インターネット計測とデータ解析 第8回

長 健二朗

2010年12月8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
- 時系列解析
- **課題 2**

今日のテーマ

インターネットの挙動を計る

- トラフィック量
- (経路情報)
- インターネット計測とプライバシー

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

経路情報

- AS 内部と AS 間の 2 階層
- トポロジーの回でやったので今回省略

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

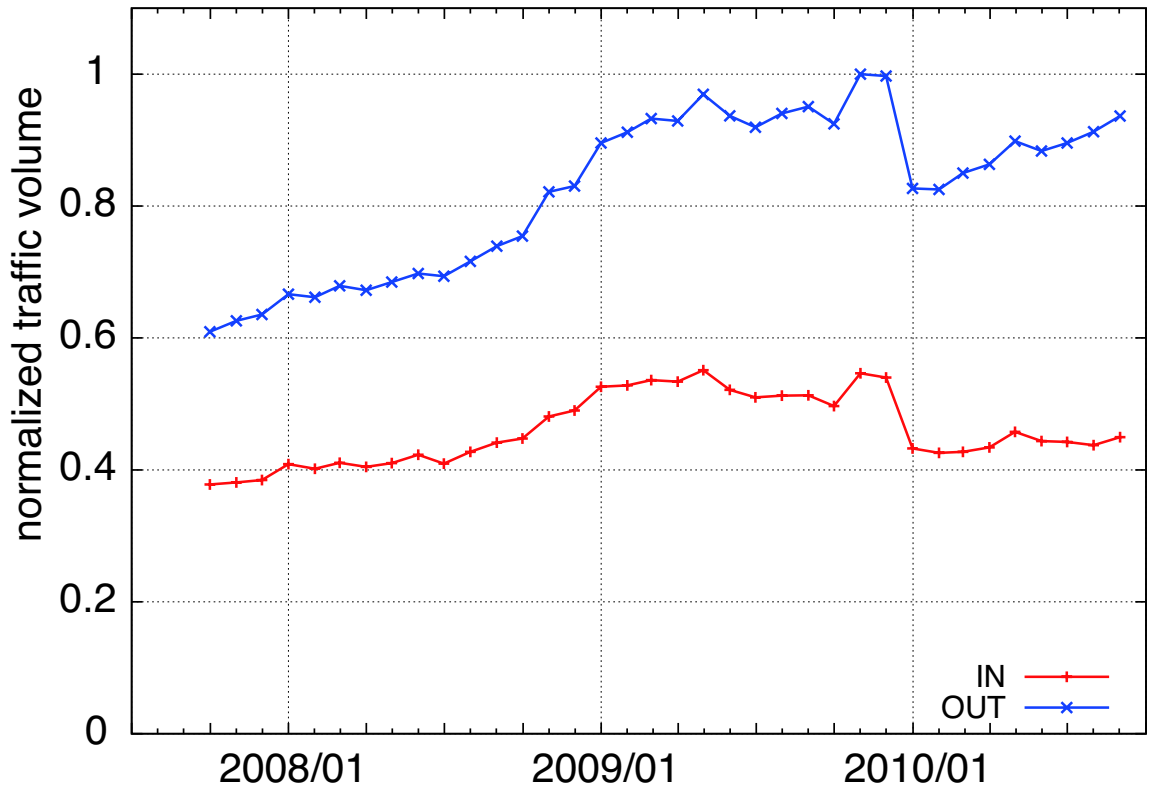
ブロードバンドトラフィックの傾向

- 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

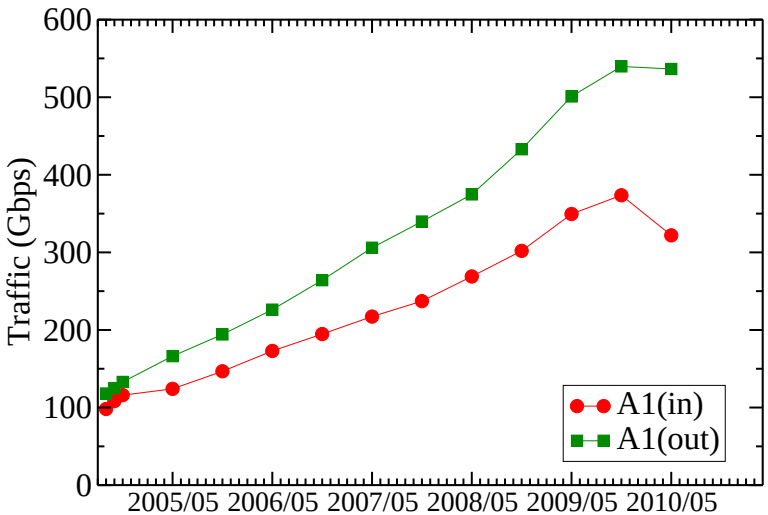


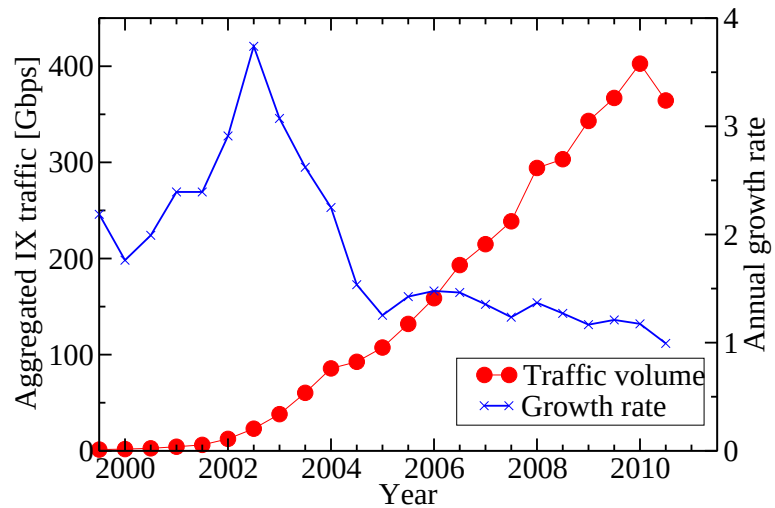
2009年と2010年のデータを比較し原因を探る

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



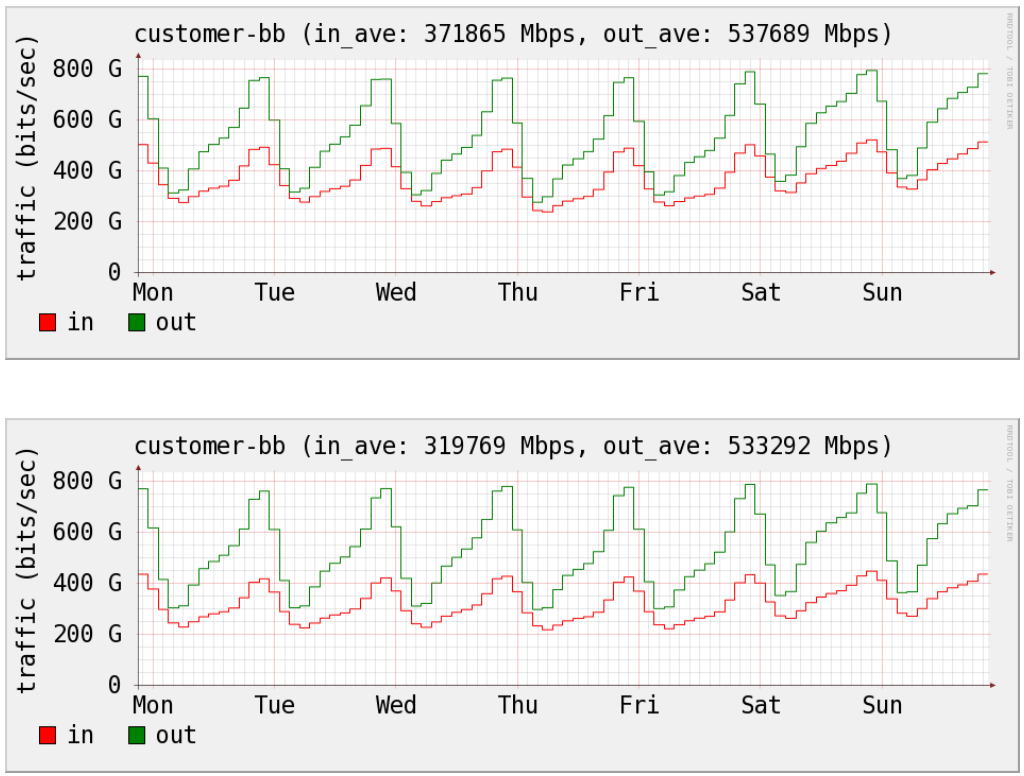


国内 ISP6 社のブロードバンドトラフィック (左) 主要 IX トラフィック (右)

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

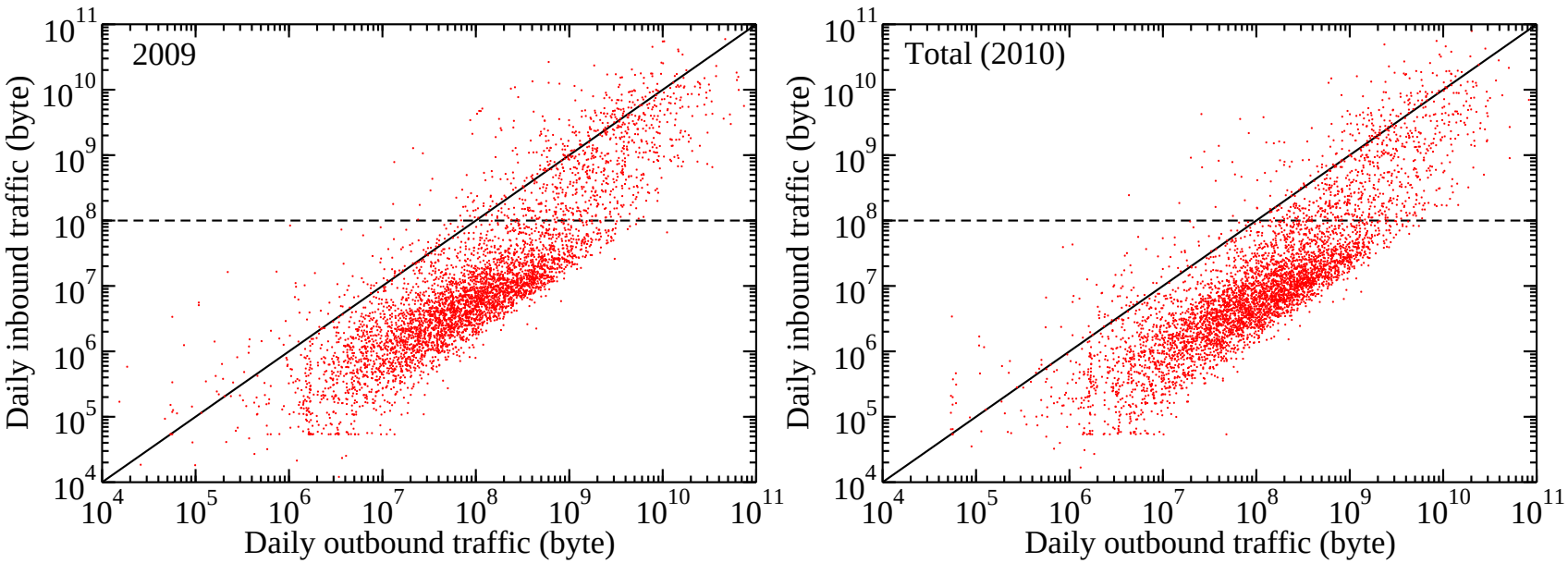
IN/OUT は ISP からの視点

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

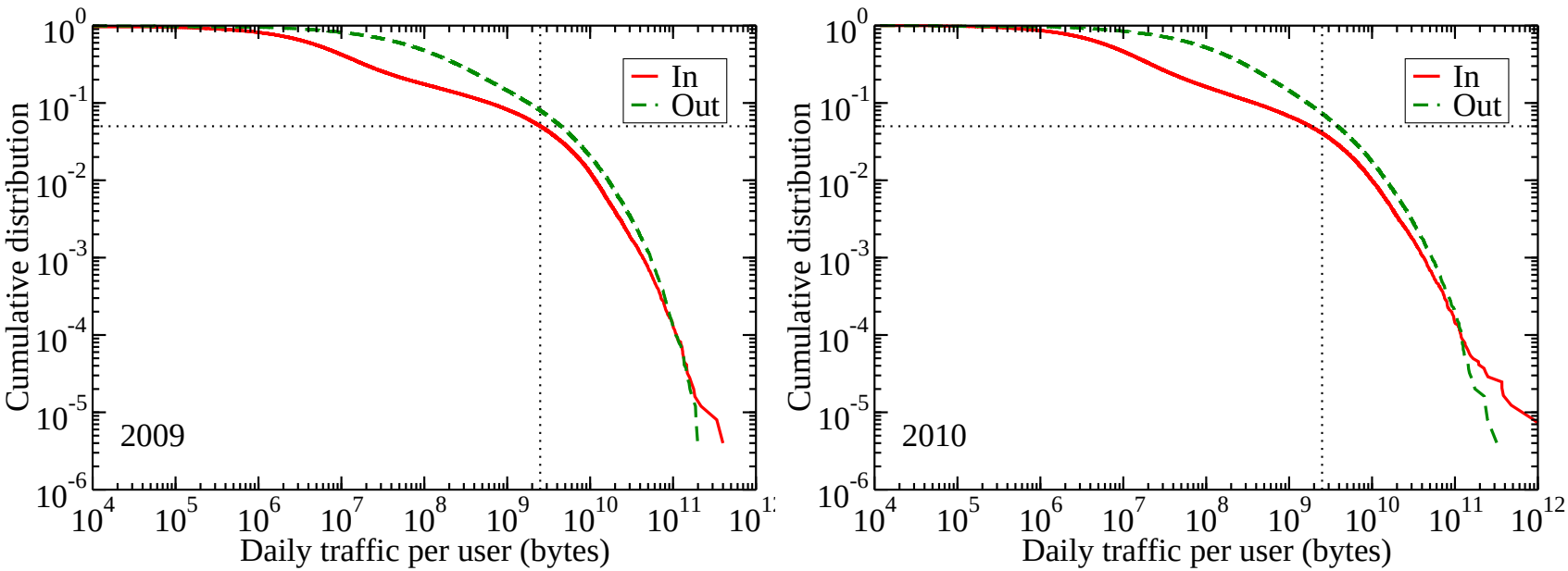
- 境界はあいまい
  - ヘビーユーザとそれ以外、クライアント型とピア型
- 利用者は両タイプのアプリケーションを異なる割合で使用



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

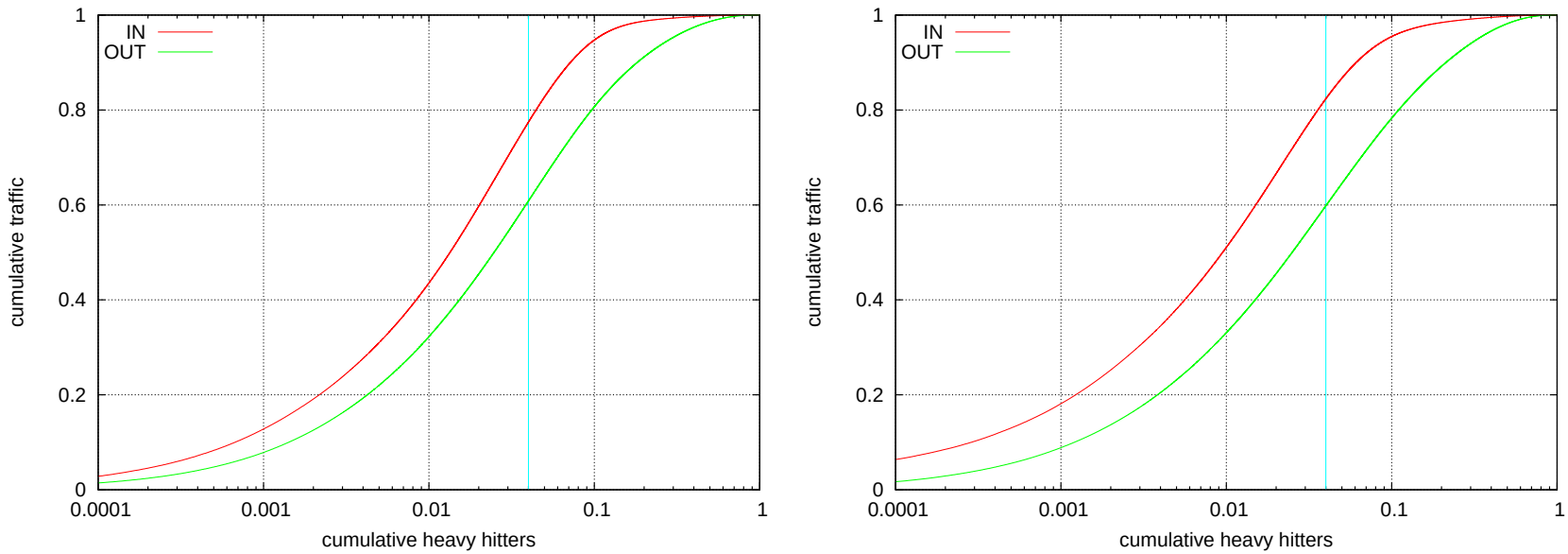
- ベキ分布的 (確率的な分布)
  - 幅広いヘビーユーザが存在
- 2010 年には IN 側でヘビーユーザの割合が若干減少
  - 100MB/日以上アップロードするユーザの総数は 20%程減少
  - 一方で、右端の極端なヘビーユーザは逆に増えている



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

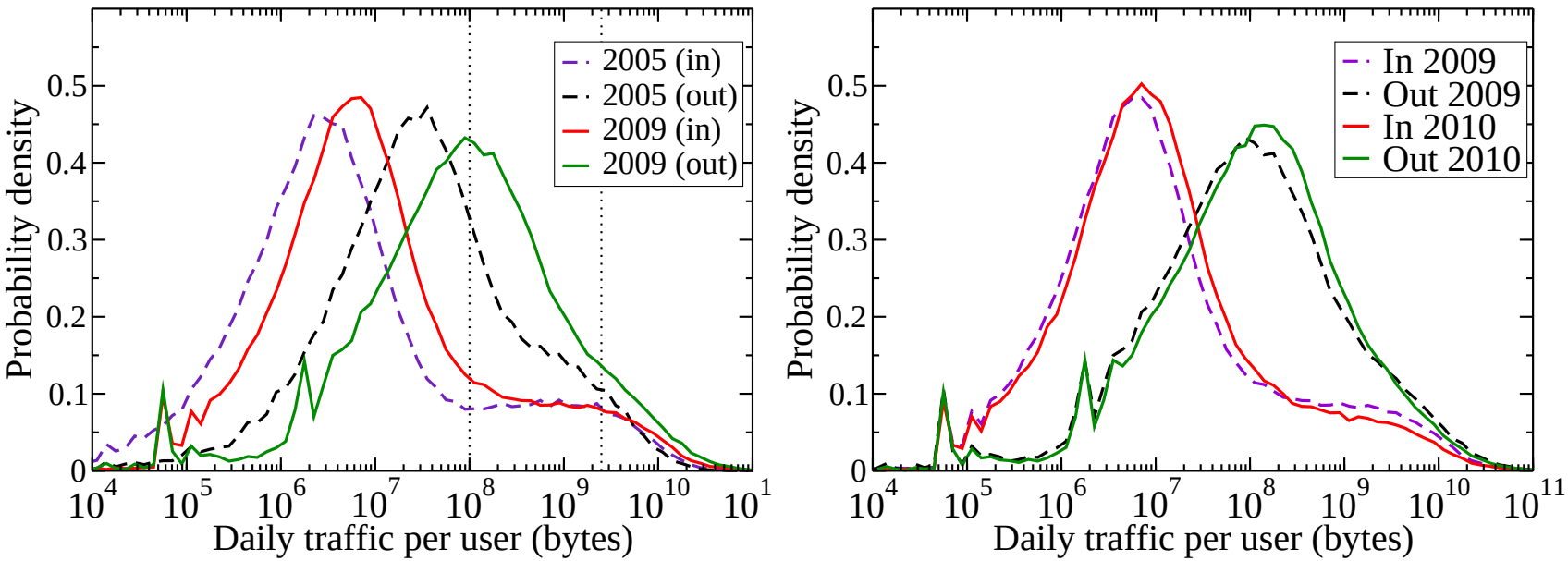


利用者間のトラフィック使用量の偏り (左)2009 (右)2010

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

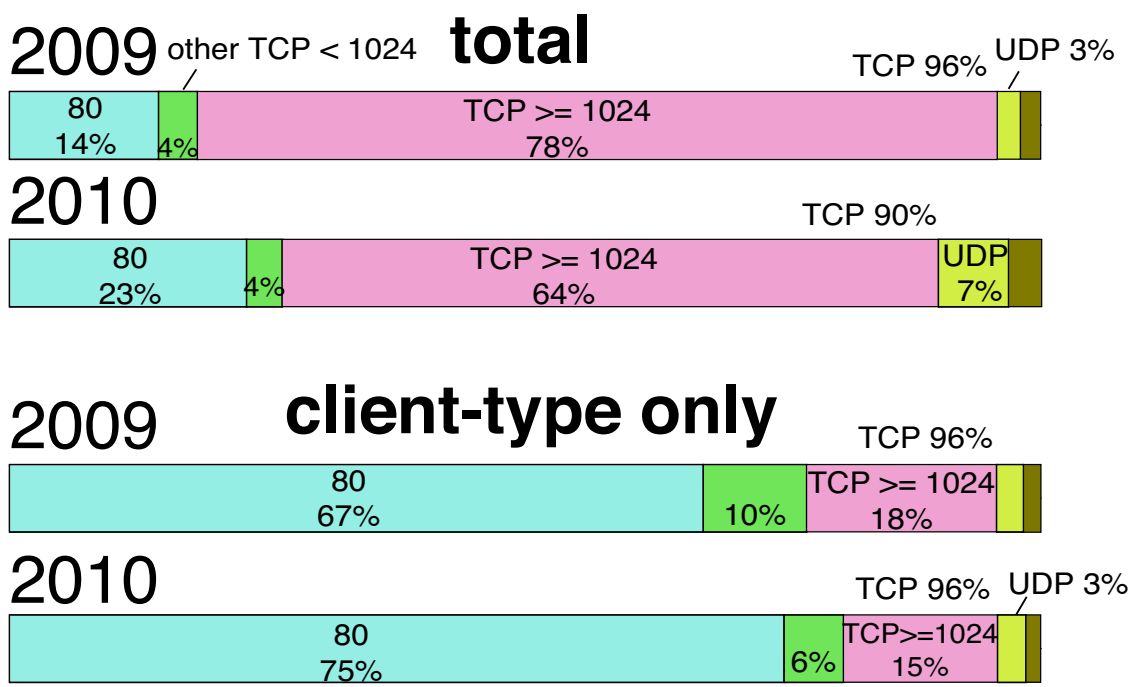


利用者の 1 日の使用量分布 (確率密度関数) (左)2005 と 2009 (右)2009 と 2010

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



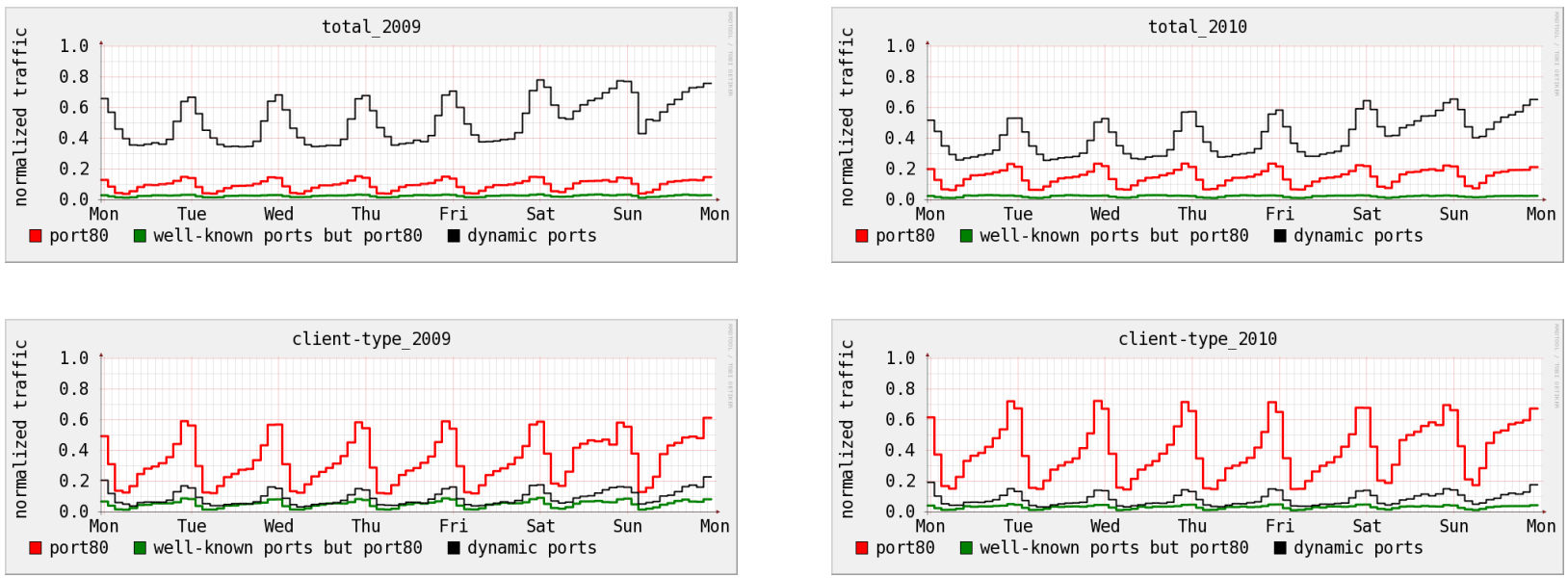
プロトコル別使用量詳細

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

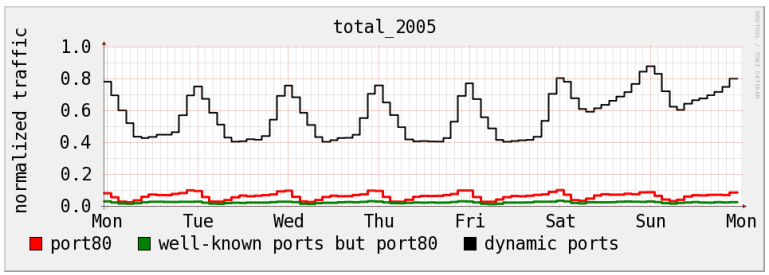
- 合計のピーク値で正規化
- 全体でも動的ポートが減って 80 番のトラフィックが増加
  - これまではクライアント型に顕著な傾向

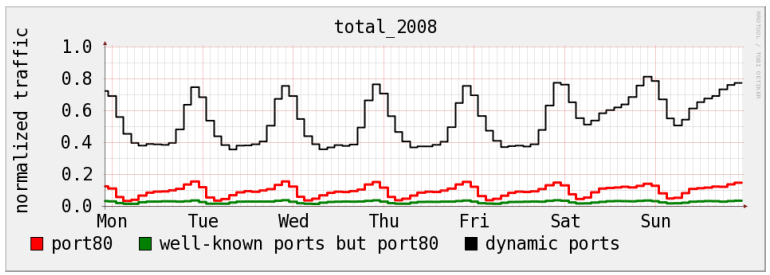


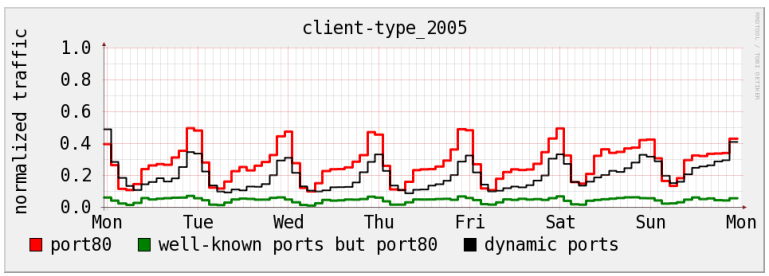
TCP ポート利用の週間推移: (上) 全体 (下) クライアント型 (左)2009 (右)2010

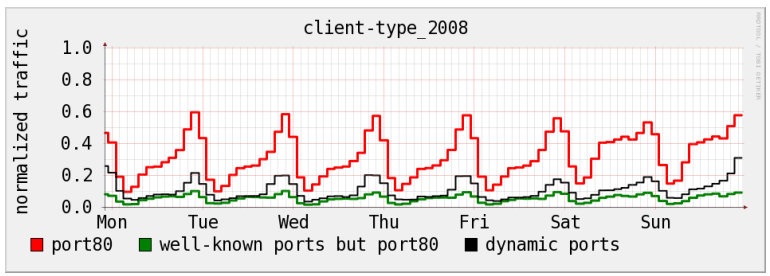
参考: 2005 年と 2008 年の比較

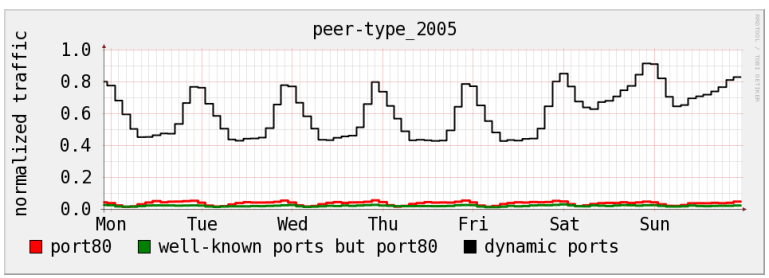
- 全体はピア型ユーザに利用を反映
- クライアント型で 80 番ポートの増加が目立った

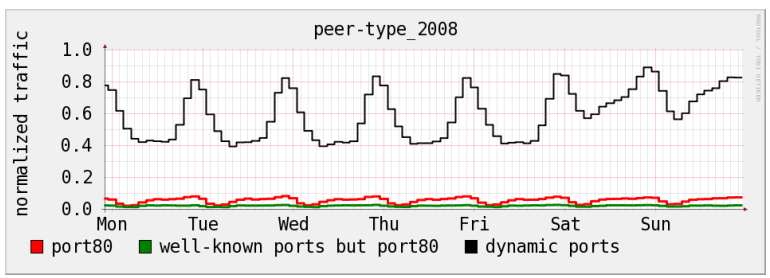












(上) 全体 (中) クライアント型 (下) ピア型 (左)2005 (右)2008

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

インターネット計測とプライバシー

計測はすべての技術の基本
計測情報の開示: 個人情報を含まない統計情報のみ開示可能
計測データからプライバシー情報が漏洩するリスク

- 計測データ中のプライバシー情報 (IP アドレスなど)
- 技術の進歩で情報の拡散や加工が容易になった
- 悪意の利用やリバースエンジニアリングの可能性

技術に法制度がついていけない現状

- ほとんどがインターネット以前に作られた制度
- 計測には法的にはグレーな部分が多い
  - 計測に対する立場の違い、技術者の認識にも大きな温度差

通信の秘密

憲法上の通信の秘密

- 政府など公権力に対する義務

電気通信事業法第 4 条第 1 項で通信の秘密

- 電気通信事業者の取扱中に係る通信の秘密は、侵してはならない

例外

- 当事者の同意がある場合
  - ウイルスチェックサービスや迷惑メールフィルタリングサービス
- 違法性阻却事由が存在し、違法とはされない場合
  - 業務上必要な正当業務行為に当たる場合
    - 例: パケット配送のためにヘッダ情報を見る
  - 緊急避難に該当する場合
    - 例: 他のサービスに支障が出ないよう対策をする

個人情報

個人を識別することができる情報

- 氏名、性別、生年月日、住所、電話番号、家族構成、職業、年収、生体情報
- IP アドレス、メールアドレス、オンライン上の ID、位置情報
- 日本の個人情報保護法　2005 年に施行
  - 5000 件以上の個人情報を扱う事業者が対象
  - 利用目的の特定、制限、適切な取得、通知義務、苦情処理

プライバシー

みだりに自分の私生活を公開されない権利、法的保証
個人の情報を自分でコントロールできる権利
プライバシー情報

- 利用したサービス、web 閲覧履歴、検索履歴、購入商品、趣味指向
- 本人が自ら公開している場合はプライバシー情報とはならない
- しかし、情報の収集、加工、第三者への提供などもプライバシーの侵害になりえる

インターネット計測とプライバシー漏洩リスク

生データ、汎用データ

- 当初の目的以外の利用が可能、いっぽうで情報漏洩リスクを伴う
- 汎用性と情報漏洩リスクのトレードオフ
  - 例えば、特定目的用にオンライン処理することでリスク減少

データの共有、公開

- 共有: 第三者への情報提供となる問題
  - 必要最小限の情報のみ共有するようなデータの加工は可能
- 公開: 幅広い利用促進、悪用されるリスク

商用トラフィックと非商用トラフィック

- 研究教育用ネットワークは比較的計測しやすい
- いっぽうで、商用トラフィックとの乖離

インフォームド コンセント

- 利用者に説明、理解と合意を得るプロセス
- 医療分野で進んでいる (倫理委員会設置など)

法的側面とモラル

- 合法であるかだけでなく、技術者のモラルが問われる
  - センシティブなデータの削除や匿名化

まとめ

インターネットの挙動を計る

- トラフィック量
- (経路情報)
- インターネット計測とプライバシー

次回予定

第 9 回 インターネットの異常や問題を計る (12/11)

- 異常検出
- スパム判定
- ベイズ理論