

インターネット計測とデータ解析 第4回

長 健二郎

2010年10月20日

前回のおさらい

インターネットの構造を計る

- ▶ インターネットアーキテクチャ
- ▶ ネットワーク階層
- ▶ 経路制御
- ▶ トポロジー
- ▶ グラフ理論

今日のテーマ

インターネットの速度を計る

- ▶ 速度計測
- ▶ 利用可能帯域の推測
- ▶ 平均 標準偏差
- ▶ 線形回帰
- ▶ 課題 1

はじめに

速度とは？

- ▶ オブジェクトの単位時間あたりの位置変化

$$v = \frac{\Delta P}{\Delta t}$$

ネットワークの速度

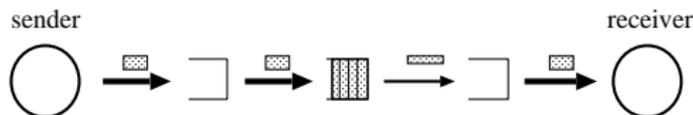
- ▶ 真空中の光の速度: $3.0 \times 10^8 \text{ km/s}$
- ▶ 光ファイバー中の伝播速度: $2.1 \times 10^8 \text{ m/s}$

データ転送レート: 正確には速度というより効率を表す

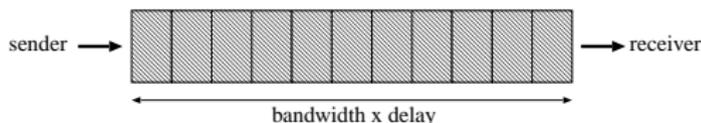
- ▶ ビットレート (bit rate): bits per second
- ▶ 関連用語 (とその混同)
 - ▶ 伝送容量 (bandwidth capacity)、帯域幅 (bandwidth)
 - ▶ スループット (throughput)

スループット計測

- ▶ 速度計測サイト: 実効転送速度を計測
 - ▶ ボトルネックはアクセス回線だと想定
 - ▶ 実際には ISP 境界などもボトルネックになる可能性
 - ▶ クロストラフィックの影響
- ▶ スループット計測ツール
 - ▶ Iperf、netperf、ixChariot など
- ▶ 輻輳: トラフィックの集中で品質低下する状態
 - ▶ 中間ルータのバッファにパケットが溜り遅延が増加

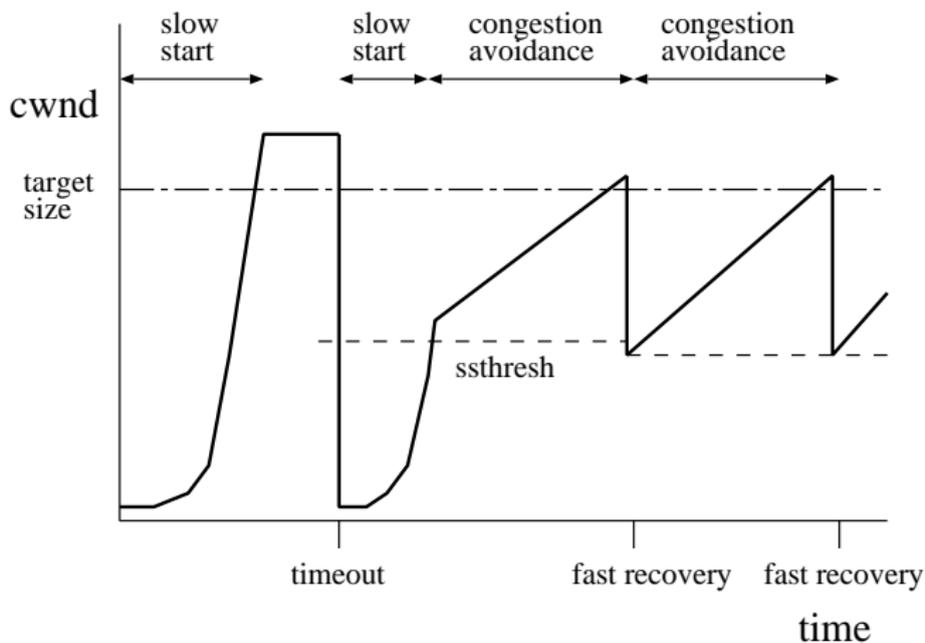


- ▶ TCP によるスループット計測
 - ▶ UDP ではバッファがオーバーフロー
 - ▶ TCP は利用可能な帯域に適應する
 - ▶ TCP バッファサイズ: $\text{delay} \times \text{bandwidth}$



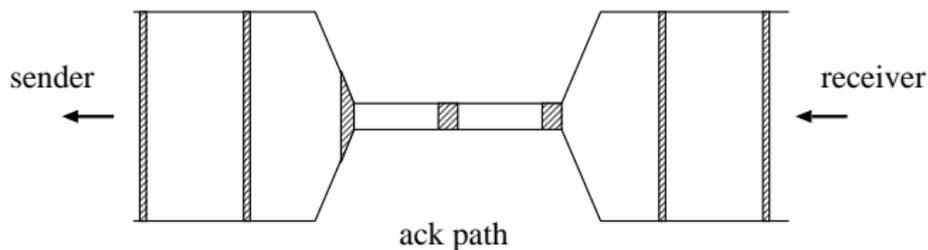
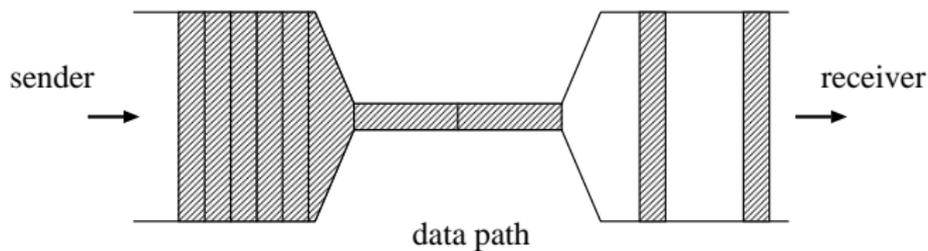
TCP congestion control

- ▶ congestion window により転送中のパケット量を調節
 - ▶ slow start/congestion avoidance
 - ▶ retransmit timeout
 - ▶ fast retransmit/fast recovery



TCP self-clocking

- ▶ ack の到着をトリガーに次のパケットを転送
- ▶ ボトルネック帯域に適應するメカニズム

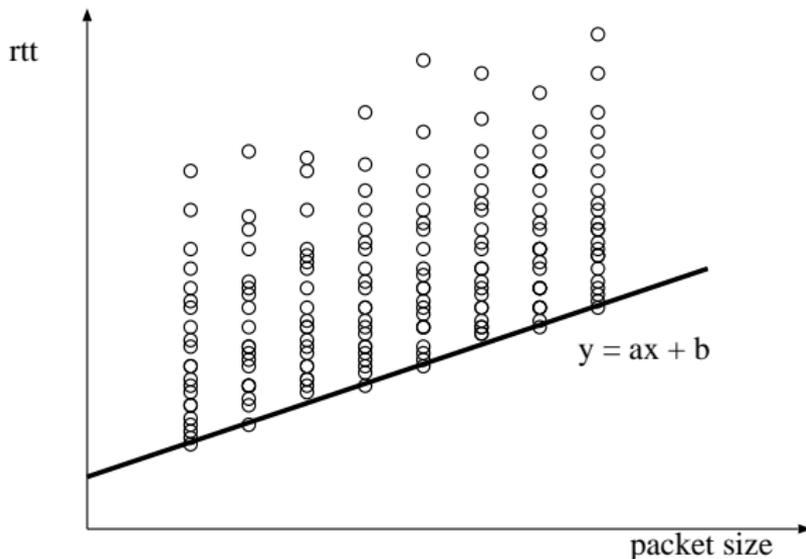


利用可能帯域の推測

- ▶ 回線を埋めることなく、利用可能帯域を推測する技術
- ▶ pathchar アルゴリズム (同様のツールが多数存在)
 - ▶ traceroute と同様に TTL を利用したホップ毎の遅延計測
 - ▶ 異なるパケットサイズで繰り返し計測
 - ▶ 各パケットサイズで最小遅延値を見つける
 - ▶ 線形回帰により伝送遅延と回線容量を推測
 - ▶ 手前のホップとの差分から、その区間の遅延と容量を求める
- ▶ 制約
 - ▶ 繰り返し計測の必要性
 - ▶ 誤差の累積: 特にボトルネック回線の先で精度低下
 - ▶ 一方向しか測定できない

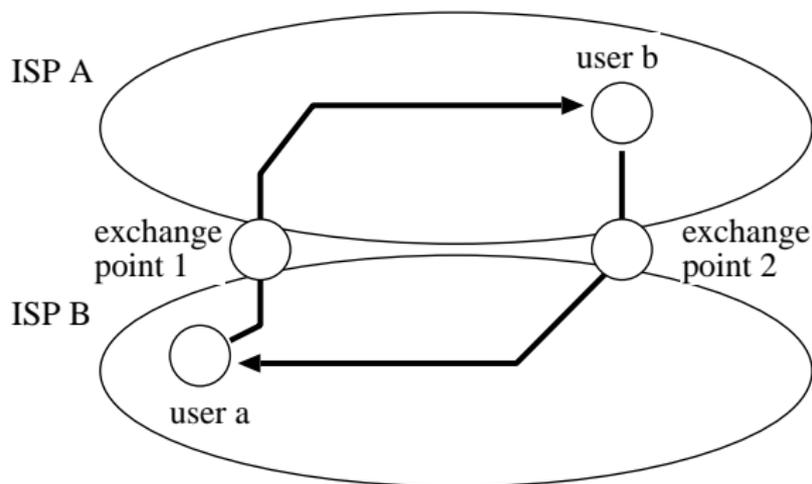
pathchar アルゴリズム

- ▶ 線形回帰による伝送遅延と回線容量の推測
 - ▶ $y = ax + b$
 - ▶ a: RTT-delta/packetsize-delta (bandwidth)
 - ▶ b: delay for packetsize 0 (propagation delay)



非対称な経路

- ▶ 事業者間の非対称経路は一般的
 - ▶ 通常、最も近い接続点で相手事業者にパケットを転送



要約統計量 (summary statistics)

標本の分布の特徴を要約して表す数値

- ▶ 位置を表す数値:
 - ▶ 平均 (mean)、中央値 (median)、最頻値 (mode)
- ▶ ばらつきを表す数値:
 - ▶ 範囲 (range)、分散 (variance)、標準偏差 (standard deviation)

位置を表す数値

- ▶ 平均 (mean):

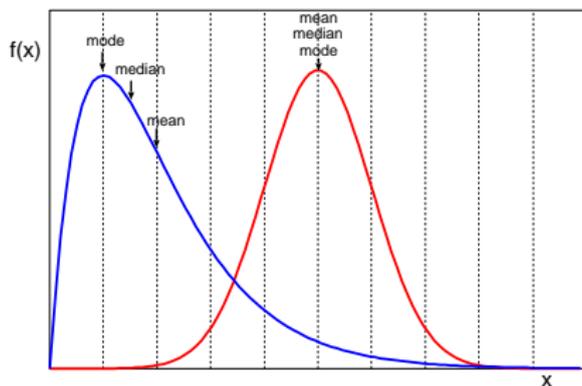
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- ▶ 中央値 (median): データの値をソートして中央にくる値

$$x_{median} = \begin{cases} x_{r+1} & m \text{ が奇数の場合, } m = 2r + 1 \\ (x_r + x_{r+1})/2 & m \text{ が偶数の場合, } m = 2r \end{cases}$$

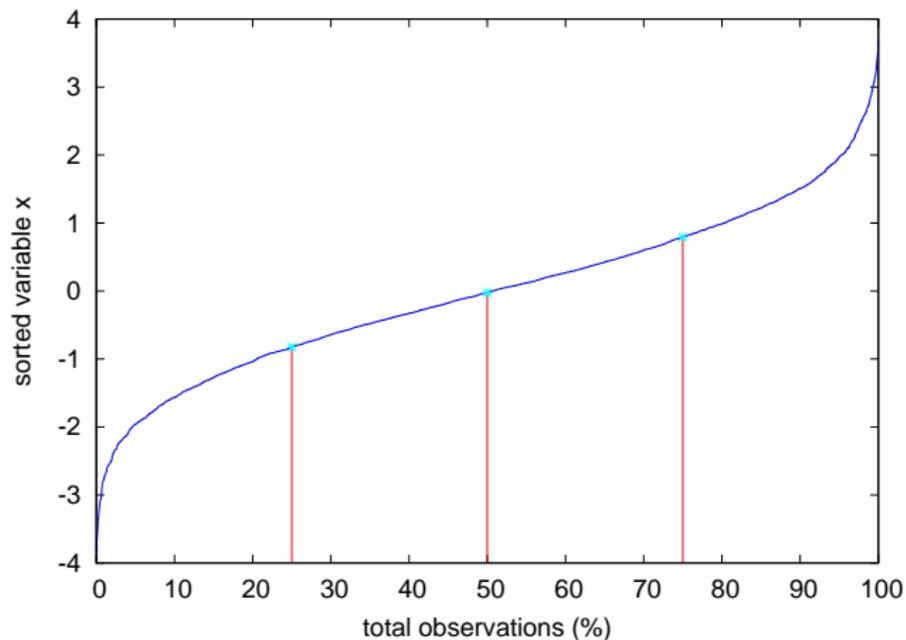
- ▶ 最頻値 (mode): 出現頻度が最も高い値

対称な分布であれば、これらは同一



パーセンタイル (percentiles)

- ▶ p th-percentile: 小さい方から数えて $p\%$ 目の値

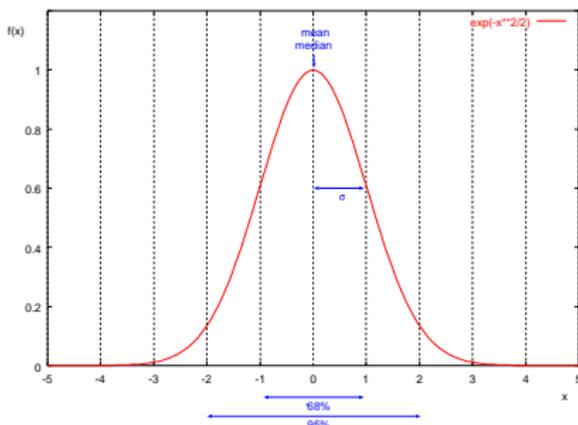


ばらつきを表す数値

- ▶ 範囲 (range): 最大値と最小値の差
- ▶ 分散 (variance):

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

- ▶ 標準偏差 (standard deviation): σ
 - ▶ 平均と同じ次元なので直接比較可能
 - ▶ 統計的なばらつきを示すのに最も良く使われる値
- ▶ 正規分布ではデータの68%は ($mean \pm stddev$)、95%は ($mean \pm 2stddev$) の範囲に入る



相関 (correlation)

- ▶ 共分散 (covariance):

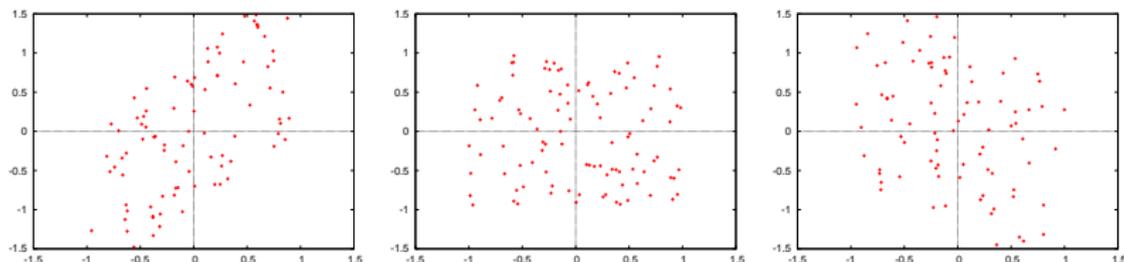
$$\sigma_{xy}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- ▶ 相関係数 (correlation coefficient):

$$\rho_{xy} = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

散布図 (scatter plots)

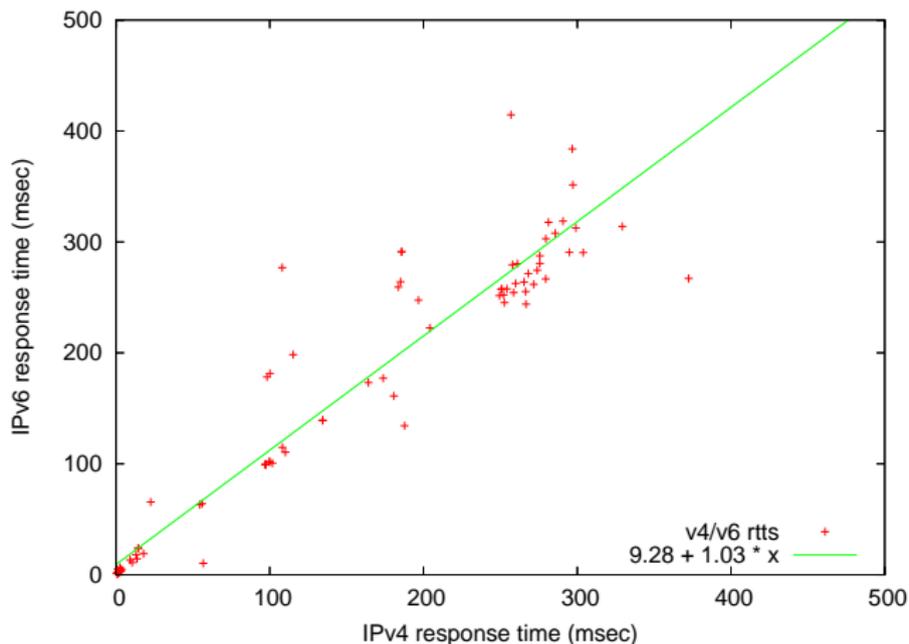
- ▶ 2つの変数の関係を見るのに有効
 - ▶ X軸: 変数 X
 - ▶ Y軸: それに対応する変数 Y の値
- ▶ 散布図で分かる事
 - ▶ XとYに関連があるか
 - ▶ 無相関、正の相関、負の相関
 - ▶ 変数 Y が変数 X に依存して変化するかどうか
 - ▶ 外れ値の存在があるか



例: (左) 正の相関 0.7 (中) 無相関 0.0 (右) 負の相関 -0.5

線形回帰 (linear regression)

- ▶ データに一次関数を当てはめる
 - ▶ 最小二乗法 (least square method): 誤差の二乗和を最小にする



最小二乗法 (least square method)

誤差の二乗和を最小にする一次関数を求める

$$f(x) = b_0 + b_1x$$

切片と傾きの求め方

$$b_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

ここで

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

$$\sum xy = \sum_{i=1}^n x_i y_i \quad \sum x^2 = \sum_{i=1}^n (x_i)^2$$

最小二乗法の導出

i 番目の変数の誤差 $e_i = y_i - (b_0 + b_1 x_i)$ 、 n 回の観測における誤差の平均は

$$\bar{e} = \frac{1}{n} \sum_i e_i = \frac{1}{n} \sum_i (y_i - (b_0 + b_1 x_i)) = \bar{y} - b_0 - b_1 \bar{x}$$

誤差平均が 0 になるようにすると $b_0 = \bar{y} - b_1 \bar{x}$

b_0 を b_1 で表現すると $e_i = y_i - \bar{y} + b_1 \bar{x} - b_1 x_i = (y_i - \bar{y}) - b_1 (x_i - \bar{x})$

誤差の二乗和 SSE は

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n [(y_i - \bar{y})^2 - 2b_1(y_i - \bar{y})(x_i - \bar{x}) + b_1^2(x_i - \bar{x})^2]$$

分散に書き直す

$$\begin{aligned} \frac{SSE}{n} &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 - 2b_1 \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) + b_1^2 \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \sigma_y^2 - 2b_1 \sigma_{xy} + b_1^2 \sigma_x^2 \end{aligned}$$

SSE を最小にする b_1 は、この式を b_1 の 2 次式とみて b_1 について微分して 0 と置く

$$\frac{1}{n} \frac{d(SSE)}{db_1} = -2\sigma_{xy} + 2b_1 \sigma_x^2 = 0$$

$$\text{すなわち } b_1 = \frac{\sigma_{xy}^2}{\sigma_x^2} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2}$$

課題 1

- ▶ 課題: 線形回帰計算と結果のプロット
- ▶ ねらい: 統計処理のプログラミング、プロット作成
- ▶ 提出〆切: 2010年11月10日

- ▶ 詳細は TA の空閑くんから説明

まとめ

インターネットの速度を計る

- ▶ 速度計測
- ▶ 利用可能帯域の推測
- ▶ 平均 標準偏差
- ▶ 線形回帰
- ▶ 課題 1

次回予定

第5回 インターネットの特徴量を計る (10/27)

- ▶ 遅延、パケットロス、ジッタ
- ▶ フロー計測
- ▶ 相関と多変量解析
- ▶ グラフによる可視化