インターネット計測とデータ解析 第11回

長 健二朗

2011年7月13日

前回のおさらい

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

今日のテーマ

インターネットの異常や問題を計る

- 異常検出
- スパム判定
- ベイズ理論

異常とは

- トラフィック異常
- 経路異常、到達性異常
- DNS 異常
- 不正侵入
- CPU 負荷異常

異常原因

- アクセス集中
- 攻撃: DoS、ウィルス/ワーム
- 障害: 機器故障、回線故障、事故、停電
- メンテナンス

異常検出

- サービスの機能低下や停止による損失の回避と低減
- 個別項目の監視: 閾値を越えるとアラート
  - パッシブ
  - アクティブ
- 異常パターン検出:
  - 既知の異常とパターンマッチング
  - IDS: Intrusion Detection System
  - 未知の異常は検出できない
  - パターンを常に更新する必要
- 統計的手法による異常検出
  - 平常時からのずれを検出
  - 一般に「平常」の学習が必要

異常への対応

- 異常を管理者に知らせる
  - 警報通知など
- 異常タイプの識別
  - 運用者が異常原因を把握するための情報提示
  - 特に統計的手法の場合、異常の原因が分かり難い
- 対応の自動化
  - フィルタリングルールの自動生成、サービス切替えなど

異常の具体例

- Flash Crowd
  - サービスへのアクセス集中 (ニュース、イベント、etc)
- DoS/DDoS
  - 特定のホストにトラフィックを集中する攻撃
  - ゾンビ PC が使われる
- scan
  - 多くの場合、脆弱性を持つホストを発見する目的
- worm/virus
  - SQL Slammer, Code Red など多数の事例
- 経路ハイジャック
  - 他人の経路を広告 (多くは設定ミス)

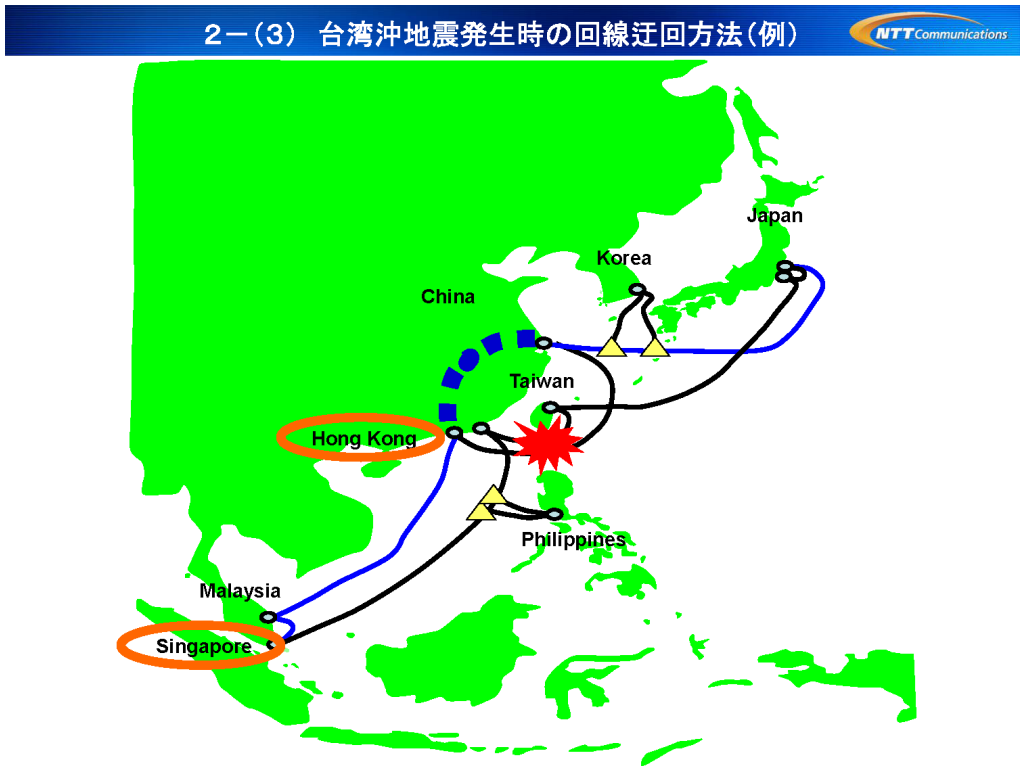
YouTube 接続のハイジャック

- 2008 年 2 月 24 日 世界中の YouTube への接続がパキスタンにリダイレクトされた事件
- 原因
  - パキスタン政府の要請で、Pakistan Telecom が国内から YouTube へ接続できないよう、BGP に YouTube の偽の経路を広告
  - 大手 ISP PCCW が、その経路を外部に伝搬
  - 結果、世界中の YouTube への接続が偽経路によってパキスタンにリダイレクトされた

参考資料:
`http://www.renesys.com/blog/2008/02/pakistan_hijacks_youtube_1.shtml`

台湾沖地震による通信障害の発生

- 2006 年 12 月 26 日台湾南西沖で M7.1 の地震発生
- 海底ケーブルが損傷、アジア向けの通信に障害が発生
- インドネシアでは一時国際向けの通信容量が 20%以下に
- 各 ISP は迂回経路でサービス復旧



出典: JANOG26 海底ケーブル、構築と運用の深イイ話

http://www.janog.gr.jp/meeting/janog26/doc/post-cable.pdf

ISP 間の接続遮断

- Tier1 ISP 同士が接続料金の負担をめぐって争いになった事例
- 2005 年 Level 3 が Cogent 側のトラフィック量が増加していると主張、無償のピアリングを解消し、有償による接続契約の変更を打診
- その他の事例
  - 2008 年 Cogent と Telia がピアリングを解消
  - 2008 年 Level 3 と Cogent がピアリングを解消
  - 2010 年 Level 3 と Comcast が対立し、交渉中

参考資料:
http://www.renesys.com/blog/2006/11/sprint-and-cogent-peer.shtml
http://wirelesswire.jp/Watching_World/201012011624.html

統計的手法による異常検出

参考資料
データマイニングの回にも紹介予定

- 時系列
- 相関
- 主成分分析
- クラスタリング
- エントロピー

スパム判定

スパム: 迷惑メール
判定手法

- 送信者による判定
  - ホワイトリスト
  - ブラックリスト
  - グレーリスト
- コンテンツによる判定
  - ベイジアンフィルタ: スパム判定手法として広く普及
  - 迷惑メールの特徴を統計的な学習手法で分析し判定
  - 学習機能により精度が向上
  - メールからトークン (単語など) を抽出し、含まれるトークンからそのメールがスパムであるかどうか判定

条件付き確率

問題

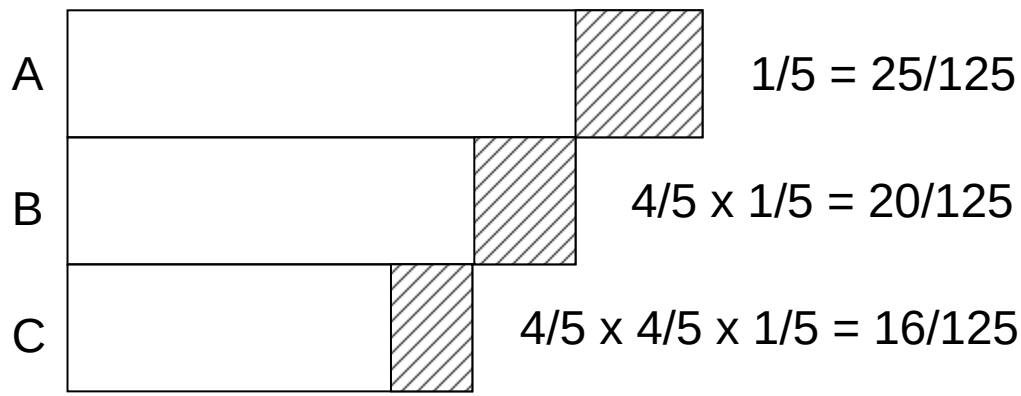
- 5 回に 1 回の割合で帽子を忘れるくせのある K 君が、正月に A、B、C 軒を順に年始回りをして家に帰ったとき、帽子を忘れてきたことに気がついた。2 軒目の家 B に忘れてきた確率を求めよ。 (昭和 51 年 早稲田大入試問題)

条件付き確率

問題

- 5 回に 1 回の割合で帽子を忘れるくせのある K 君が、正月に A、B,C 軒を順に年始回りをして家に帰ったとき、帽子を忘れてきたことに気がついた。2 軒目の家 B に忘れてきた確率を求めよ。 (昭和 51 年 早稲田大入試問題)

解



B で帽子を忘れた確率 / いずれかの場所で帽子を忘れた確率 $= 20/61$

ベイズ理論 (Bayes' theorem)

条件付き確率

- ある事象 A が起こるという条件の下で別の事象 B の起こる確率: $P(B|A)$
  - 全ての場合を事象 A として、そのうち B の起こる事象 $(A \cap B)$ を求める

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

ベイズの定理

- 上記の例とは逆に、A という原因で B が起こったときに、その原因が起こる確率を求める: $P(A|B)$
  - $P(A)$: 原因 A の存在確率 (事前確率)
  - $P(A|B)$: B が起こった場合の原因 A の確率 (事後確率)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$

ベイズ理論の応用

観測結果から原因の確率を推測する: 多くの工学的応用

- 通信: ノイズの加わった受信信号から送信信号を求める
- 医学: 検査結果から実際に疾患を持つ可能性を求める
- スパム判定: 届いたメールの文面から迷惑メールであるか求める

病気検査の例

問題

- ある病気に掛かっている人口割合は 50/1000、この病気の検査は、この病気の患者の 90%が陽性が出るが、患者でない人も 10%は陽性反応がでる。
  あるひとがこの検査で陽性反応が出た場合、本当にこの病気にかかっている確率はいくらか？

病気検査の例

問題

- ある病気に掛かっている人口割合は 50/1000、この病気の検査は、この病気の患者の 90%は陽性が出るが、患者でない人も 10%は陽性反応がでる。
  あるひとがこの検査で陽性反応が出た場合、本当にこの病気にかかっている確率はいくらか？

解

病気にかかっている確率: $P(D) = 50/1000 = 0.05$
陽性反応が出る確率: $P(R) = P(D \cap R) + P(\bar{D} \cap R)$
陽性反応が出た場合、病気である事後確率

$$\begin{aligned} P(D|R) &= \frac{P(D \cap R)}{P(R)} \\ &= (0.05 \times 0.9)/(0.05 \times 0.9 + 0.95 \times 0.1) = 0.321 \end{aligned}$$

迷惑メール判定

- 迷惑メール (SPAM) とそうでないメール (HAM) を用意
- 迷惑メールに多く含まれる単語について
  - SPAM がこの単語を含む条件つき確率
  - HAM がこの単語を含む条件つき確率
- を計算しておき、この単語を含む未知のメールが SPAM である事後確率を求める

例: ある単語 A に関して、$P(A|S) = 0.3$, $P(A|H) = 0.01$, $\frac{P(H)}{P(S)} = 2$ の場合に $P(S|A)$ を求める

$$
\begin{aligned}
P(S|A) &= \frac{P(S)P(A|S)}{P(S)P(A|S) + P(H)P(A|H)} \\
&= \frac{P(A|S)}{P(A|S) + P(A|H)P(H)/P(S)} \\
&= \frac{0.3}{0.3 + 0.01 \times 2} = 0.94
\end{aligned}
$$

単純ベイズ分類器 (naive Bayesian classifier)

- 実際には、複数のトークンを利用
  - トークン同士の組合せを考慮すると膨大なデータが必要
- 単純ベイズ分類器: 各トークンが独立と仮定
  - 独立でない場合でも、実際には有効な場合が多い
  - 学習ステップ:
    - 判定済み学習サンプルから各トークンがスパムに含まれる確率を推定
  - 予測ステップ:
    - 判定が未知のメールに対し、含まれるトークンの推定スパム確率からメールがスパムである事後確率を計算、判定
- 学習ステップはトークン毎に独立計算なので簡単
- トークンスパム確率から結合スパム確率の算出にベイズの結合確率を使う

単純ベイズ分類器 (もう少し詳しく)

トークンを $x_1, x_2, \ldots, x_n$ とする。 これらが出現したとき SPAM である事後確率は

$$P(S|x_1, \ldots, x_n) = \frac{P(S)P(x_1, \ldots, x_n|S)}{P(x_1, \ldots, x_n)}$$

分子の部分は、これらのトークンが出現し、かつ SPAM である同時確率なので、以下のように書け、条件つき確率の定義を繰り返し適用すると

$$\begin{aligned} P(S, x_1, \ldots, x_n) &= P(S)P(x_1, \ldots, x_n|S) \\ &= P(S)P(x_1|S)P(x_2, \ldots, x_n|S, x_1) \\ &= P(S)P(x_1|S)P(x_2|S, x_1)P(x_3, \ldots, x_n|S, x_1, x_2) \end{aligned}$$

ここで、各トークンが条件付きで他のトークンと独立だと仮定すると

$$P(x_i|S, x_j) = P(x_i|S)$$

すると上記の同時確率は

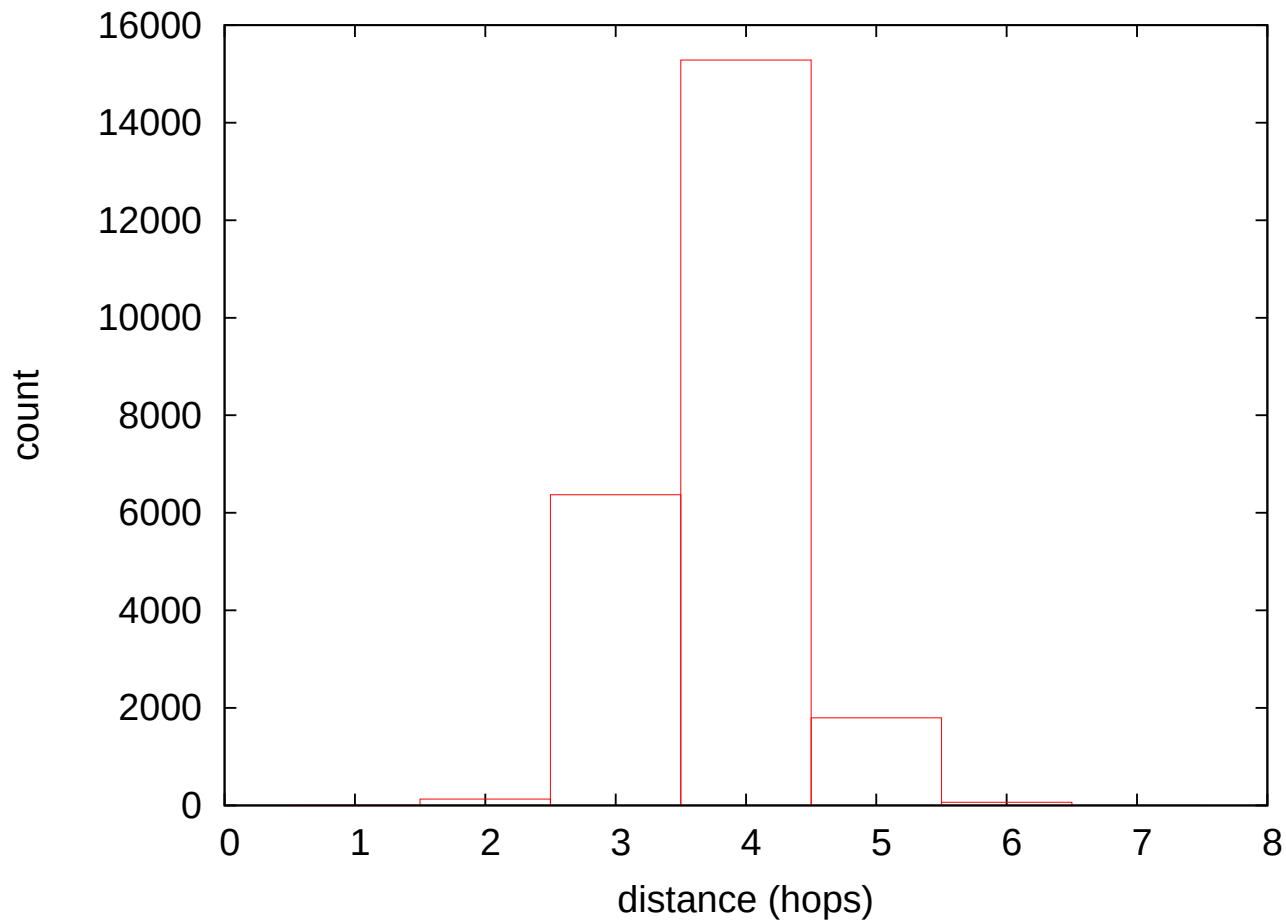
$$P(S, x_1, \ldots, x_n) = P(S)P(x_1|S)P(x_2|S)\cdots P(x_n|S) = P(S)\prod_{i=1}^{n} P(x_i|S)$$

したがって、各トークンが独立だとの仮定の下で、SPAM である事後確率は

$$P(S|x_1, \ldots, x_n) = \frac{P(S)\prod_{i=1}^{n} P(x_i|S)}{P(S)\prod_{i=1}^{n} P(x_i|S) + P(H)\prod_{i=1}^{n} P(x_i|H)}$$

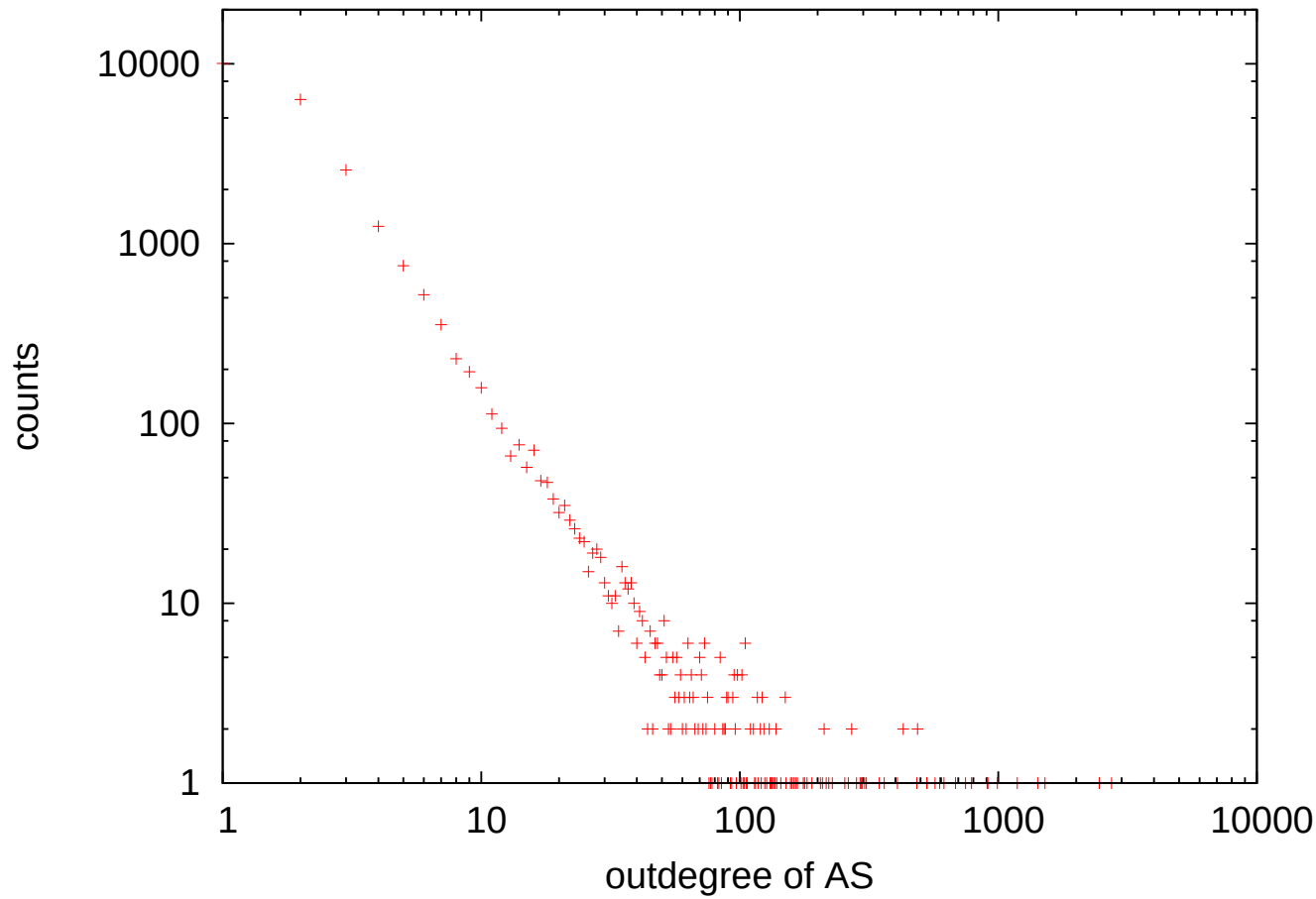
課題 2: 最短経路木の計算、距離の分布と次数分布のプロット 解答

慶應 (38635) から他の AS への距離 (ホップ数) の分布のプロット



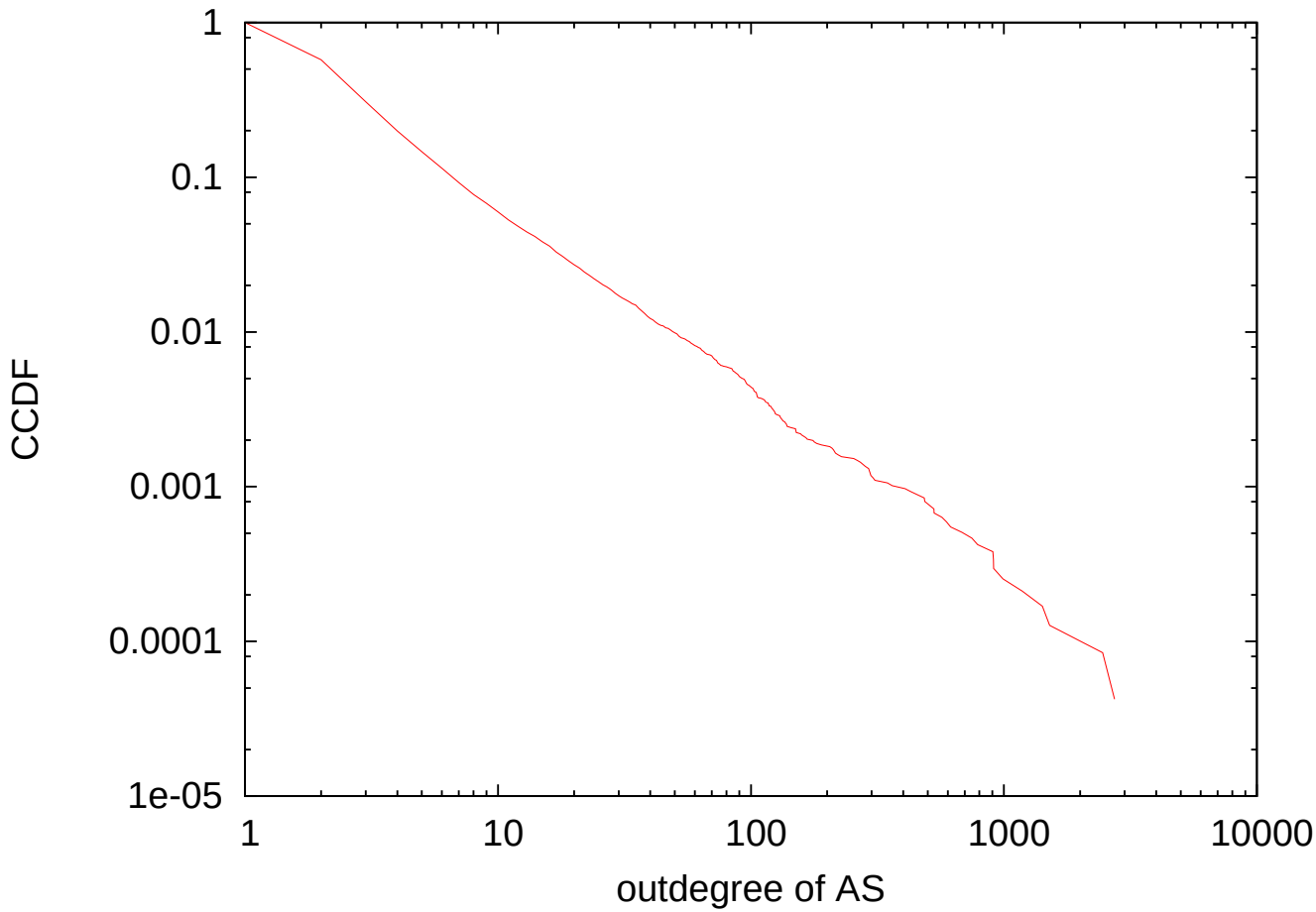
課題 2 解答

AS の次数分布の散布図



課題 2 解答

AS の次数分布の CCDF プロット



最終レポート

課題 A 追加参考資料

- 訪問パターン解析 by SA 加藤さん

まとめ

インターネットの異常や問題を計る

- 異常検出
- スパム判定
- ベイズ理論

次回予定

第 12 回 データマイニング (7/20)

- パターン抽出
- クラス分類
- クラスタリング
- 演習:クラスタリング