インターネット計測とデータ解析 第10回

長 健二朗

2011年7月6日

前回のおさらい

インターネットの時間変化を計る

- インターネットと時刻
- ネットワークタイムプロトコル
- 時系列解析
- 演習:時系列解析

今日のテーマ

インターネットのトラフィック量を計る

- トラフィック計測
- 演習:トラフィック量解析

インターネットの挙動を計る

トラフィック量

- ネットワーク計測の基本指標
- 収集方法
  - SNMP によるインターフェイスカウンタ値の収集
  - NetFlow などの flow 計測
  - パケットキャプチャリング
- 個別回線の計測とデータの集約
  - 加算可能性: 平均値は加算可能、最大値等は加算できない
  - ダブルカウントの問題

今回は、具体例としてブロードバンドトラフィック計測を紹介

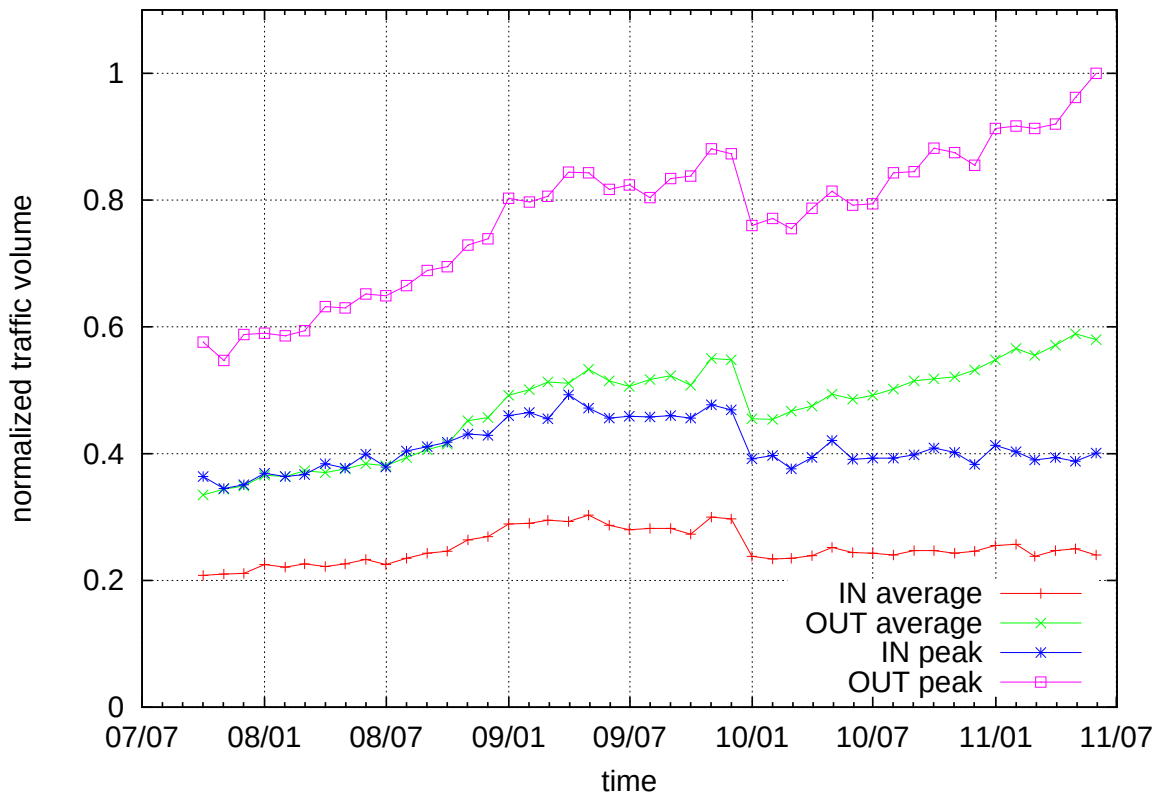
ブロードバンドトラフィックの傾向

- 過去 5 年ほどは年率 30%程度の安定した伸び
- しかし、過去のデータをもとに将来の予測は難しい
  - 一部のヘビーユーザの挙動が大きく影響
  - 技術以外の社会的要因等で利用の仕方が大きく変わる可能性

2010年1月に大きな変化

実際、2010年1月に20%近く急減

- これまでにも変動はあったがここまで長期的影響は初めて
- 改正著作権法 (ダウンロード違法化) の影響か？
  - 罰則規定のない改正なので、ここまでの影響は予想外

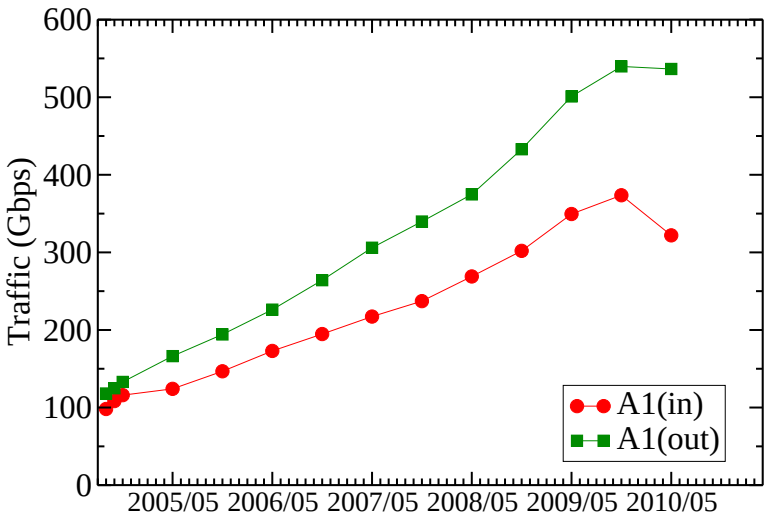


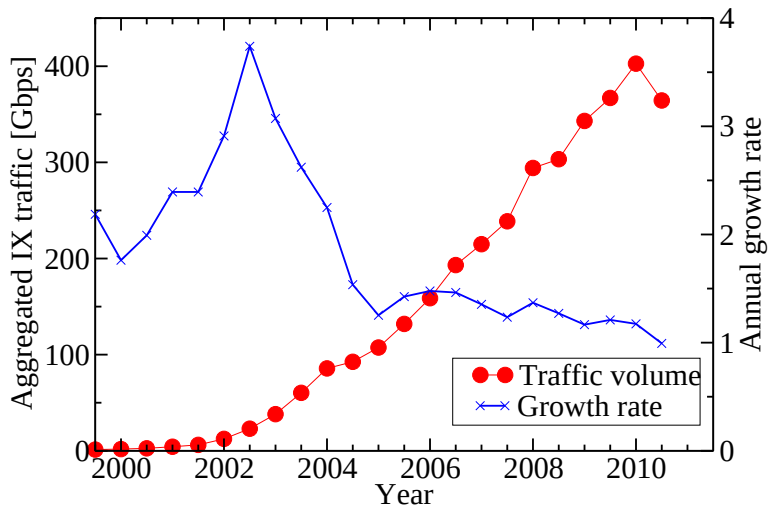
2009年と2010年のデータを比較し原因を探る

国内全体の傾向

総務省「我が国のインターネットにおけるトラヒックの集計・試算」

- １月のトラフィック減少は日本全体で観測されている



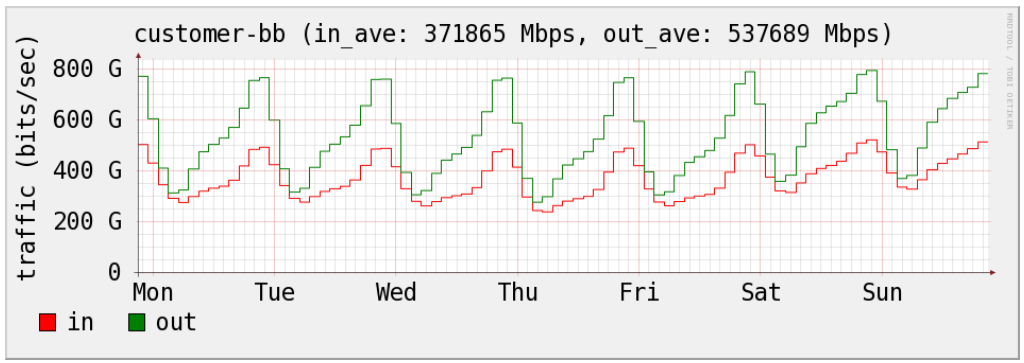


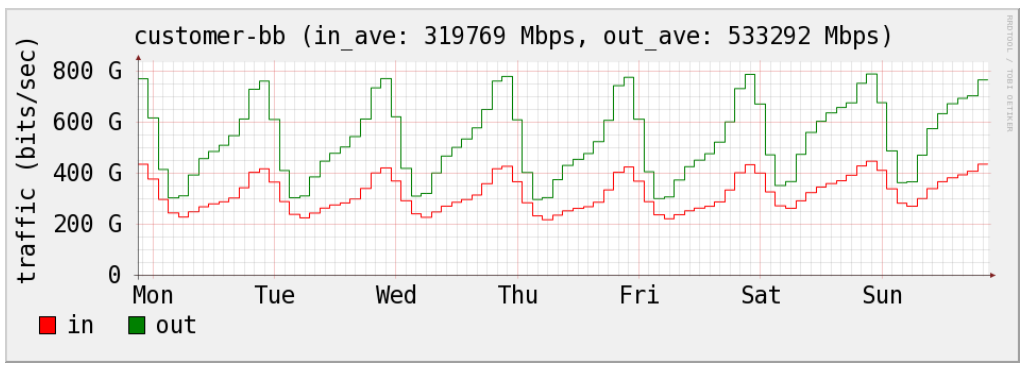
国内 ISP6 社のブロードバンドトラフィック (左) 主要 IX トラフィック (右)

ブロードバンド週間トラフィックの変化

- 家庭利用のトラフィックパターン (ピークは 21-23 時)
- 2005 年頃は IN/OUT はほぼ同量 (P2P トラフィックが支配的)
- 除々に OUT(利用者のダウンロード) が大きく

P2P ファイル共有から web サービスへのシフトが窺える





ブロードバンド週間トラフィック: 2009(上) 2010(下)

ブロードバンド利用者別データの解析

- IIJ が運用するブロードバンドサービスが対象
- Sampled NetFlow 形式のデータ
  - FTTH/DSL ブロードバンド顧客収容ルータ
- 1 週間分のデータ
  - 2009 年 5 月と 2010 年 5 月の比較
  - 平日と休日でパターンが異なる、7 で割った 1 日平均を使用

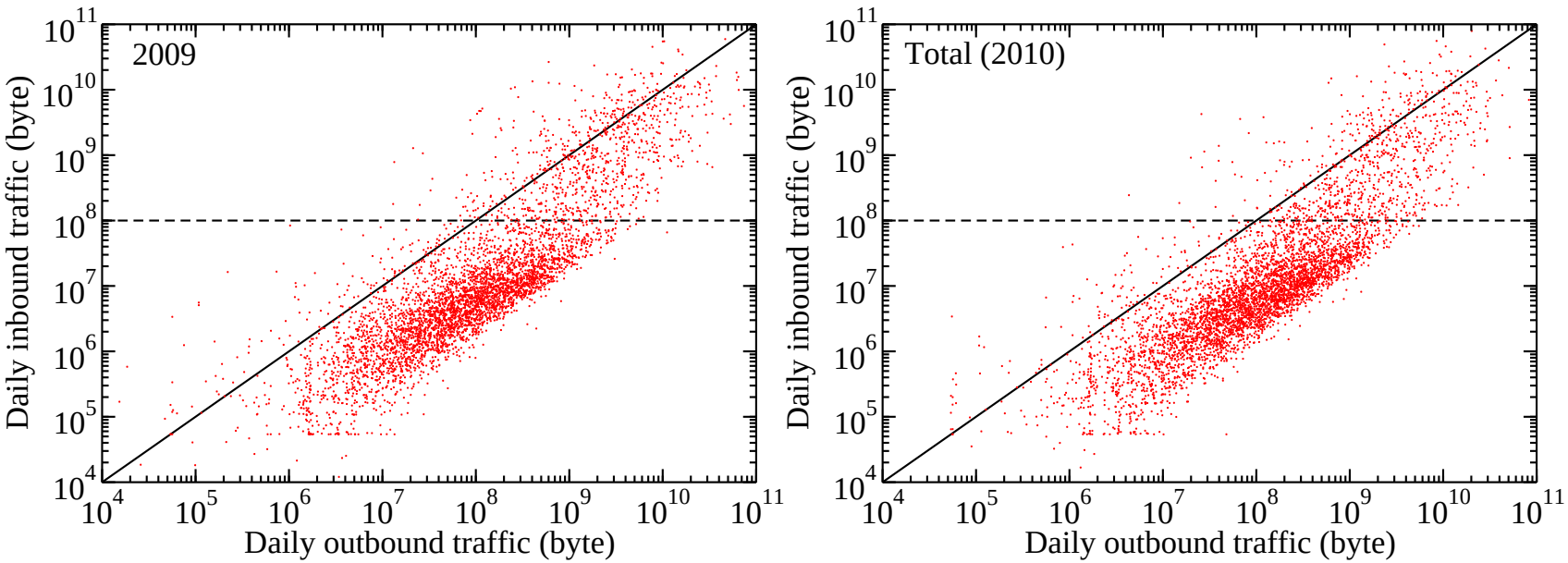
IN/OUT は ISP からの視点

利用者ごとの IN/OUT 使用量

5000 ユーザをランダムサンプリングし IN/OUT をプロット

2 つのクラスタ: クライアント型一般ユーザとピア型ヘビーユーザ

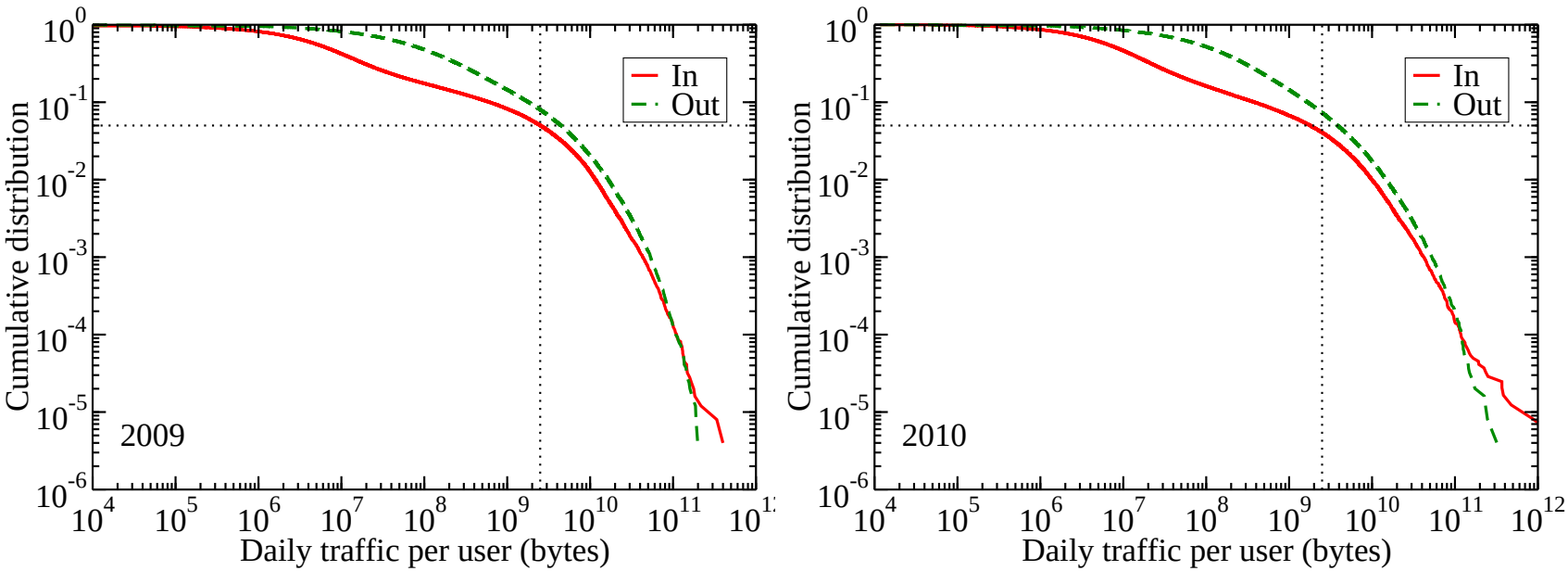
- 境界はあいまい
  - ヘビーユーザとそれ以外、クライアント型とピア型
- 利用者は両タイプのアプリケーションを異なる割合で使用



利用者ごとの IN/OUT 使用量 (左)2009 (右)2010

トラフィック使用量のユーザ分布

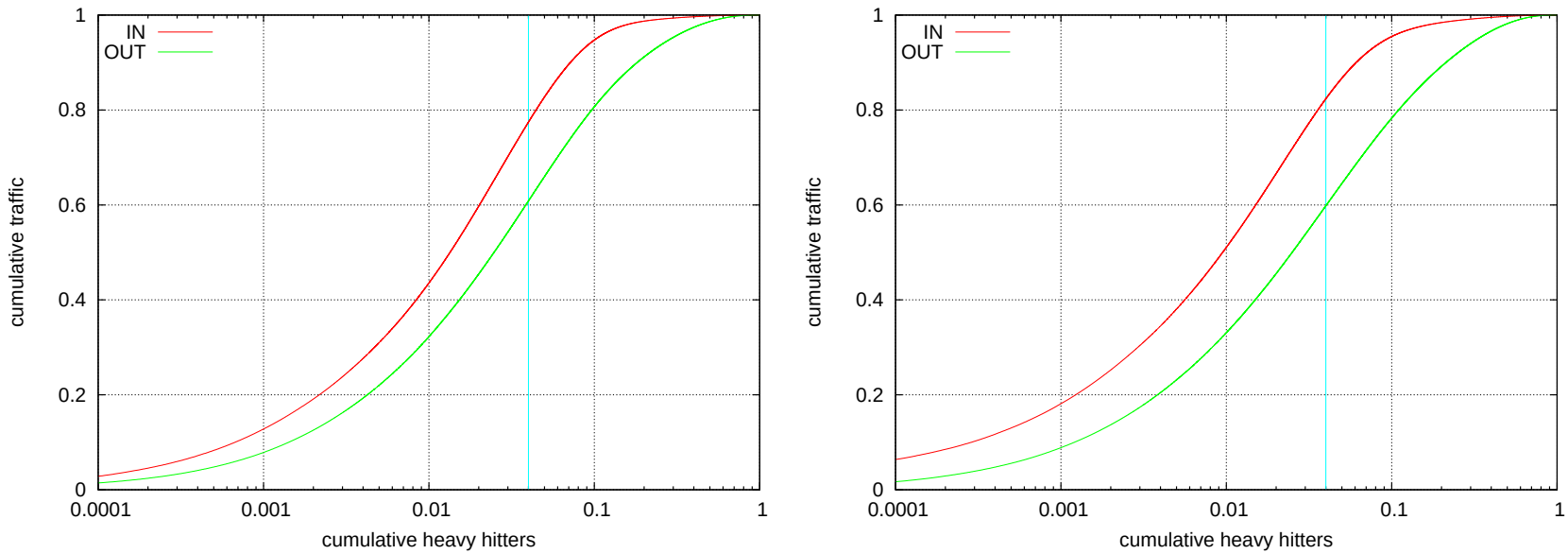
- ベキ分布的 (確率的な分布)
  - 幅広いヘビーユーザが存在
- 2010 年には IN 側でヘビーユーザの割合が若干減少
  - 100MB/日以上アップロードするユーザの総数は 20%程減少
  - 一方で、右端の極端なヘビーユーザは逆に増えている



トラフィック使用量の相補累積分布: (左)2009 (右)2010

利用者間のトラフィック使用量の偏り

- ユーザ別の使用量に大きな偏り
  - 2010 年: 上位 10%の利用者が OUT の 78%、IN の 96%を占める
- 2009 年と比較すると IN 側の偏りが拡大
  - ヘビーユーザ総数は減ったが、極端なヘビーユーザは増えた

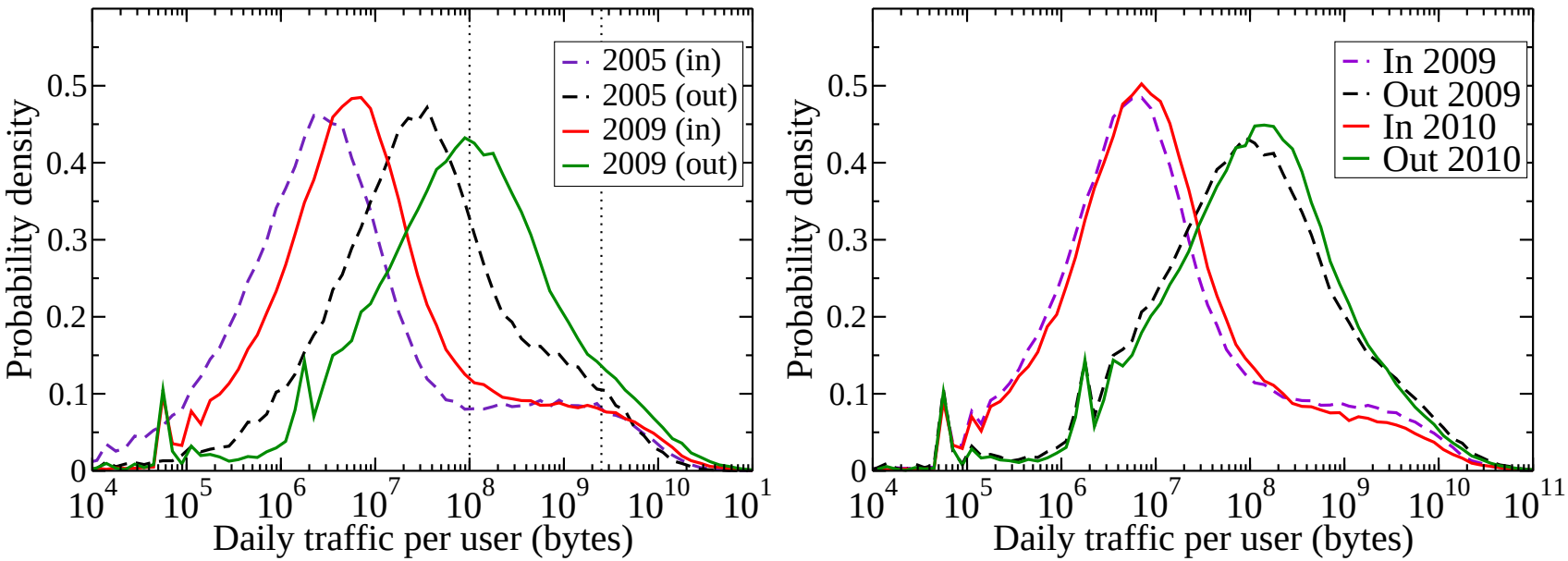


利用者間のトラフィック使用量の偏り (左)2009 (右)2010

利用者ごとの１日の使用量

- IN/OUT の各分布は２つの対数正規分布から成る
  - ダウンロードがひと桁多いクライアント型グループ
  - 利用量の多い IN/OUT 対称的なピア型グループ

	IN (MB/day)		OUT (MB/day)	
	mean	mode	mean	mode
2005	430	3.5	447	32
2009	556	6	971	114
2010	469	7	910	145

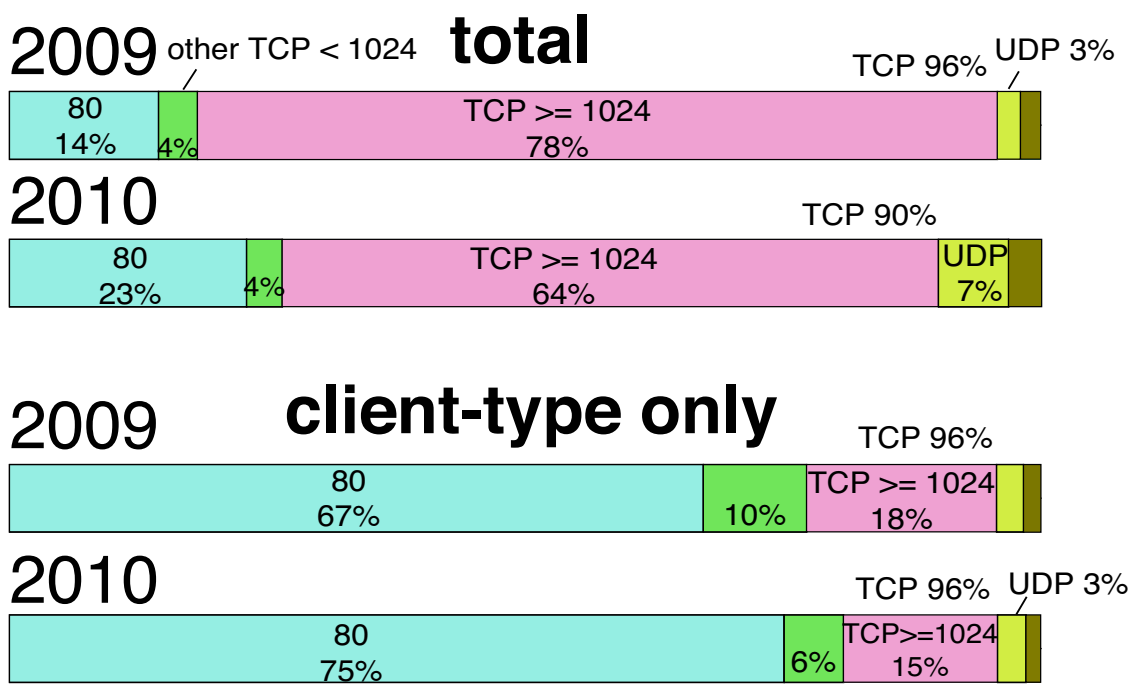


利用者の 1 日の使用量分布 (確率密度関数) (左)2005 と 2009 (右)2009 と 2010

プロトコル別使用量

アップロード 100MB/日でピア型とクライアント型を分類

- ポート番号: min(sport, dport)
  - 一般に、well-known ポートはクライアントサーバ型アプリケーション、動的ポートは P2P の可能性が高い
- 全体でみるとほとんどは TCP の動的ポート
- TCP80 番ポートが増加傾向
  - 2010 年に動的ポート同士の通信は 25%程減少、そのうち 1/3 は 80 番ポートに移行



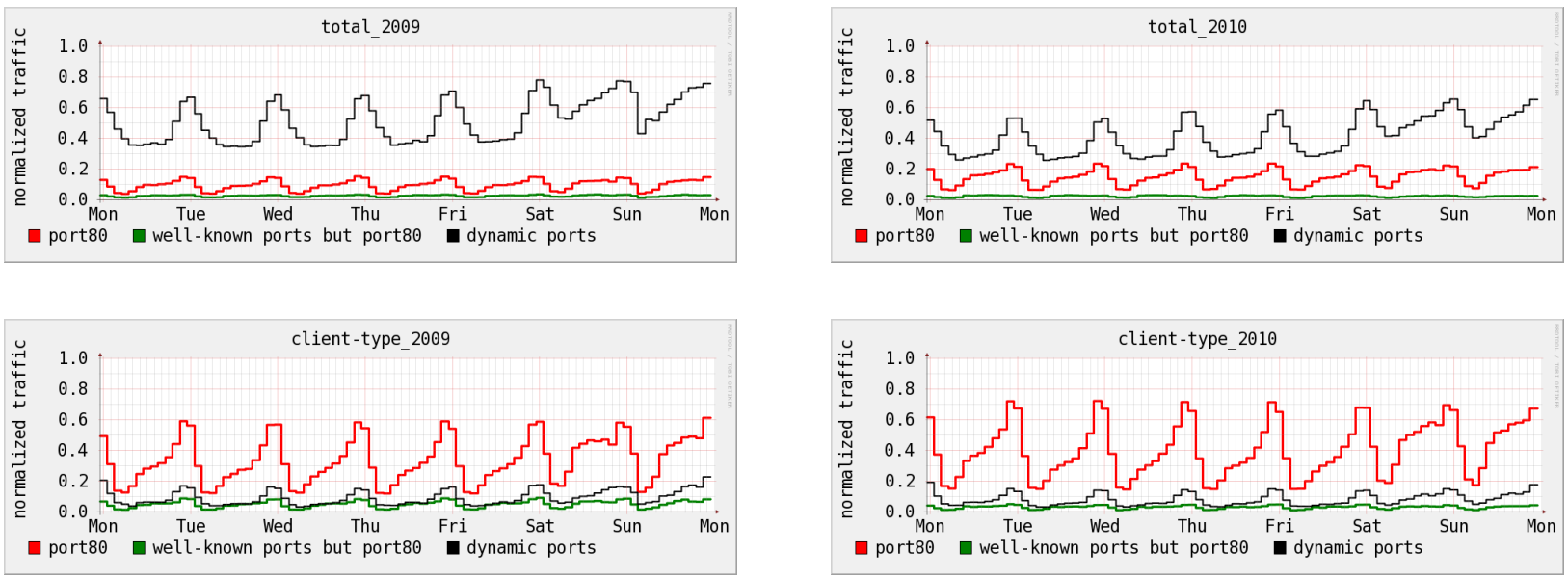
プロトコル別使用量詳細

protocol port	2009 total (%)	2009 client type	2010 **total** (%)	2010 client type
TCP *	**95.80**	**95.73**	**90.09**	**95.82**
(< 1024)	*18.23*	*77.31*	*26.46*	*80.87*
80 (http)	14.46	67.30	23.00	75.12
554 (rtsp)	1.48	6.89	1.15	2.45
443 (https)	0.64	1.91	0.98	2.28
20 (ftp-data)	0.19	0.17	0.18	0.07
(>= 1024)	*77.57*	*18.42*	*63.63*	*14.95*
1935 (rtmp)	0.36	1.51	1.04	2.91
6346 (gnutella)	1.10	0.60	0.86	0.33
6699 (winmx)	0.70	0.24	0.65	0.17
8084	0.00	0.00	0.61	0.00
UDP	**2.24**	**2.60**	**6.79**	**2.76**
ESP	**1.87**	**1.55**	**2.91**	**1.30**
GRE	**0.07**	**0.08**	**0.14**	**0.06**
IP-IP	**0.01**	**0.00**	**0.04**	**0.01**
ICMP	**0.02**	**0.05**	**0.02**	**0.04**

TCP ポート利用の週間推移

3 つに分類: 80 番, その他の well-known ポート, 動的ポート

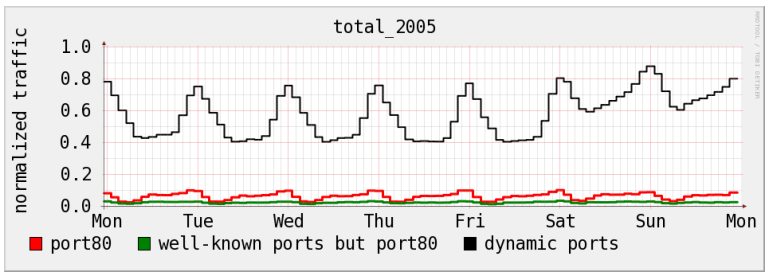
- 合計のピーク値で正規化
- 全体でも動的ポートが減って 80 番のトラフィックが増加
  - これまではクライアント型に顕著な傾向

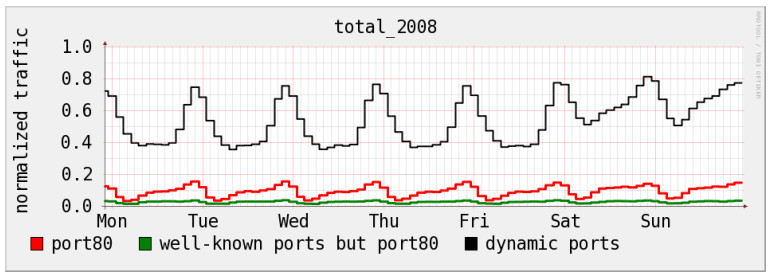


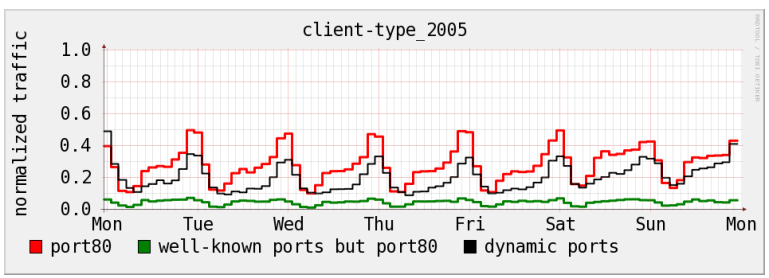
TCP ポート利用の週間推移: (上) 全体 (下) クライアント型 (左)2009 (右)2010

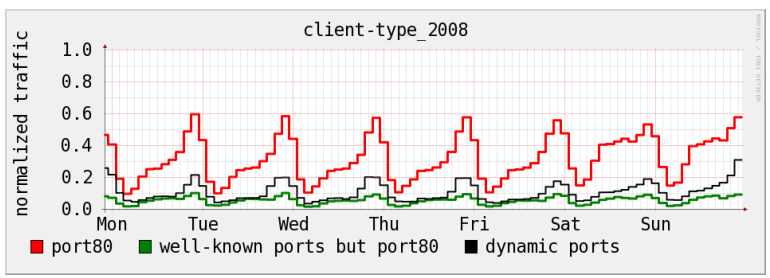
参考: 2005 年と 2008 年の比較

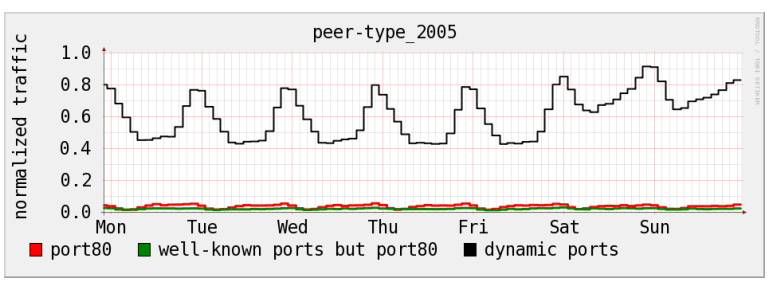
- 全体はピア型ユーザに利用を反映
- クライアント型で 80 番ポートの増加が目立った

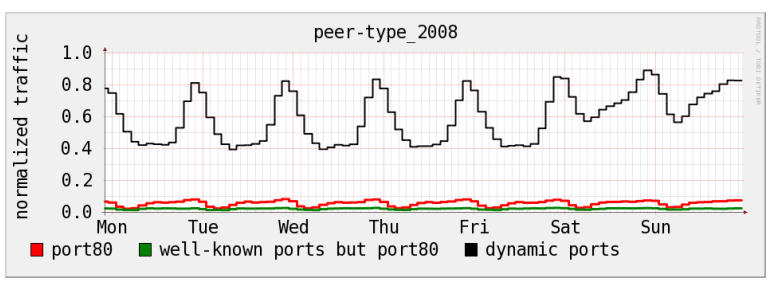












(上) 全体 (中) クライアント型 (下) ピア型 (左)2005 (右)2008

まとめ

- ブロードバンドトラフィック
  - 過去 5 年は年率 30%程で安定した伸びをしていた
  - 2010 年 1 月に急減
- トラフィックパターンの変化傾向
  - 全体でみると依然 P2P ファイル共有が支配的
  - しかし、web ベースのサービスへのシフトが明確に
  - 各利用者は多様なアプリケーションを異なる割合で使用
- 2010 年に入っての特徴
  - いままでの傾向に大きな変化はない
  - ヘビーユーザのトラフィック変動がこれまでより大きい
  - ヘビーユーザや動的ポート同士の通信が単純に減った訳ではない
    - ヘビーユーザ数は 20%程減少、一方で極端なヘビーユーザは増加
    - 動的ポート同士の通信は 25%程減少、そのうち 1/3 は 80 番ポートに移行
  - これまでは、一般ユーザの動向に顕著だった web サービスへのシフトが、今回、ヘビーユーザにも広がった

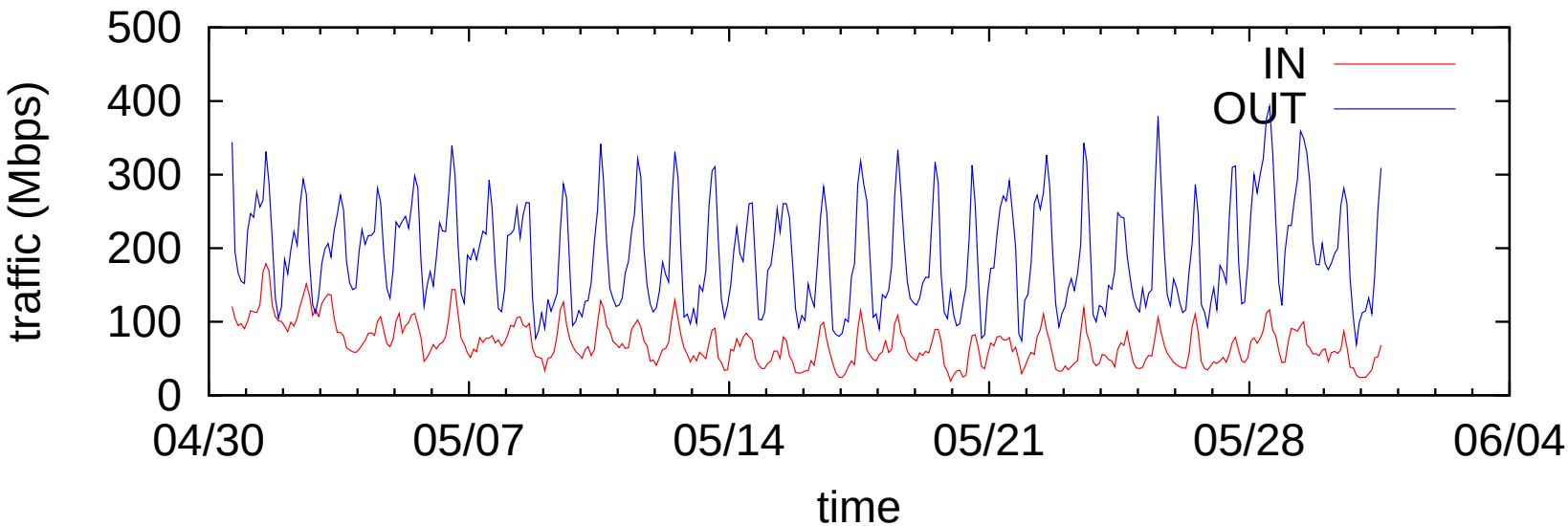
改正著作権法の影響の考察

- 以前から P2P ファイル共有から web サービスへシフトする流れ
  - ネットのビデオコンテンツのユーザ層の広がり
  - 代替技術として web ベースのサービスの成熟
  - P2P ファイル共有使用リスクに対する社会的認識の変化
- 改正著作権法を契機に、この流れが加速した
  - 例え:地震で地滑りが起こった、本当の原因は地盤の緩み
- 世界的にも同様の事例が報告されている
  - 2009 年スウェーデンの著作権強化でトラフィック半減など

演習: トラフィック解析

演習用データ: ifoctets.txt

- あるブロードバンド収容ルータのインターフェイスカウンタ値
- 2011 年 5 月の 1ヶ月分、2 時間粒度
- format: unix_time IN(bytes/sec) OUT(bytes/sec)



最終レポートについて

- A, B からひとつ選択
  - A. SFC Web アクセスログ解析
  - B. 自由課題
- 8 ページ以内
- pdf ファイルで提出
- 提出〆切: 2011 年 7 月 30 日 (土) 23:59

最終レポート 選択テーマ

A. SFC Web アクセスログ解析
- SFC Web アクセスログ (weblog-20110516-20110522.txt)
- 上記ログを元に訪問者毎のアクセスパターンを抽出したデータ (visit-pattern-201105.txt, idmap-201105.txt)
- 小課題
  1. 訪問者のアクセスパターンから、訪問者毎のアクセス数分布の CDF と CCDF の 2 つのプロットを作成せよ。
  2. 訪問者毎のアクセスパターンに関して、機械的な自動アクセスを統計的に除外する手法を考案し、その手法の利点と欠点について考察せよ。
     (注: この課題にはひとつの正解があるわけではない。 どのような方法を使っても誤判定の可能性はあるので、比較的簡単に自動アクセスを除外する手法を考えればよい。)
  3. アクセスログに関する自由分析。
     データを元になんらかの分析と考察を行い、分析手法の説明と結果に対する考察を記述する。できれば、SFC の Web サイトのデザインに関して、何らかの改善提案ができるとよい。

B. 自由課題
- 授業内容と関連するテーマを自分で選んでレポート
- 必ずしもネットワーク計測でなくてもよいが、何らかのデータ解析を行い、考察すること

訪問者のアクセスパターンについて

訪問者のアクセスパターンデータは以下のように作った

- weblog-20110516-20110522.txt をアクセス時間順にソートし、weblog-sorted-20110516-20110522.txt を作成。(オリジナルのログでは、アクセス時間順序が前後している場合がある。以下の処理を簡単にするため、まず時間順にソートする。)

```
% ./sort-before.rb weblog-20110516-20110522.txt | sort -n -k1,1 -s | \
  ./sort-after.rb > weblog-sorted-20110516-20110522.txt
```

`sort-before.rb`: アクセスタイムを `unix time` にして行頭にプリペンドする
`sort-after.rb`: 行頭の `unix time` を削除

- 同一 IP アドレスから、15 分以内の間隔でアクセスがあるものを、ひとつの訪問と見て、アクセスパターンを集計する。スクリプト (visit-pattern.rb) を利用。
- データ
  - visit-pattern-201105.txt: 訪問者毎のアクセスパターン

    `format: start_time stay_time(sec) number_of_access ## list_of_url_id`
  - idmap-201105.txt: コンテンツの ID と URL の対応マップ

    `format: id(rank) url number_of_hits`

訪問者毎のアクセスパターン

```
2011-05-16T12:00:00 180 22 ## 3 10 12 1 2 7 8 5 9 6 72 77 161 122 43 50 1 97 63 141 36 104
2011-05-16T12:00:04 1 2 ## 667 2
2011-05-16T12:00:07 347 4 ## 276 359 40 70
2011-05-16T12:00:08 0 1 ## 40
2011-05-16T12:00:14 3 9 ## 1 2 9 5 8 7 6 4 11
2011-05-16T12:00:27 108 14 ## 385 2 5 6 1 7 9 8 411 383 419 208 4 11
2011-05-16T12:00:29 16 13 ## 1 2 2 5 7 9 8 6 4 11 25 33 36
2011-05-16T12:00:31 7 13 ## 3 10 10 12 1 2 5 8 7 9 6 4 11
2011-05-16T12:00:35 21 24 ## 3 10 12 15 14 19 20 21 18 13 6 31 52 75 2 5 8 7 9 31 3 1 205 212
2011-05-16T12:00:41 0 1 ## 3
2011-05-16T12:00:44 1 6 ## 41 2 5 9 7 8
2011-05-16T12:00:52 0 1 ## 1
2011-05-16T12:00:54 252 3 ## 1 4 56
2011-05-16T12:01:00 0 1 ## 1
2011-05-16T12:01:14 36 3 ## 1 4 25
2011-05-16T12:01:31 30 11 ## 4 3 10 12 15 14 20 19 21 18 13
2011-05-16T12:01:31 337 3 ## 55 213 67
2011-05-16T12:01:42 0 1 ## 24
2011-05-16T12:01:50 55 6 ## 1 72 79 186 40 55
```

コンテンツの ID と URL の対応マップ

```
1 /top.html 30480
2 /css/main.css 26124
3 / 23791
4 /students_soukan/ 23728
5 /images/keio_logo.gif 22968
6 /favicon.ico 21832
7 /images/gaibu.gif 21610
8 /images/notice.gif 18261
9 /images/rss.gif 18239
10 /css/top.css 14892
11 /images/new.gif 13179
12 /js/cookie.js 12438
13 /images/copy.gif 8598
14 /images/keiou.gif 8240
15 /images/pen.gif 8120
16 /files/61/Graduate_School_of_Media_and_Governance_Guidebook2011.pdf 7408
17 * 7057
18 /images/htm_a.gif 5161
19 /images/flash_a.gif 5157
20 /images/flash_b.gif 5152
21 /images/htm_b.gif 5141
22 /images/notice-new.gif 3844
```

まとめ

インターネットのトラフィック量を計る

- トラフィック計測
- 演習:トラフィック量解析

次回予定

第 11 回 インターネットの異常や問題を計る (7/13)

- 異常検出
- スパム判定
- ベイズ理論
- 演習:異常検出