アルゴリズムとプログラミング実践講座

http://akashi.ci.i.u-tokyo.ac.jp/mary/lectures/algorithm/

火曜　13:00 -- 14:30
I-REF 棟 6階　ヒロビー

稲葉真理　with 浅井大史・手塚宏史

先週のレポートから
二次元 random walk　の分布
ヒートマップ

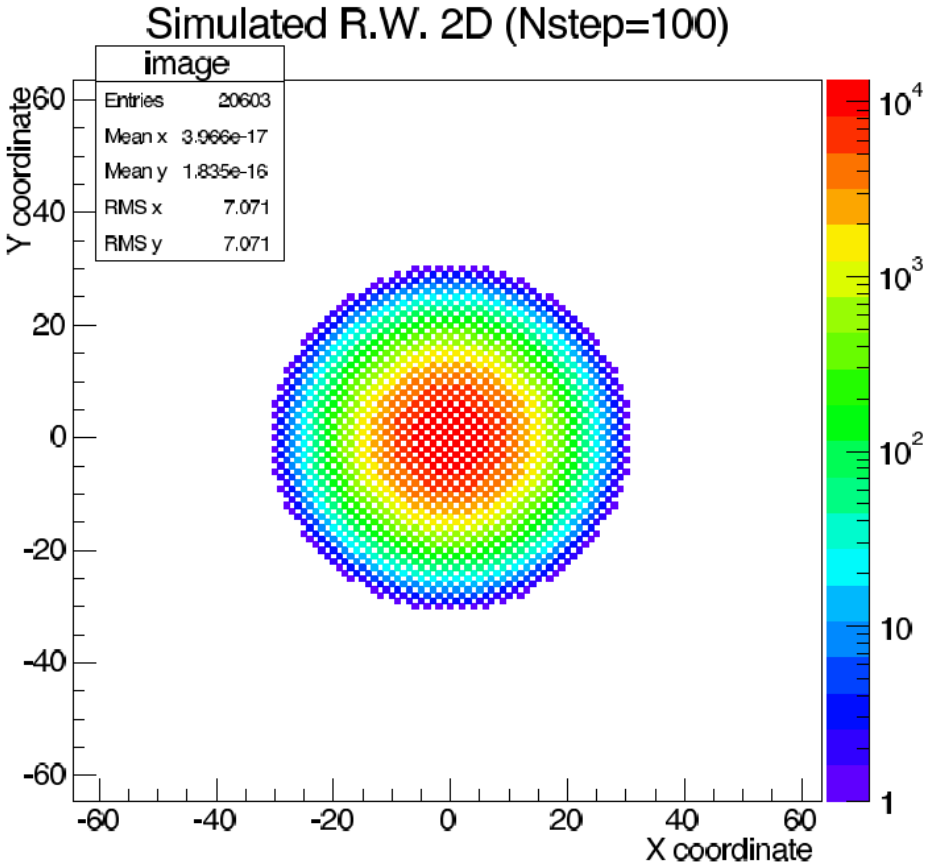


図 8: 式 4 により求めた $P_{100}(x, y)$ に試行回数をかけたもの。

先週のレポートから
二次元 random walk　の軌跡
アニメーション ２つ

animation.py (Python)

http://d.hatena.ne.jp/xef/20120629/p2
を参考にして作成したそうです。

Walk2Dcanvas.html　(JavaScript)

先週のふりかえり

現実の世界を反映するような
数理モデルが欲しい

現実世界の（役に立つ）
ネットワークの３つの特徴
（これも多数決　←　便利と思う人が多い）

- スモールワールド性
  → 任意の２ノードの平均距離は短い

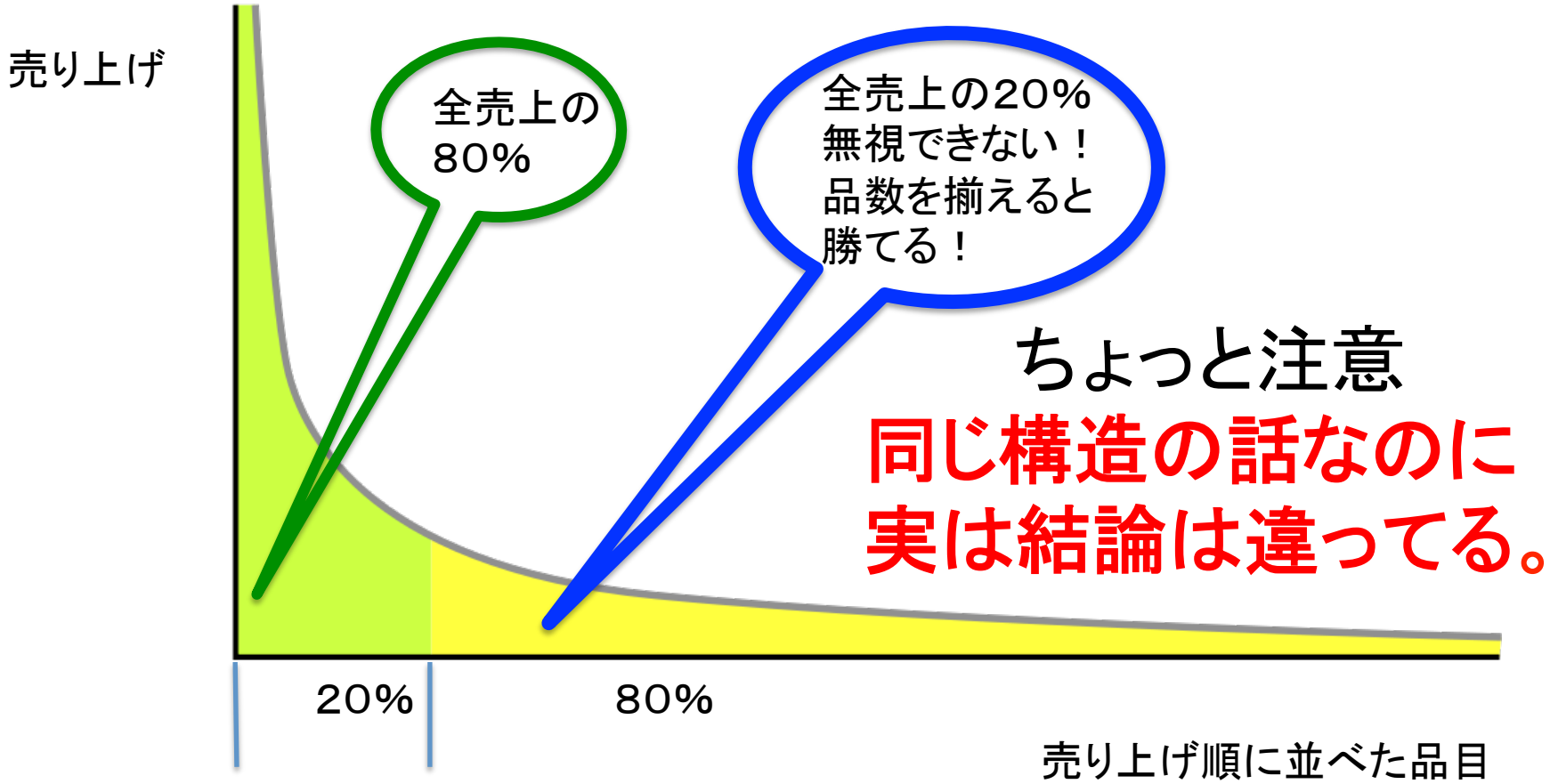
- スケールフリー性
  → ノードの次数がベキ乗則

- クラスター性
  → あるノードと、隣接ノードの隣接ノードとの間は
  エッジがある確率が高い

べき乗則

選択と集中

amazon など　Long　Tail



ちょっと注意

同じ構造の話なのに
実は結論は違ってる。

グラフは、http://ja.wikipedia.org/wiki/ロングテール　より

べき分布

- 金持ちが金を稼げる
- 平均は意味をもたない
- 横軸、縦軸ともに対数にすれば直線
- 成長していくときに、構造が維持される

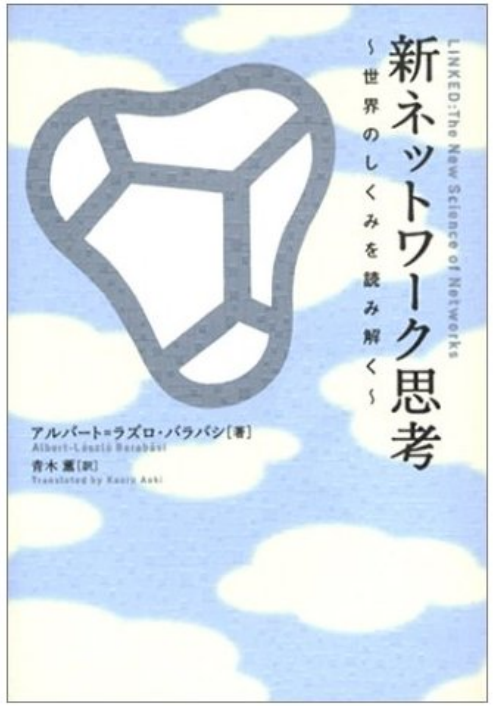
- スケールフリーと名前をつけた　←　はやった
- 話題になっているベキ分布は、-2 ～-3乗

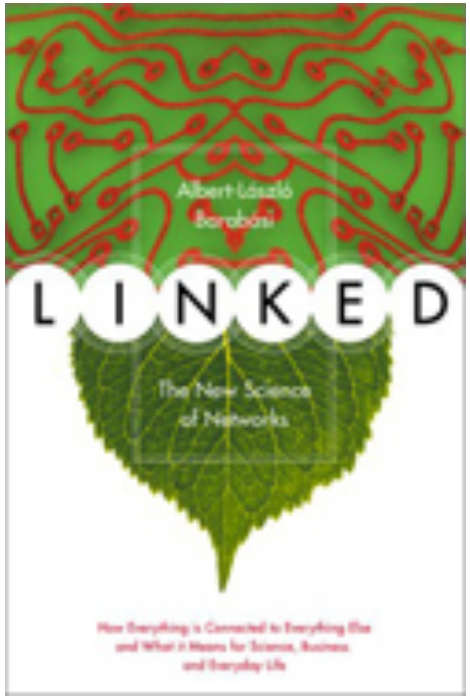
先週の補足 （本日のタネ本）

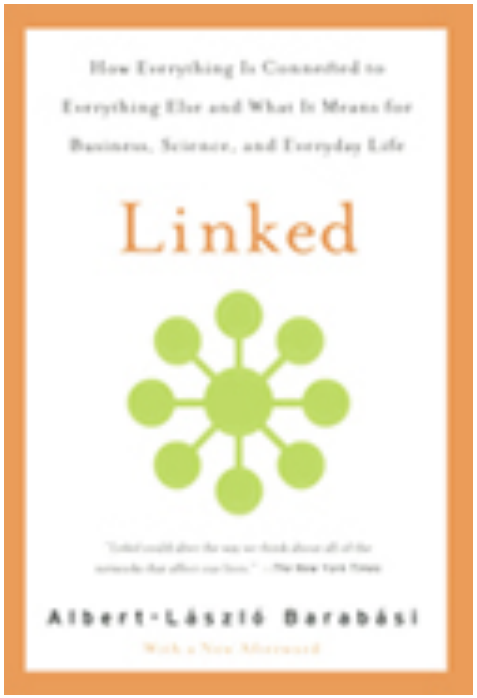
新ネットワーク思考

LINKED: The New Science of Networks

Albert-Lazlo Barabasi

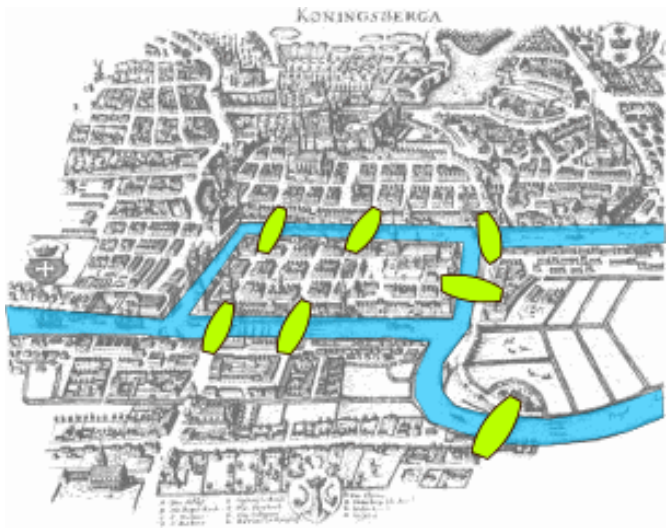


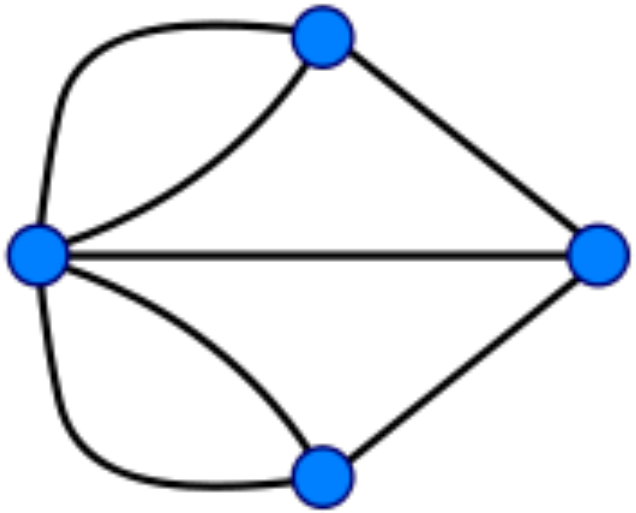




グラフ理論の歴史

1736年　オイラー　（ケーニヒスベルグの橋）





20世紀までのグラフ理論の目標：
さまざまなグラフを見つけて、性質を明らかにする

ランダムグラフ （1959 年が最初）
エルデシュ と レーニィ

今までと違うグラフ理論

ネットワークの形成 がテーマ

現実世界での
ネットワークは「ランダム」と仮定

注： エルデシュとレーニィにとっては、
数学的に面白いことが大事だった

ランダムグラフの性質

- ノードの平均次数が1をこえると、巨大クラスターができる

- ランダムグラフの次数は　ポアソン分布
（ピークの両側で分布は急速に減少）

→　ほとんどのノードは平均次数

弱い絆の強さ [Granovetter ’72]
(The strength of weak ties)

社会学

就職した時に力になったのは、親しい友人ではなく、「ちょっとした知り合い」だった

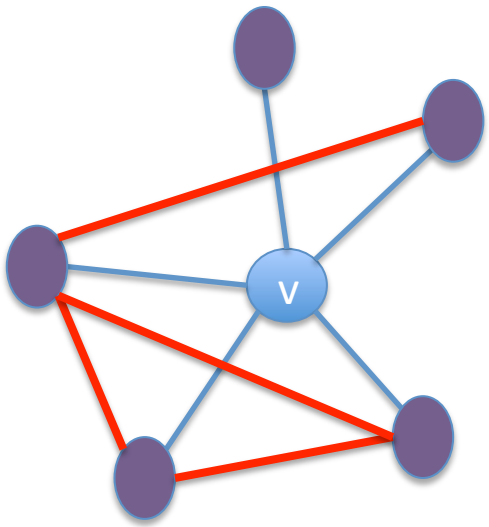
社会では、人はクラスターに属し、

外の世界とのコミュニケーションは、「弱い絆」に支配されている

社会システムの中には
クラスター構造が存在する

「自分の友達と、自分の友達は、けっこう、友達同士のことが多いよね？」

ある頂点に着目



頂点　v のクラスター係数 Cv

$$= \frac{\text{vの隣接ノード間のエッジの総数}}{\text{完全グラフの場合のエッジ数}}$$

グラフの　クラスター係数　C

Cv　の平均

WS モデル

秩序の高いクラスターが存在し
少数の（ランダムな）遠距離リンクで
互いに結ばれている

→　ものすごく次数の高いノード　（ハブ）は
　　めったなことでは生まれない。
　　（バラバシの web 地図と矛盾）

ケヴィンベーコンゲーム

「ハリウッド俳優は　共演をたどると、２～３でベーコンに到達」 (John Stuart Show)

→　TV　見ていた学生が

IMDb.com (Internet Movie Data base) を利用して

１週間で　作成

「The Oracle of Bacon」 →　人気サイトに

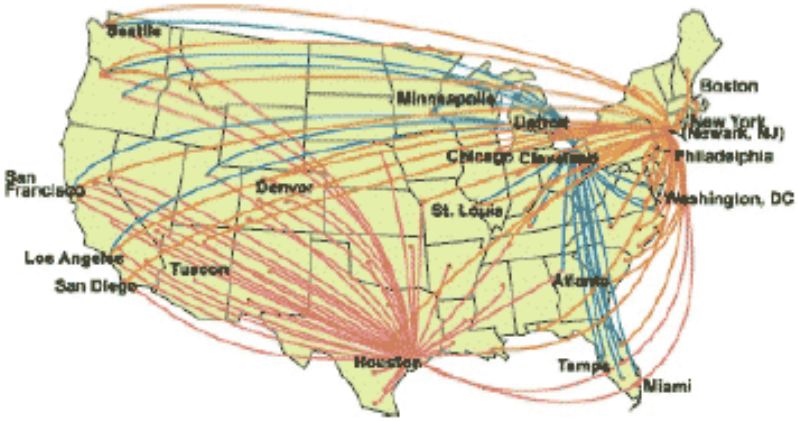
Observation

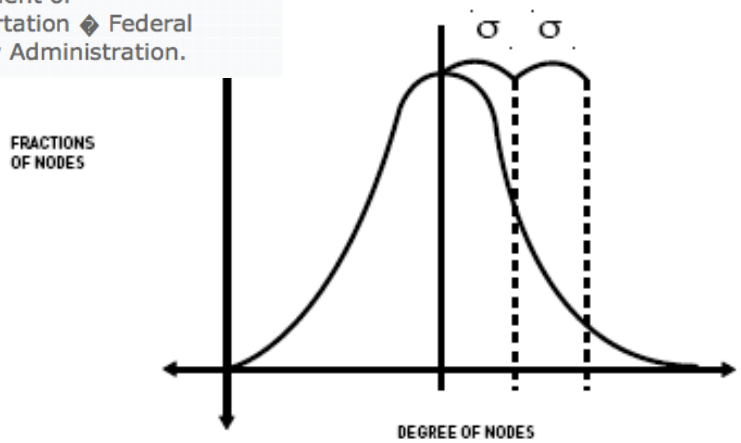
- ベーコンだけが特別リンクが多いわけではなかった。
- ハリウッドの俳優どうしだと　たいてい　３以下でつながる
- 出演数が多くても、ハリウッドの中心に近いとは限らない

ランダムネットワークと
スケールフリーネットワークの例

ハイウェイ

Item 3128 / United States Department of Transportation ◆ Federal Highway Administration. NATIONAL HIGHWAY SYSTEM MAP (2002). Courtesy of United States Department of Transportation ◆ Federal Highway Administration.

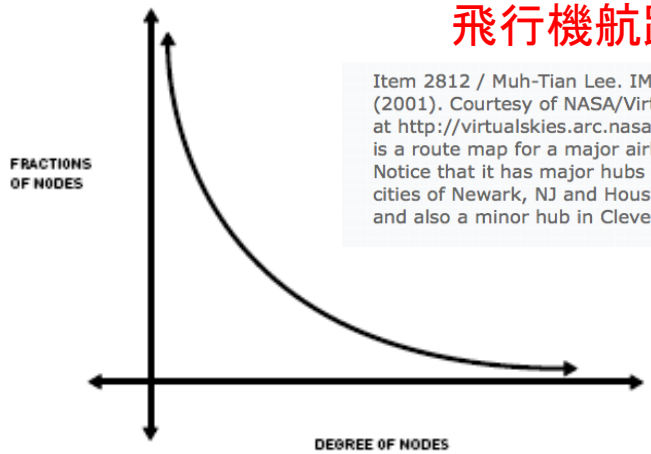




Binomial Distribution

飛行機航路

Item 2812 / Muh-Tian Lee. IMAGE (2001). Courtesy of NASA/Virtual Skies at http://virtualskies.arc.nasa.gov. This is a route map for a major airline. Notice that it has major hubs in the cities of Newark, NJ and Houston, TX and also a minor hub in Cleveland, OH.



http://www.learner.org/courses/mathilluminated/units/11/textbook/05.php　より

べき則

- 相転移と繰り込み群 （私には理解不能）
  無秩序が秩序になる臨界点の近くはべき則
- WWW
- ベーコン（俳優）
- チップ配線

Observation
べき則があるところに、ハブがある

関係は？

成長するネットワーク

新しいノードができるとき、
ランダムに、既存のノードに接続

→　べき則にならなかった
→　旧いノードがリンクを獲得

BAモデル （スケールフリーネットワーク）

（小さい）完全グラフから始める。
時刻 t のネットワークがに対して、時刻 t+1 で
A. 成長
　ノードを一つだけ増やす
B. 優先的選択
　新しいノードを既存のノードとニリンクで結ぶ。
　結ぶ先のノードは、リンク数に比例して選ぶ

→ ベキ則

スケールフリーネットワーク

- 金持ちはもっと金持ちに
- 進化（成長）するネットワークのモデル

WWW,
ハリウッド、
細胞内代謝ネットワーク、
経済ネットワーク、・・・

ところで・・・（はなもげら）

Scale Free ネットワークで、
新参者が、なぜ、勝てるのか？

具体的には、Google　がなぜ？

→　**適応度モデル**
　　「ボーズ・アインシュタイン凝縮」
　　単一原子理想気体の量子論（だそうです）
　　ネットワークとボーズ気体は同じ振る舞い

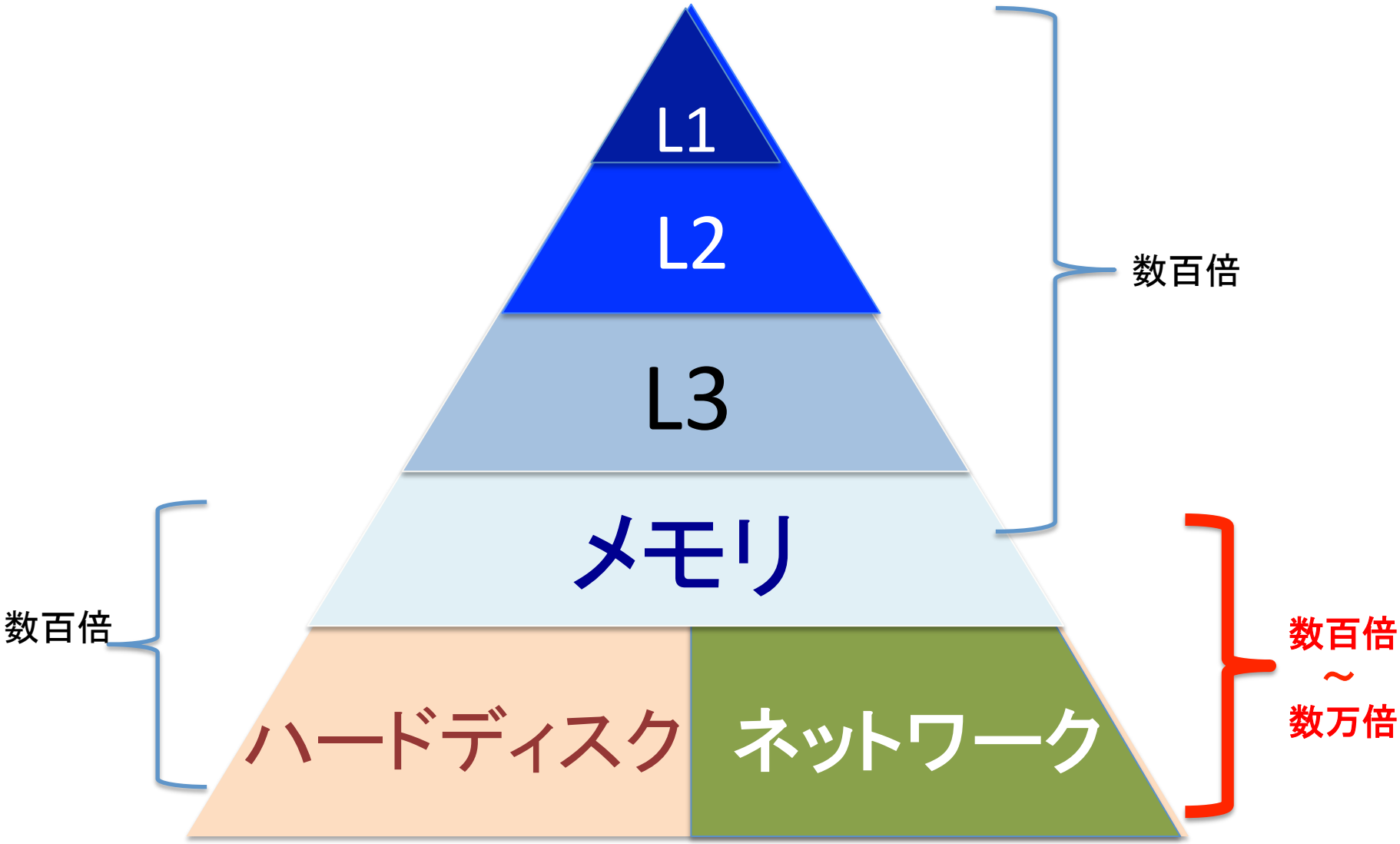
「一人勝ち」する物がいる　(スケールフリーじゃなくなる，MS）

1999　ファルートス３兄弟

(インターネットの）ルータの構成は
スケールフリー。

それまで、 ランダムネットワークで
モデル化していたものは、間違っていた。

インターネットもべき則



the Internet

すごいなと思うもの

IP　の世界

the Internet→　IPという風呂敷 (by 村井純）

（ether, serial , FDDI, ATM, 全部包む)

The Internet はなぜはやったか？

- さまざまなプロトコル
  - DECnet
  - SNA (System Network Architecture) IBM
  - XNS (Xerox) --- 仕様書 $100
  - FNA (Fujitsu)
  - HNA (Hitachi)
- Gateway を作ってプロトコル変換していた。
- internet --- network をつなぐ （おりじなる）

なにがすごいか？

- 自律分散

  決まってるのはプロトコルだけ

  郵便・電話・放送と違う　ー　法律・国

- 寿命の長さ

  感想

  - なんとかなるもんだなぁ

ちょっと歴史の話

1957 ARPA 設立(DOD)

1966 ARPANET プラン

1968 ARPA　56Kbps パケット交換

1969 4 ノード稼動開始

1972 telnet の仕様

1973 ftp の仕様

1975 TCP protocol

1975 UNIX

1983 ARPANET を分割。(オペレーションコスト)

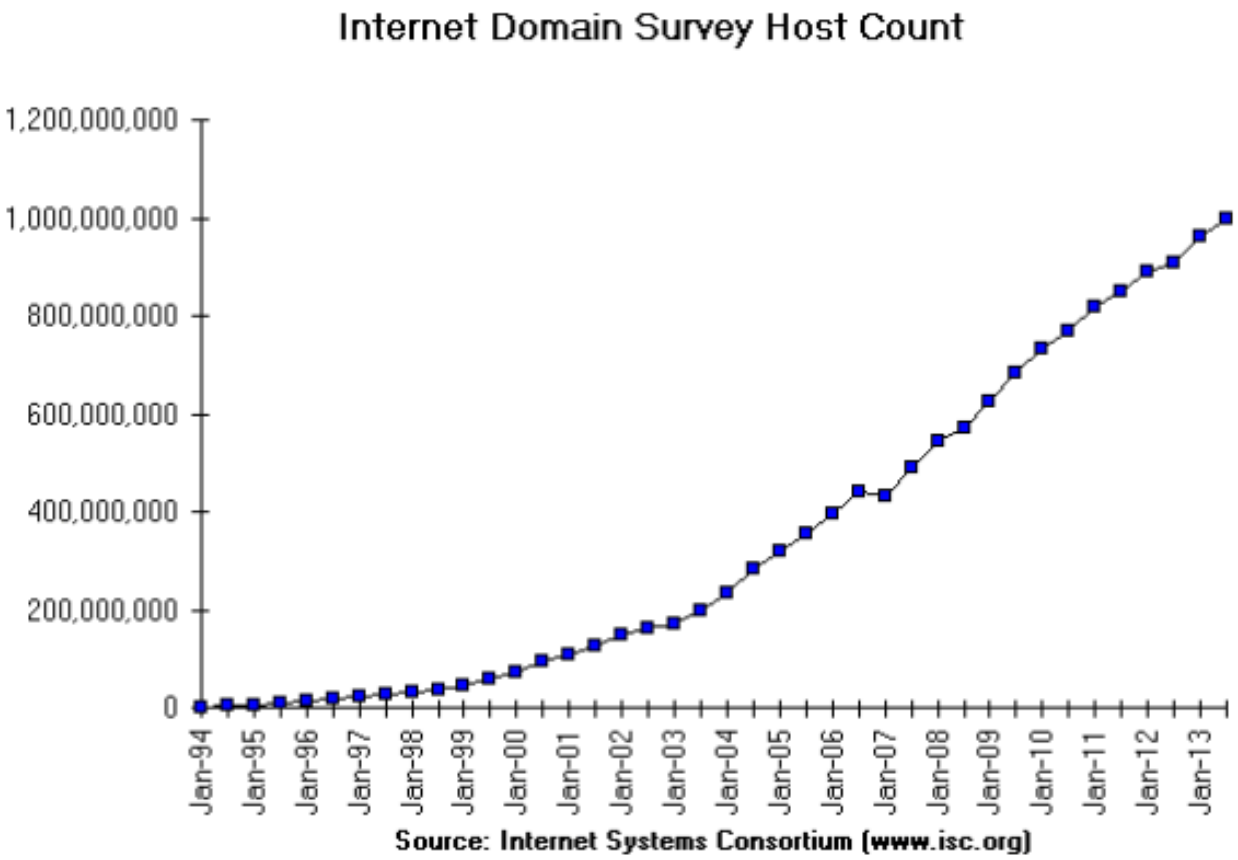
1987 UUnet (ISP)

1992 Web

1992 ISOC（Internet Society　→　USからの独立）

接続ホスト数

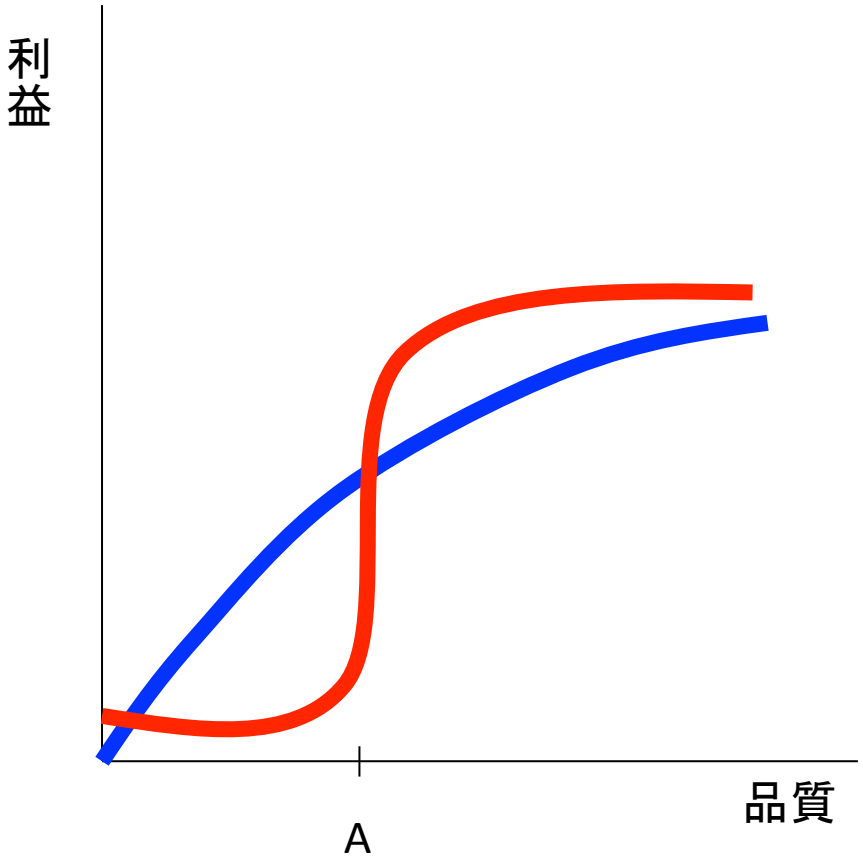
1969	4
1984	1,000
1988	20,000
1989	10 万
1995	480 万
2000	5600万



電話網との違い

- 電話は国（あるいは国に認められた独占企業）が管理。（ある地域の設計主体は一つ）
- 品質に対する要求
- 経路制御
- 電話では端末（電話機）はすることがほとんどない。

Social Optimality



電話：　品質が悪すぎると、
お話にならない。

B/A以上の利用は許容しないほうが全体の利益が大きい。

（B:回線容量）

（古い）internet

使いたい人がいれば使いたいだけわける。

Best-Effort

電話網との違い

- 全体を管理する人がいない。

  関係者のハードもソフトも想定できない

  ヘテロな環境

  しかも、端末の作業がとても多い

  （たとえば、さい

  → よってたつのはプロトコル

日本では？

大学・研究機関

最初の位置づけ　ーー　研究用

法律との闘い（？）があったらしい

プロバイダ

最初はメールとBBS

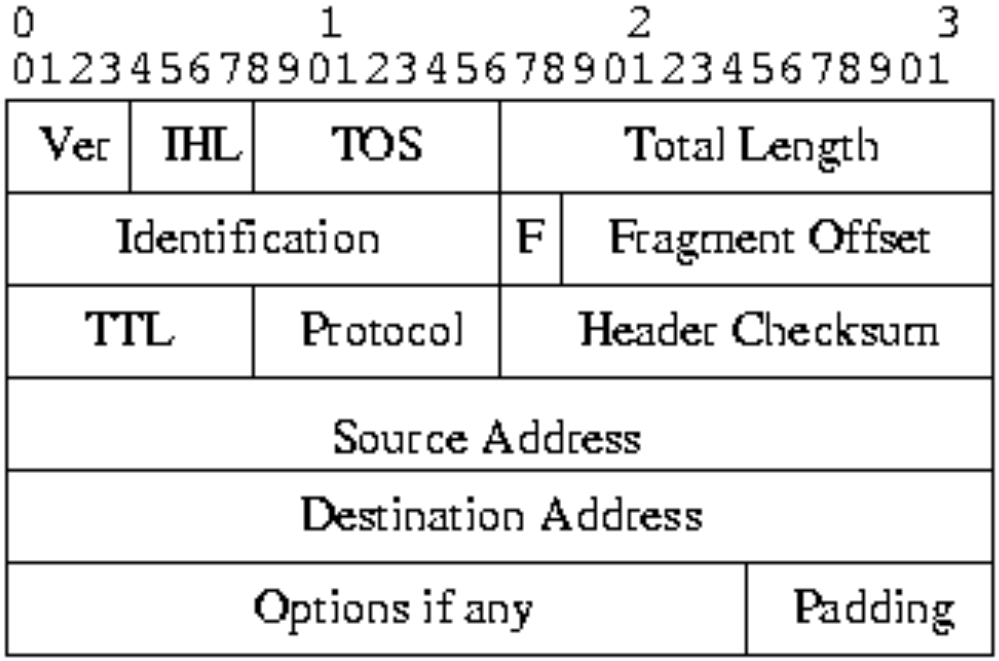
メールの相互乗り入れ　←　ま、いいじゃん。

あっというま（すごく驚いた）

IP の経路制御の基本

（あくまで基本）

- 各ルータが経路表を持っている
- 各パケットにあて先が書いてある



IP　の経路制御の基本

（あくまで基本）

- 各ルータが経路表を持っている
- 各パケットにあて先が書いてある

難しいのは、経路表の管理
ばんばん変わる。

経路制御

うんと昔

手動・静的

ちょっと昔

すべてのルータがお話・動的

今

階層化

経路表を小さく保つための努力

- CIDR

  Classless Internet　Domain Routing

背景　アドレス枯渇

1985　subnet 導入

1993　CIDR

N - O
PAC - PNI

933
F

019005

ゲイルスバーグの
春を愛す

番号	返す日	番号	返す日
6569	48. 3.-6		
	48. 5. 31		

933
F

019005

忘れずに期限までに返しましょう.
読み終わらないときは電話かハガキで登録番号と返す日をお知らせくだされば もう7日間借りられます.

上尾市図書館

上尾市本町三丁目1－1
電話 上尾(71)6422

図書返却期限

年 月 日	年 月 日	年 月 日
48. 3.-6		
48. 5. 31		
48. 7. 12		
48. 9. 27		
52. 6.-5		

上尾市図書館

<記入例>

<u>貸出カード</u>

Zwiebach, B.

A first course in strings theory

受入番号 04SCM00287

氏　名	所属	貸出月日	返却月日
△△△△	M1	2005.4.1	

M2
D1
D2
D3
研究生等

今週の余計なお世話

本を読んでみるのも
素敵です

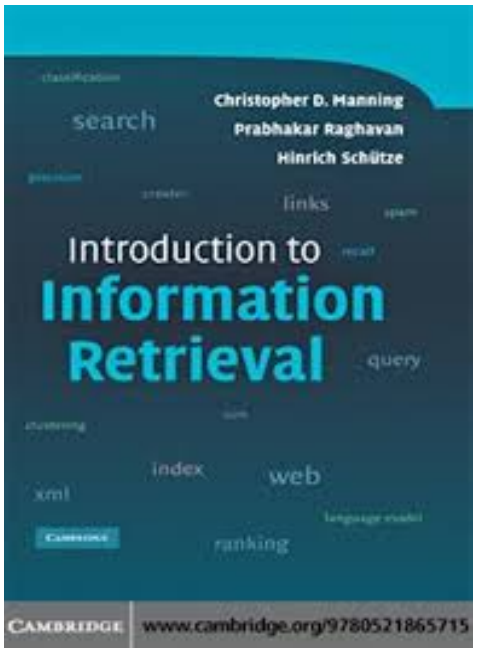
以下、Web 検索のタネ本

http://nlp.stanford.edu/IR-book/

情報検索の基礎

Introduction to Information Retrieval

C.D.Manning, P.Raghavan, H.Schutze



web の特徴

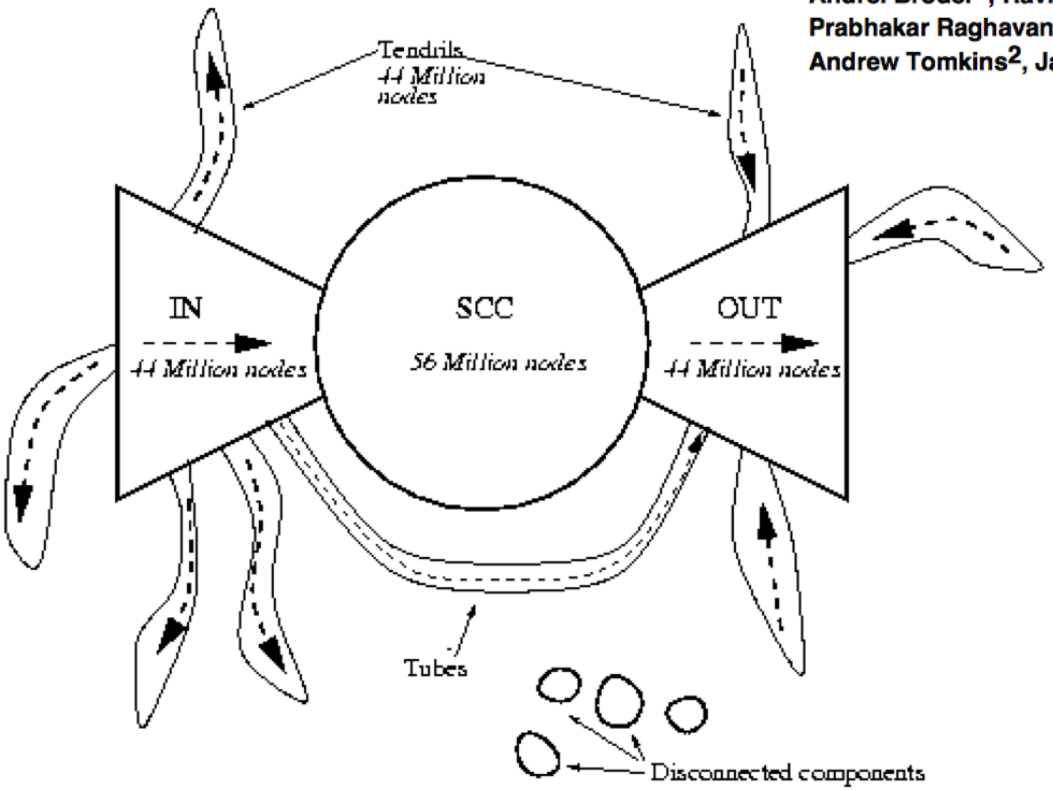
server と client が　protocol で会話
browser は、universal resource locator を指定
（階層的パス　→　コンテンツが返される）
◎browser は、わからない記述は無視

検索
- フルテキスト索引検索エンジン
- カテゴリー分類

→　　authoritativeness を測る必要

Graph Structure in the web[2003]
bowtie 構造 (有向グラフ)

Andrei Broder[1], Ravi Kumar[2], Farzin Maghoul[1],
Prabhakar Raghavan[2], Sridhar Rajagopalan[2], Raymie Stata[3],
Andrew Tomkins[2], Janet Wiener[3]



Region	SCC	IN	OUT	TENDRILS	DISC.	Total
Size	56,463,993	43,343,168	43,166,185	43,797,944	16,777,756	203,549,046

http://www.cis.upenn.edu/~mkearns/teaching/NetworkedLife/broder.pdf

Spammer vs Search Engine

paid inclusion (お金で順位をあげる)

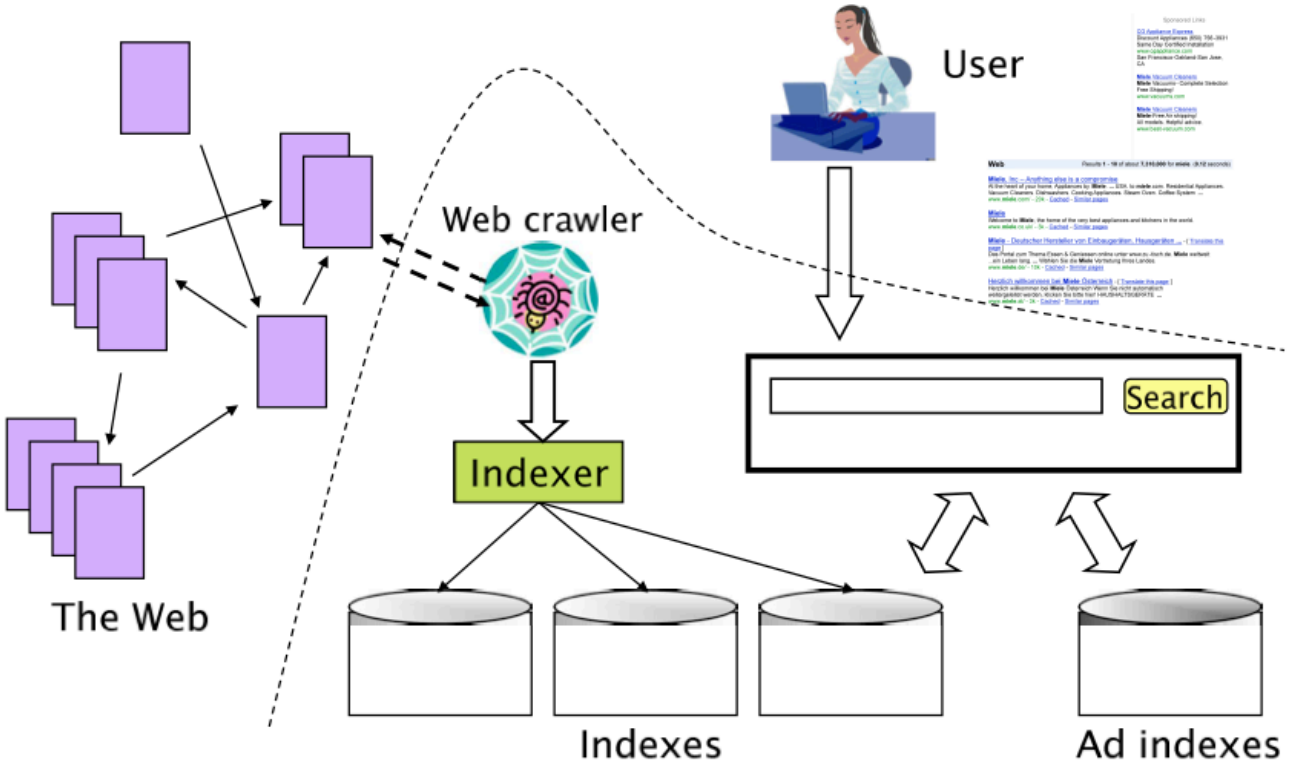
text の使い方に関しては、
SEO (Search Engine Optimization) と
Adversarial Information Retrieval

リンクの解析

query の分類

- 情報型
  一般的な情報 （白血病・プロバンス・群馬）
  いくつかのページから情報を得る
- 探索型
  あるサイトを探したい（ルフトハンザ、JAL）
  1番上に、欲しいサイトがでてほしい。
- 取引型
  購入、予約、ダウンロードなど。
  取引のサービスのインターフェースがほしい

検索エンジンのコンポーネント



► **Figure 19.7** The various components of a web search engine.

リンク解析

text で検索をすると困ること:
例えば、（当時）yahoo.com ページは “portal” という「単語」を含まなかった。

→　ページB をさす「anchor text」は、ページB の説明になってる （使える）
Journal of the ACM.

→　ページA から貼られた、ページ　Bへのリンクは、
A の作者のBに関する「推薦」とみなせる

リンク解析を使った代表的なアルゴリズム

- Page Rank
- HITS

Page Rank

ページの移動を 状態遷移と考える

定常状態で、(私は)どこにいそう？

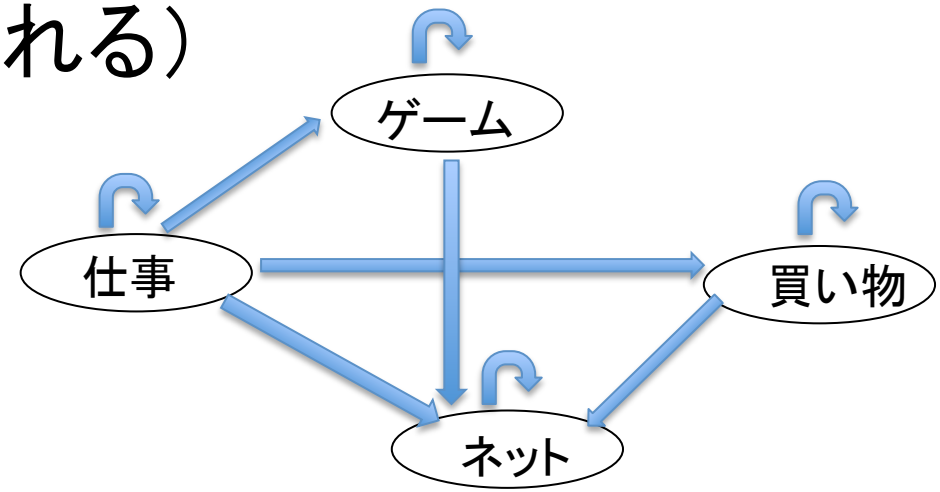
あるページ(state)にいるとき、移動は2種類

1. リンクに従う
   - (例えば)等確率で、貼られているリンク先に移動

2. URL 直打ち （テレポート）
   - 確率を別途定義
   - (例えば)全ノードに等確率で移動

（復習）マルコフ連鎖

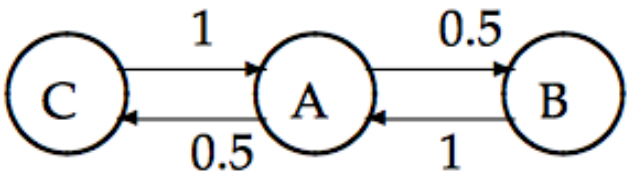
- 既約 （全部、どの状態にも、たどりつける）
- 有限再帰 （もどってこれる）
- 非周期的



の場合は、**定常分布 をただ一つ持つ**

（どういう分布から始めても、その状態に落ち着く）

単純なマルコフチェーンの例



遷移確率行列

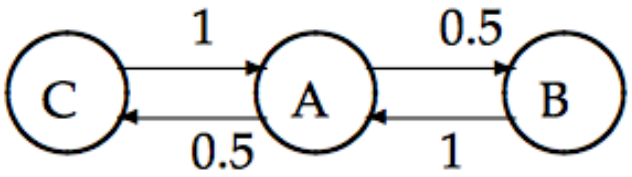
$$\begin{pmatrix} 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

各行を、足すと1

確率αでテレポートするとする
0<=α<=1（0.1くらいが多い）

遷移確率行列

テレポート:　等確率で、どこかのノードに行く
URL 直打ちを想定　(下の場合は各1/3)

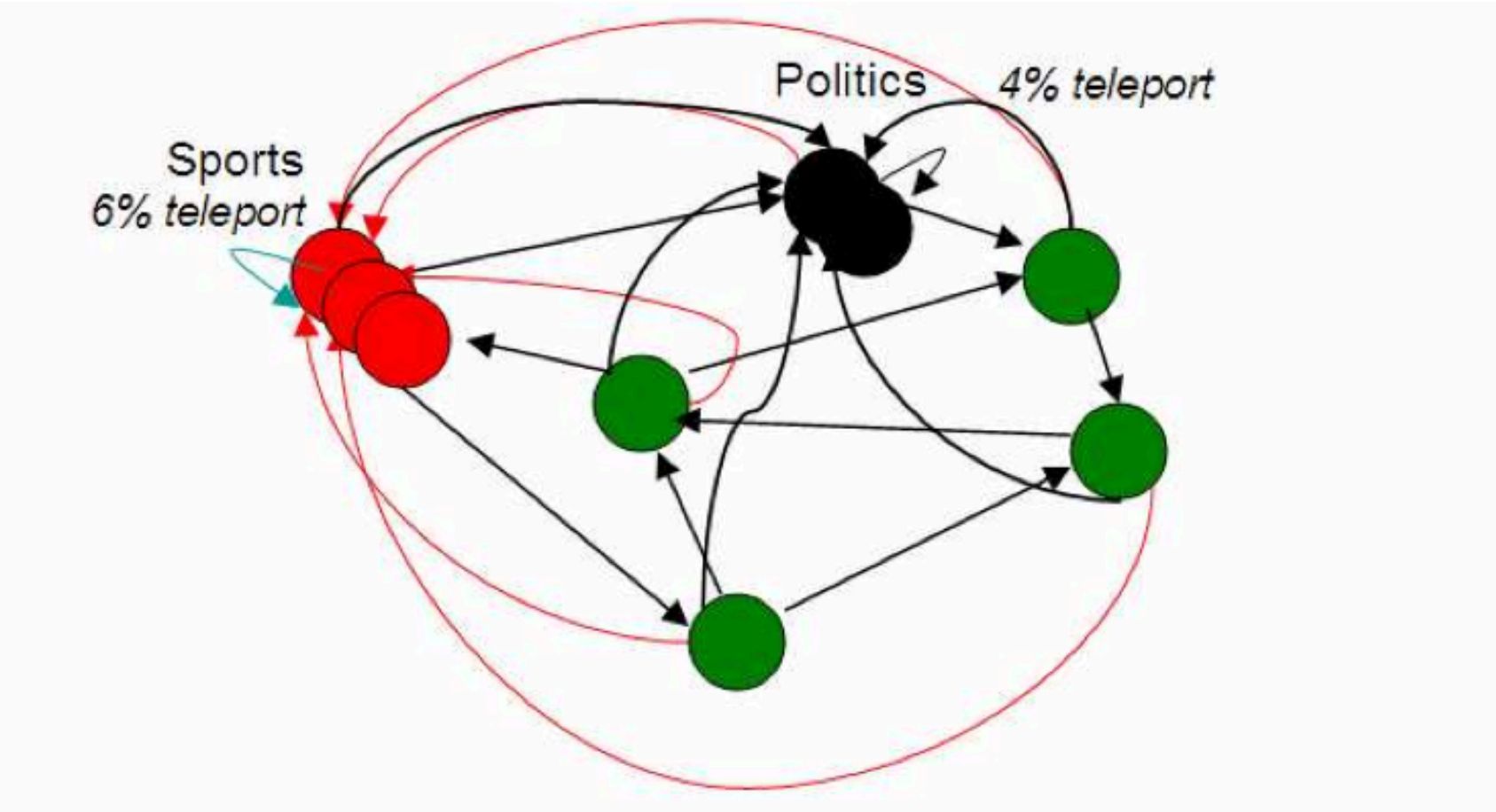


遷移確率行列

$$\begin{pmatrix} \alpha/n & 0.5(1-\alpha)+\alpha/n & 0.5(1-\alpha)+\alpha/n \\ (1-\alpha)+\alpha/n & \alpha/n & \alpha/n \\ (1-\alpha)+\alpha/n & \alpha/n & \alpha/n \end{pmatrix} = \begin{pmatrix} \alpha/3 & 0.5-\alpha/6 & 0.5-\alpha/6 \\ 1-2\alpha/3 & \alpha/3 & \alpha/3 \\ 1-2\alpha/3 & \alpha/3 & \alpha/3 \end{pmatrix}$$

各行を、足すと1

Topic Specific な Page Rank



HITS
Hypertext Induced Topic Selection

色んな意見があるけど、どれがいいの？
　単語数による多数決はうまくいかない (Spammer, SEO)

→ 「みんな」が「良いと言う物」がいいかも。

→ 「みんな」が「リンクを貼ってる物」がいいかも

→ 「色々リンクを貼る人」が「リンク・・（略」がいいかも

→ 「色々『リンク・・（略』を貼る人」が「リンク・・（略」がいいかも

Hub と Authority

良いHubページ:

多くの 良いAuthority ページをさしている

良い Authority ページ:

多くの良い Hub ページから指されている

$$h(v) \leftarrow \sum_{v \mapsto y} a(y)$$

$$a(v) \leftarrow \sum_{y \mapsto v} h(y).$$

線形代数

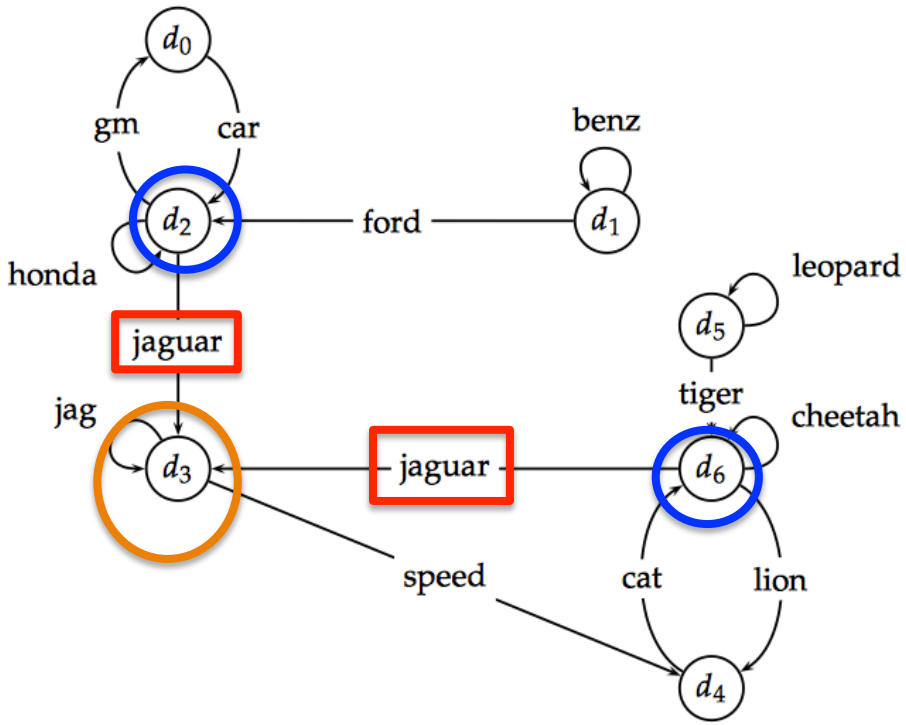
$$\vec{h} \leftarrow AA^T\vec{h}$$
$$\vec{a} \leftarrow A^TA\vec{a}.$$

$$\vec{h} = (1/\lambda_h)AA^T\vec{h}$$
$$\vec{a} = (1/\lambda_a)A^TA\vec{a}.$$

ただし、λ_hは　AA^T の固有値

例

jaguar という anchor text の重みを2倍にしたとき



0	0	1	0	0	0	0
0	1	1	0	0	0	0
1	0	1	2	0	0	0
0	0	0	1	1	0	0
0	0	0	0	0	0	1
0	0	0	0	0	1	1
0	0	0	2	1	0	1

$$\vec{h} = (0.03 \quad 0.04 \quad 0.33 \quad 0.18 \quad 0.04 \quad 0.04 \quad 0.35)$$

$$\vec{a} = (0.10 \quad 0.01 \quad 0.12 \quad 0.47 \quad 0.16 \quad 0.01 \quad 0.13)$$

レポート・成績について

- ほぼ毎回、プログラミング課題を出題する予定
  - 効率の良い計算機実験のためのツールを使ってみる
  - アルゴリズムの実装
  - ライブラリの利用・・・　など

- ３回以上、レポートのファイルとプログラムのソースを添付し、メールで提出のこと
  - E-mail: algorithm2014@edu.jar.jp
  - サブジェクト 「アルゴリズムとプログラム実践講座・レポート」
  - 学生証番号と名前は、 メールの本文にも書いてください。
  - **〆切:6/8 (日)深夜 （講評の都合上。〆切後も受付ます）**
  - プログラムやレポートは（見本として）公開することがあります。適宜、作者名やコピーライトをいれておいてください。公開不可の場合は、プログラムの冒頭にその旨、コメントをいれておいてください。
  - 質問・作問提案も歓迎 （作問については採用の場合は別途加点）
  - サンプルプログラムは「初心者向け」です。 **上級者は無視してください。**

推奨環境など

- Linux, Mac,（Windows+Cygwin）
- 仮想マシン環境（VMware, VirtualBox, Parallels)
  - 余裕があれば、いろいろな組み合わせを試して比較してみると面白いと思います

- 言語
  - 自由。ただし、一般的でない言語については、上記いずれかのOS 上にインストール可能なもの

第六回　課題

〆切:6/8 (日)深夜
（講評の都合上。〆切後も受付ます）

来週は課題の出題はお休みです。

Web グラフの解析

適当な Web グラフを拾ってきて
何か解析してみる。
e.x. 次数の分布・ PageRank・ HITS など

SNAP **Stanford Large Network Dataset Collection**
http://snap.stanford.edu/data/web-Google.html
他に、お薦めがあったら、教えてください。