンドが変化していることに注意が必要である．例えば期間限定の大量通信，パケットルーティングの経路変更，主要サーバの頻繁な構成変更などが (B) の期間中に行われている．これらをふまえると，$F(n)$ が時期やトレンドの変化に依存せずに，ネットワークの動的性質を示せたことを示唆している．

		中/長レンジ (sec)		長レンジ (sec)	
サーバ	向き	α_f	区間 (log)	α_f	区間 (log)
Web	WL	0.69	[2.0, 3.8]	0.91	[3.8, 6.5]
	LW	0.70	[1.0, 3.3]	1.18	[3.3, 5.0]
	LL	0.53	[1.5, 3.0]	1.16	[3.8, 6.0]
Mail	WL	0.88	[0.5, 6.5]	-	-
	LW	0.78	[1.0, 3.3]	1.29	[3.3, 5.0]
	LL	0.72	[1.0, 6.5]	—	—
DNS	WL	0.85	[1.0, 4.0]	0.70	[4.5, 6.5]
	LW	0.80	[1.0, 4.0]	1.05	[4.0, 6.0]
	LL	0.84	[1.0, 3.5]	0.67	[3.5, 5.0]

表 2: 接続数に対するサービス別スケーリング係数 α_f(weekly detrended)

		中/長レンジ (sec)		長レンジ (sec)	
サーバ	向き	α_f	区間 (log)	α_f	区間 (log)
Web	WL	0.50	[1.0, 4.0]	0.76	[4.0, 6.5]
	LW	0.55	[1.0, 3.8]	0.84	[3.8, 6.0]
	LL	0.58	[1.0, 5.5]	—	—
Mail	WL	0.47	[1.0, 4.0]	-	-
	LW	0.60	[1.0, 4.8]	1.00	[4.8, 6.0]
	LL	0.76	[1.1, 2.5]	0.56	[2.5, 6.0]
DNS	WL	0.56	[1.5, 4.0]	0.68	[4.0, 6.5]
	LW	0.70	[1.0, 4.0]	1.00	[4.0, 6.5]
	LL	0.60	[1.0, 5.0]	1.40	[5.5, 6.5]

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 α_f(weekly detrended)

DNS では，WL の長レンジと LL の中レンジについて，期間 (A) と期間 (B) のスケーリング則が一致しなかった．期間 (B) の後期では，DNS のサーバ構成の頻繁な変更や，ネットワークルーティングの変更などが行われたため，その影響を受けた可能性がある．また，DNS 通信は LAN 内の他のサーバの構成に依存していると考えられるため，ネットワークの変更の影響を受けて結果的に時期依存性が大きくなった可能性がある．

期間 (C) についても同様に解析および比較を実施したところ，期間 (A), (B) の比較と同様の結果を得た．

なお，別のユニバーサルな特性として，場所依存性が挙げられる．この検証としては，佐賀大学のメイルサーバのログ解析 (2008/5/9〜2008/9/30 実施) の結果 [2] を利用できる．KEK の結果（図 1,2 参照）と比較したところ，係数 α_f が概ねべき則を示す点では一致が見られたものの，折れ曲がりのレンジなどが不一致であった．この原因として，佐賀大学のメイルサーバは大学在籍の教職員向けである一方，期間 (A)，(B)，(C) で測定した SMTP 通信は，例えば機器のログを定期的に送信するなど用途が多彩であることなどを挙げることができる．これを踏まえ，例えば KEK の教職員向けメイルサーバに限定するなど，条件を統一して比較することが必要である．

	WEB	SMTP	Others	DNS
WEB	-	×	×	○
SMTP	×	-	×	○
Others	×	×	-	△
DNS	△	△	△	-

表 4: 接続数に対する共相関係数 $\rho_{i,j}$. 表の上三角は LW の通信，下三角は WL 通信に対する結果を示す．○は有意な相関が検知されたこと，△ は若干の相関が検知されたこと，× は相関が検知されなかったことをあらわす．

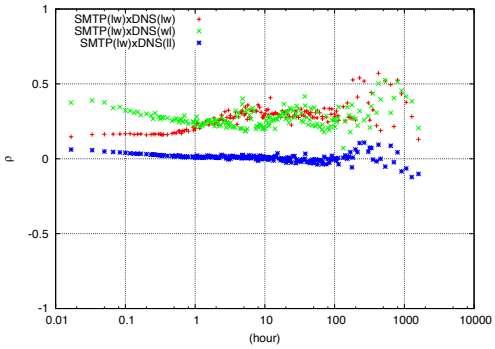

図 3: SMTP-DNS 共相関係数

5　揺らぎの共相関分析

次に，相関係数について解析する．ファイアウォールで観測される通信は，end-to-end の通信の他に，インターネットのサーバ群の有機的な通信の混合として観測される．トレンドを除去した各サービスの相関係数は，通信揺らぎの起源が異なる場合にはゼロとなる．すなわち，共相関係数を測ることで，起源を共通にする揺らぎを検知できる．主要なインターネットサーバとして，Web，Mail，DNS を抽出してその通信揺らぎの相関係数を観測することで，サーバ間の連携を FW をまたぐ通信について検知できるかについて検討する．接続数に着目すると， 表 1 の区分から得られる係数 ρ は ${}_{12}C_2 = 66$ 個あるが，統計的に有意な相関が認められるのは主として DNS がかかわる通信となる．

表 4 は，LW 及び WL の接続数における共相関係数を測った結果である．Web，Mail，Other の間では，相関は認められないが DNS との間には相関が認められた．即ち，DNS サーバは，直接人がアクセスすることがなく，WEB ページの URL や電子メイルの送信先などのアドレス解決のためにインターネット・アプリケーションから連動してクエリが出されるインターネットの力学的性質の反映である．一方，Web サーバのアクセスや SMTP 送信の行為は主として人に起因しており独立な操作の反映である．

ここでは，インターネット通信の連携の性質をさらに詳しく調べるために Web(LW，WL)，Mail(LW，WL)，DNS(LW，WL，LL) との間の共相関係数について解析する．図 3 は，Mail(LW) と DNS の共相関係数 ρ を示している．DNS(LW) と DNS(WL) の間に強い相関を示すが，一方，DNS(LL) との相関は見られない．DNS(LW) との相関は送信時にアドレス解決のための DNS を参照していることに起因する．DNS(WL) との相関は，Web(LW，WL) や Mail(WL) と DNS(WL) の間には見られない（非常に弱い相関）．

6　まとめと討論

インターネットのサービスの混合信号に対する DFA 解析を行なった．異なる時期と期間における解析を行なうことで，結果の時期依存性がないことを確認した．また，インターネットサービス（プロトコル）の同時刻共相関係数を計測することで，ファイアウォールをまたぐサーバ間の連携およびその強さを検知することができた．共相関係数は時間スケールの関数として得られるため，時間スケールの依存性を詳細分析を行なうことで，その力学的な性質を検知することができると期待される．今後、電子メイルの転送やプロキシーなどの相関を検知できるかについて検討を進める。

謝辞

解析に使用したインターネット通信のデータは，高エネルギー加速器研究機構（KEK）のファイアウォールのログを使用した．また，科研費（基盤 (B) 20360045）のサポートを受けました．

参考文献

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

[2] 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)

[3] 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)

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

[5] 柴田章博，村上直，第 15 回交通流のシンポジウム論文集 65-68 (2009)