Text Processing on the Web

Week 1

Orientation

Intro to Web Search

The material for these slides are borrowed heavily from the precursor of this course by Tat-Seng Chua as well as slides from the accompanying recommended texts Baldi et al. and Manning et al.

Web phenomenon

Pre 2000

- Exponential Growth
- Static HTML, primarily text
- Pull technology
- Placement of web interfaces to DBs
- High value E-Commerce systems

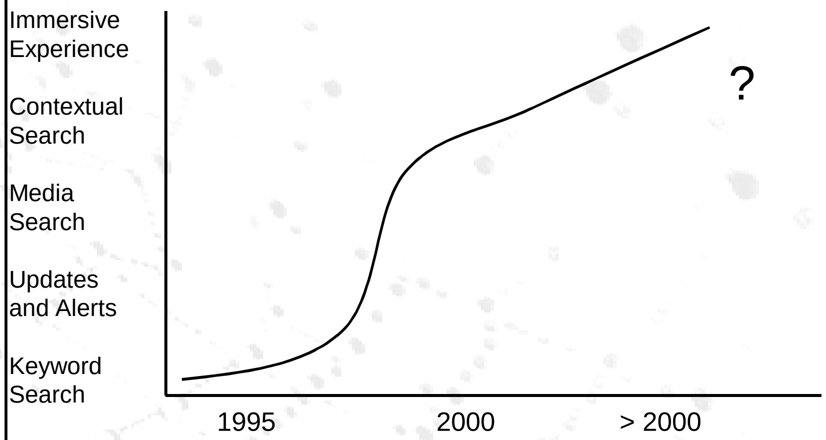
- Here, only talking about text

Web phenomenon

Post 2000

- The fat pipe
- Trend towards other media
  - Flickr, Youtube, Myspace
- Social Media
  - Del.icio.us, Digg,
  - Blogs, wikis, folksonomies
- Push technology
  - RSS, alerting
- Catering for mobile devices
- Web as application
  - Google Spreadsheets, PIM
- (The long tail)

What's next?



- What do you think?
  Speak out on the forum!!

Orientation

- Teaching Staff
- Course Overview
  - Course website review
- Continuous Assessment

Teaching staff

- Lecturer:

  Min-Yen Kan (“Min”)

  kanmy@comp.nus.edu.sg

  Office: AS6 05-12

  6516-1885

  Hours: before class

  Hobbies:
  rock climbing,
  ballroom dancing,
  and inline skating…

Lost in Hakodate, Japan

- Go over website now
- Remember to cover Academic Honesty

Continuous Assessment

- Possible Assignments (55%)
  - Both need to be demo'ed
  1. Passage Retrieval System
     - Working system to retrieve passages in scientific articles
  2. Summarization System
     - Query summarization of individual scientific articles
- Exam (40%)
  - Essay and algorithm development
  - Open book

- Late Policy
  - Intentionally set very harsh
- Academic Honesty
  - I trust you, so please reciprocate
  - Punishment will be harsh

Web Basics

Baldi *et al.* (Chapter 2)

Everyone knows the web…

- Assume you know basic HTML
- Assume you know its relation to SGML and XML
- Assume you know DTDs

Let's quickly go over some other aspects of the web

- HTTP specifics
- Log files

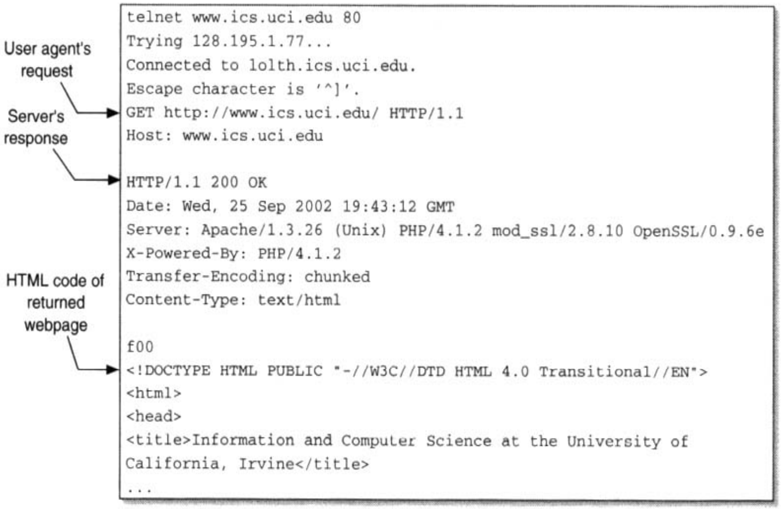
Components of the Web

The Internet and WWW are distinct

What is the web? Three components:

1. Resources:
   - Conceptual mappings to concrete or abstract entities, which do not change in the short term
   - ex: comp website (web pages and other kinds of files)
2. Resource identifiers (hyperlinks):
   - Strings of characters represent generalized addresses that may contain instructions for accessing the identified resource
   - http://www.comp.nus.edu.sg/ is used to identify the comp homepage
3. Transfer protocols:
   - Conventions that regulate the communication between a browser (web user agent) and a server

Methods in HTTP




```
telnet www.ics.uci.edu 80
Trying 128.195.1.77...
Connected to lolth.ics.uci.edu.
Escape character is '^]'.
GET http://www.ics.uci.edu/ HTTP/1.1
Host: www.ics.uci.edu

HTTP/1.1 200 OK
Date: Wed, 25 Sep 2002 19:43:12 GMT
Server: Apache/1.3.26 (Unix) PHP/4.1.2 mod_ssl/2.8.10 OpenSSL/0.9.6e
X-Powered-By: PHP/4.1.2
Transfer-Encoding: chunked
Content-Type: text/html

f00
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
<html>
<head>
<title>Information and Computer Science at the University of
California, Irvine</title>
...
```

Figure 2.4 Example of the use of the GET method in an HTTP 1.1 session.

GET	Retrieve an entity identified by a request URI (fetch a web page or file)
HEAD	Identical to GET but just return header
POST	Append enclosed entity. The supplied URI will handle the entity (e.g., used to post a message to a newsgroup)
PUT	Store an enclosed entity under the supplied URI (e.g., store a Web page or file with the server)

Server Log Files

- Server Transfer Log: transactions between a browser and server are logged
  - IP address, the time of the request
  - Method of the request (GET, HEAD, POST…)
  - Status code, a response from the server
  - Size in byte of the transaction
- Referrer Log: where the request originated
- Agent Log: browser software making the request (spider)
- Error Log: request resulted in errors (404)

- Success 2xx
  - 200 OK
  - 201 Created
  - 202 Accepted
  - 203 Partial Info

  …
- Redirection 3xx
  - 301 Moved

  …
- Error 4xx
  - 400 Bad request, syntax
  - 401 Unauthorized
  - 402 Payment required
  - 403 Forbidden
  - 404 Not Found
- Internal Error 5xx
  - 500 Internal Error
  - 501 Not Implemented
  - 502 Service temporarily overloaded
  - 503 Gateway timeout

Why do you see more 403s than 401s?

Server Log Analysis

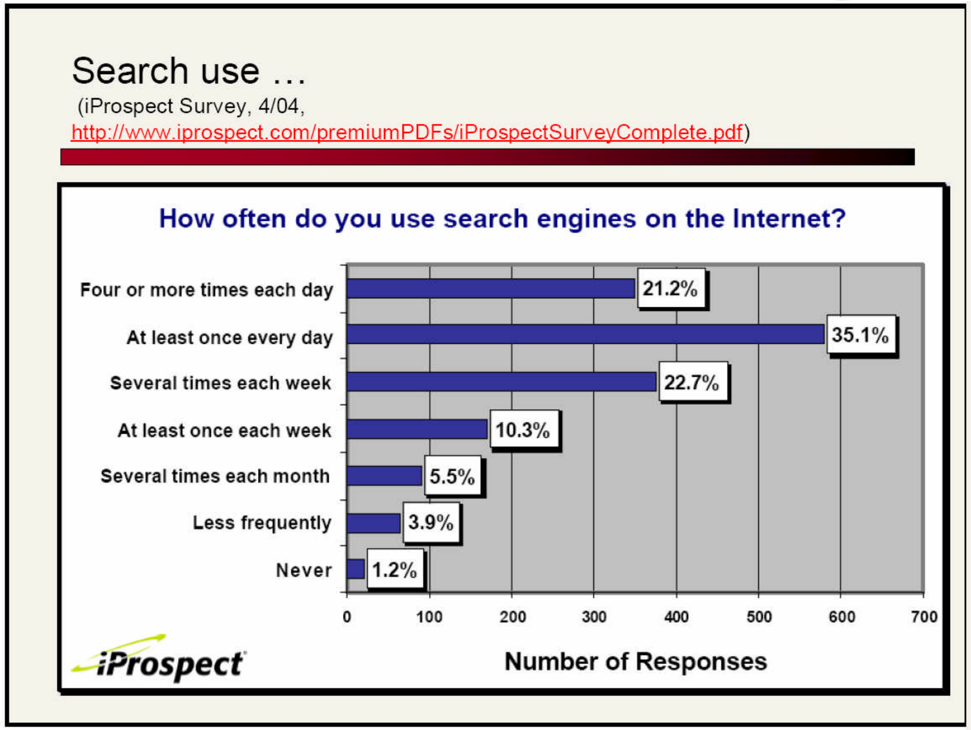
- Most and least visited web pages
- Entry and exit pages
- Referrals from other sites or search engines
- What are the searched keywords
- How many clicks/page views a page received
- Error reports, like broken links

Search Engines

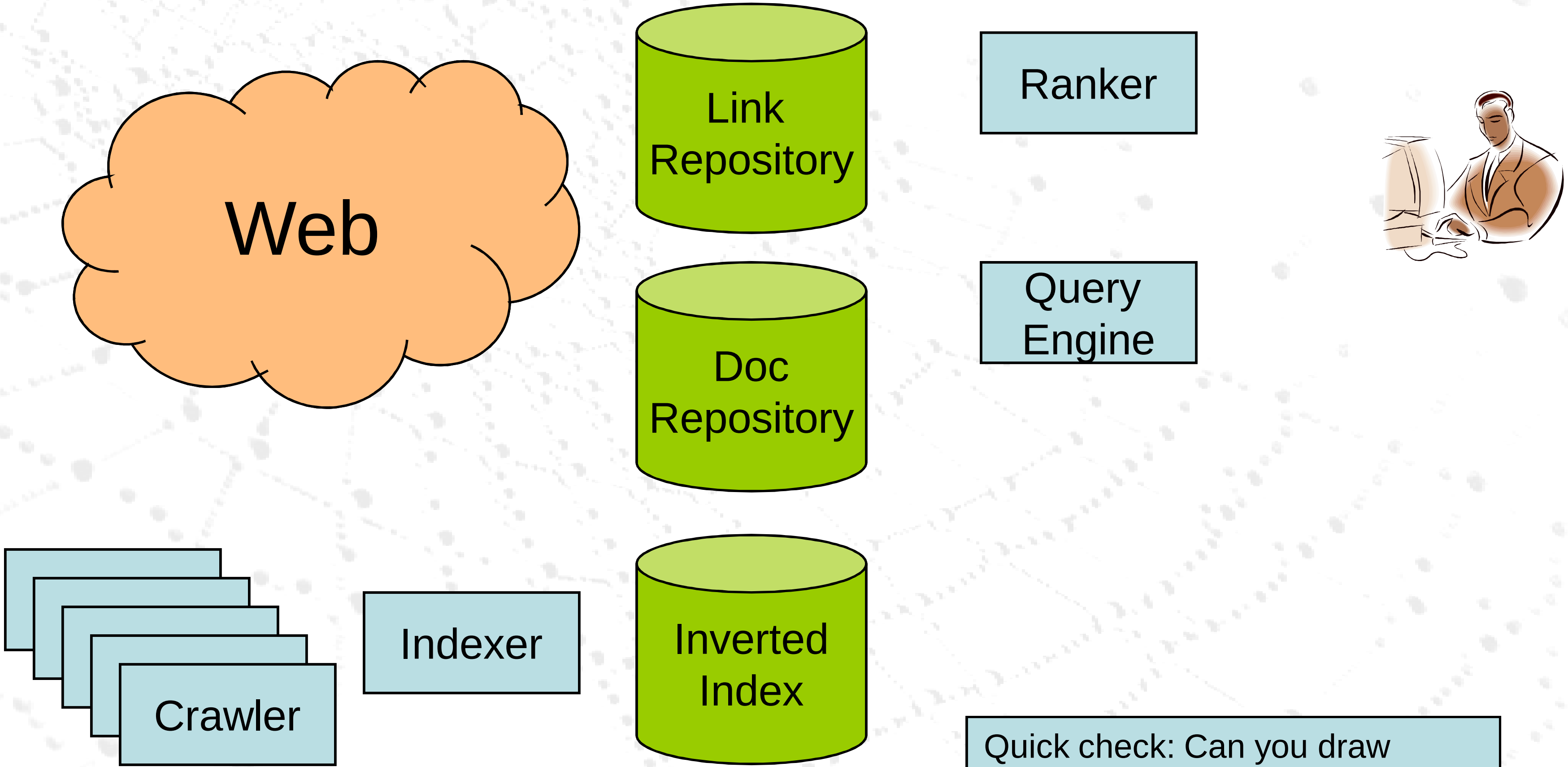
- According to Pew Internet Project Report (2002), search engines are the most popular way to locate information online
- About 33 million U.S. Internet users query on search engines on a typical day.
- More than 80% have used search engines
- Search Engines are measured by coverage and recency

Search engines are critical to the web

- No incentive in creating content unless it can be found
- SE make aggregation of niche interests possible
- Topological argument, to be discussed today



The anatomy of a search engine



Quick check: Can you draw
the links and label them yourself?

Web Crawler

- A crawler is a program that picks up a page and follows all the links on that page
- Crawler = Spider
- Types of crawler:
  - Breadth First
  - Depth First
- Focused Crawlers
  - Look for specific type of documents

Simple-Crawler (S_0, D, E)

Q = S_0

while Q not **empty**

  u = Dequeue(Q)

  D(u) = Fetch(u)

  Store(D, (d(u),u))

  L = Parse(d(u))

  for each v in L

    Store (E, (u,v))

    if not (v **in** D or v **in** Q)

      Enqueue(Q,v)

Problems

- Document downloading is problematic
- Crawlers should respect robots.txt
- Crawling as (D)DoS attacks
- Spider traps: alias hostnames, or server redirection with dynamically generated pages
  - Up to 40% are duplicates
- Dynamic nature of the web: static versus dynamic sites

“Like taking a picture of a living scene with some objects at rest and others moving”

The Web Graph

Baldi *et al.* (Chapter 3)

Outline

- Computing the size of the Web
- Adversarial IR / SEO
- Actual topology of the Web
- Models of Web structure evolution

Coverage

Overlap analysis used for estimating the size of the indexable web

- Caveat: Indexed = in doc database, first *n* words

- W: set of webpages
- Wa, Wb: pages crawled by two independent engines a and b
- P(Wa), P(Wb): probabilities that a page was crawled by a or b
- $P(Wa)=|Wa| / |W|$
- $P(Wb)=|Wb| / |W|$

Overlap Analysis

$P(Wa \cap Wb | Wb) = P(Wa \cap Wb) / P(Wb)$

$= |Wa \cap Wb| / |Wb|$

If a and b are independent:

$P(Wa \cap Wb) = P(Wa)*P(Wb)$

$P(Wa \cap Wb | Wb) = P(Wa)*P(Wb)/P(Wb)$

$= |Wa| * |Wb| / |Wb|$

$= |Wa| / |W|$

$=P(Wa)$

Overlap Analysis

Using |*W*| = |*Wa*|/ *P(Wa),* researchers found:

- Web had at least 320 million pages in 1997
- 60% of web was covered by six major engines
- Maximum coverage of a single engine was 1/3 of the web

Problems

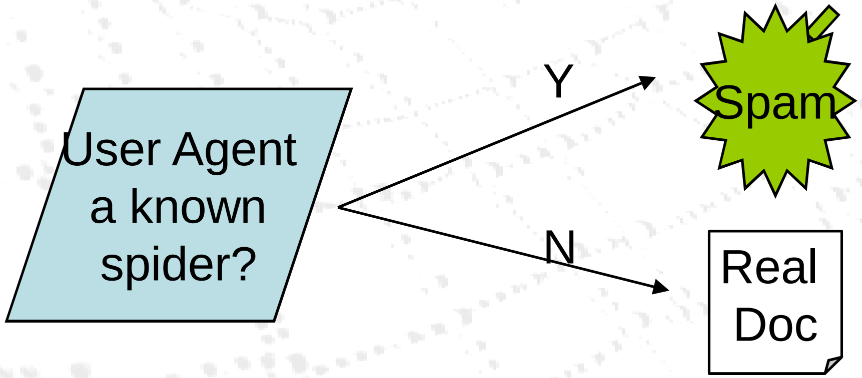
- Doesn't explicitly account for popularity of pages
- Which queries to use for estimation? Best to be random, but this is hard.
  - Use random local query logs
  - Other alternatives?

Adversarial IR

Rise of Spam

- E-commerce on the web
- Search Engine Optimization
- Cost per Impression (CPI/CPM), Cost per Click (CPC)

- Cloaking
  - Different info depending on browser-agent

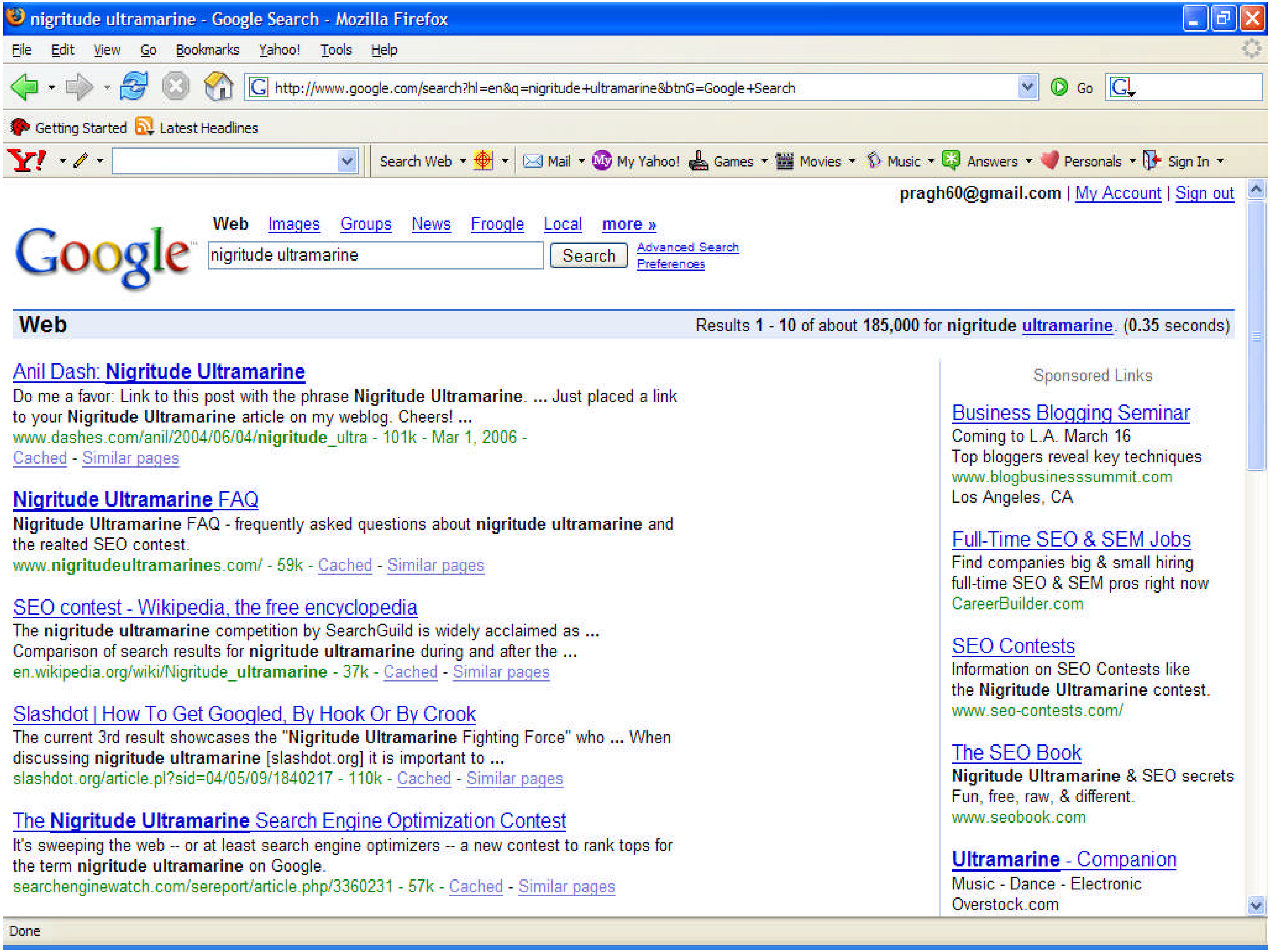


More Adversarial IR

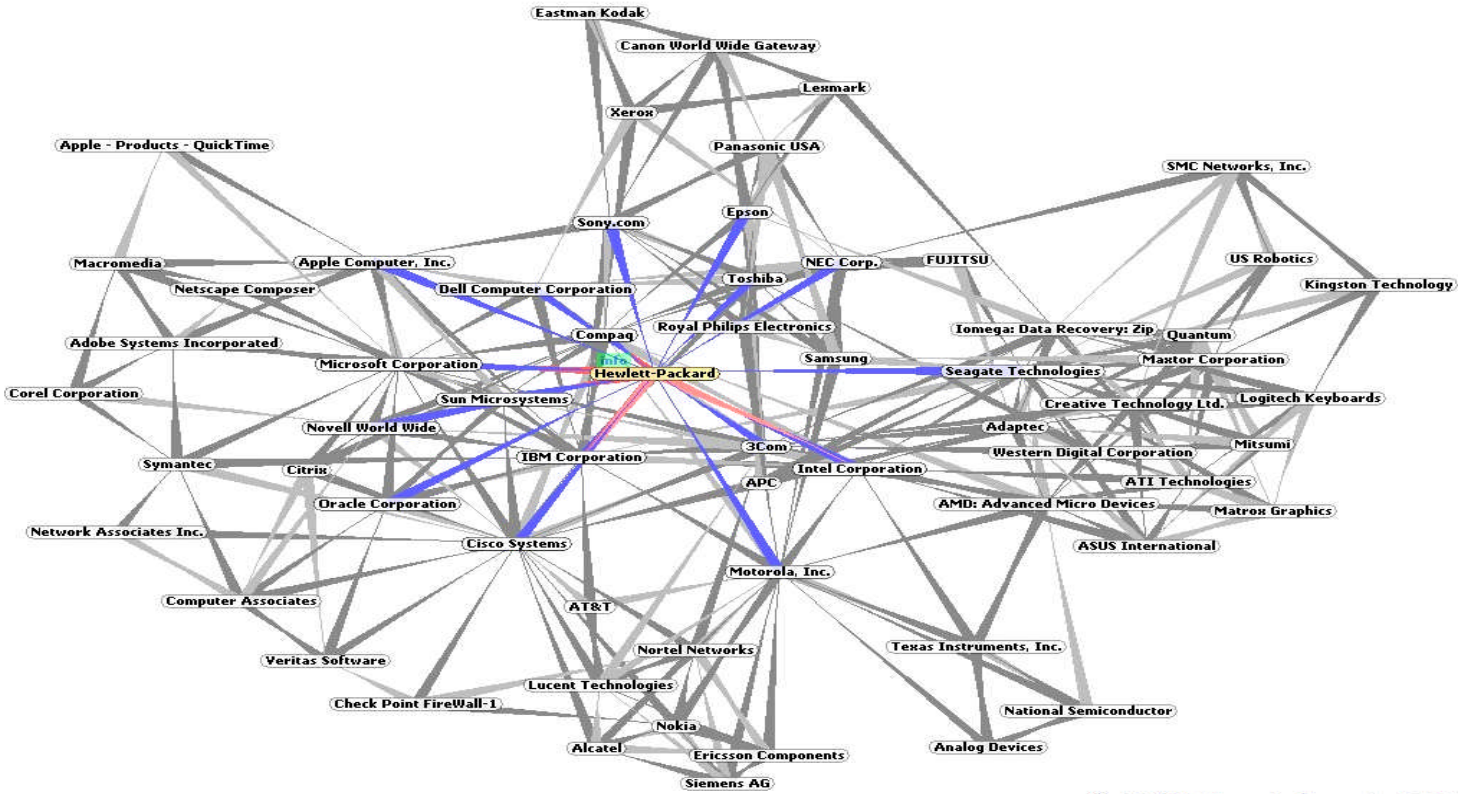
Other methods

- Doorway / bridge pages
  - Pages optimized for a single keyword that re-directs to the real target page
  - No actual content. E.g., “Click here for widgets”
- Click spam – targeting user logs
  - Putting a competitor out of business

- How to combat?
  - Link analysis (few weeks down the road)
  - An arms race; Web 2.0 exacerbated
  - Try it yourself. Contests abound


nigritude ultramarine - Google Search - Mozilla Firefox
File Edit View Go Bookmarks Yahoo! Tools Help
http://www.google.com/search?hl=en&q=nigritude+ultramarine&btnG=Google+Search
Go
Getting Started Latest Headlines
Search Web Mail My Yahoo! Games Movies Music Answers Personals Sign In
pragh60@gmail.com | My Account | Sign out
Google
Web Images Groups News Froogle Local more »
nigritude ultramarine
Search
Advanced Search
Preferences
Web
Results 1 - 10 of about 185,000 for nigritude ultramarine. (0.35 seconds)
Anil Dash: Nigritude Ultramarine
Do me a favor: Link to this post with the phrase Nigritude Ultramarine. ... Just placed a link to your Nigritude Ultramarine article on my weblog. Cheers! ...
www.dashes.com/anil/2004/06/04/nigritude_ultra - 101k - Mar 1, 2006 - Cached - Similar pages
Nigritude Ultramarine FAQ
Nigritude Ultramarine FAQ - frequently asked questions about nigritude ultramarine and the realted SEO contest.
www.nigritudeultramarines.com/ - 59k - Cached - Similar pages
SEO contest - Wikipedia, the free encyclopedia
The nigritude ultramarine competition by SearchGuild is widely acclaimed as ... Comparison of search results for nigritude ultramarine during and after the ...
en.wikipedia.org/wiki/Nigritude_ultramarine - 37k - Cached - Similar pages
Slashdot | How To Get Googled, By Hook Or By Crook
The current 3rd result showcases the "Nigritude Ultramarine Fighting Force" who ... When discussing nigritude ultramarine [slashdot.org] it is important to ...
slashdot.org/article.pl?sid=04/05/09/1840217 - 110k - Cached - Similar pages
The Nigritude Ultramarine Search Engine Optimization Contest
It's sweeping the web -- or at least search engine optimizers -- a new contest to rank tops for the term nigritude ultramarine on Google.
searchenginewatch.com/sereport/article.php/3360231 - 57k - Cached - Similar pages
Sponsored Links
Business Blogging Seminar
Coming to L.A. March 16
Top bloggers reveal key techniques
www.blogbusinesssummit.com
Los Angeles, CA
Full-Time SEO & SEM Jobs
Find companies big & small hiring full-time SEO & SEM pros right now
CareerBuilder.com
SEO Contests
Information on SEO Contests like the Nigritude Ultramarine contest.
www.seo-contests.com/
The SEO Book
Nigritude Ultramarine & SEO secrets
Fun, free, raw, & different.
www.seobook.com
Ultramarine - Companion
Music - Dance - Electronic
Overstock.com
Done

Web Graph



http://www.touchgraph.com/TGGoogleBrowser.html

Properties of Web Graphs

- Connectivity follows a power law distribution
- The graph is sparse
  - $|E| = O(n)$ or at least $o(n^2)$
  - Average number of hyperlinks per page roughly a constant
- A small world graph

Power Law Connectivity

- Distribution of number of connections per node follows a power law distribution
- Study at Notre Dame University reported
  - γ = 2.45 for outdegree distribution
  - γ = 2.1 for indegree distribution

Note: in contrast, random graphs have a Poisson distribution if p is large.

- – Decays exponentially fast to 0 as *k* increases towards its maximum value *n-1*

Examples of networks with Power Law Distribution

- Internet at the router and interdomain level
- Citation network
- Collaboration network of actors
- Networks associated with metabolic pathways
- Networks formed by interacting genes and proteins
- Network of nervous system connection in C. elegans

Small World Networks

- It is a ‘small world’
  - Millions of people. Yet, separated by “six degrees” of acquaintance relationships
  - Popularized by Milgram’s famous experiment
- Mathematically
  - Diameter of graph is small (log *N*) as compared to overall size
    - 3. Property seems interesting given ‘sparse’ nature of graph but …
    - This property is ‘natural’ in ‘pure’ random graphs

The small world of WWW

- Empirical study of Web-graph reveals small-world property
  - Average distance (d) in simulated web:

    $d = 0.35 + 2.06 \log(n)$

    e.g. $n = 10^9$, $d \sim= 19$
  - Graph generated using power-law model
  - Diameter properties inferred from sampling
    - Calculation of max. diameter computationally demanding for large values of *n*

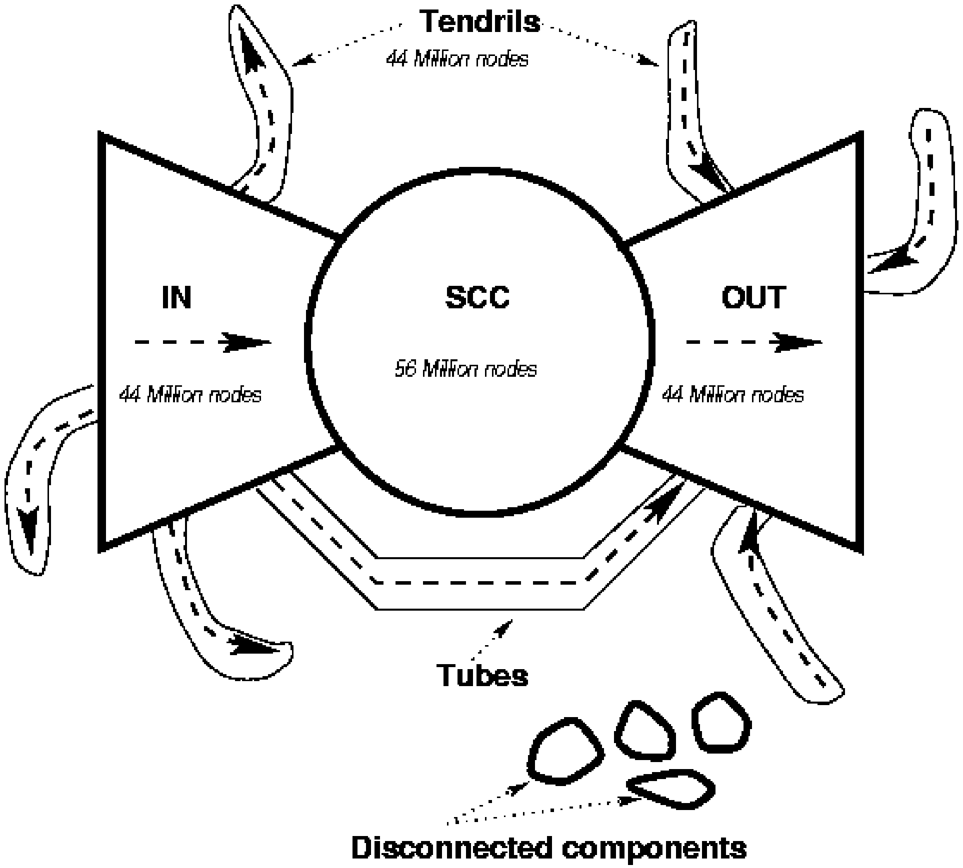
Implications for Web

- Logarithmic scaling of diameter makes future growth of web manageable
  - 10-fold increase of web pages results in only 2 more additional ‘clicks’, but …
  - Users may not take shortest path, may use bookmarks or just get distracted on the way
  - Therefore search engines play a crucial role

Pagerank

- In-degree as first approximation
- Random surfer model
- Teleportation to get to and out of traps / isolated structures
- Return to this later in Week 5

Web Topology (ca. 2000)



Max Diameter (in SCC, 16; IN-to-OUT, up to 500)

Probability of connection of random 2 pages: 24%, if connected, average path length of 16

Models for the Web Graph

- Stochastic models that can explain or atleast partially reproduce properties of the web graph
  - The model should follow the power law distribution properties
  - Represent the connectivity of the web
  - Maintain the small world property

Web Page Growth

- Empirical studies observe a power law distribution of site sizes
  - Size includes size of the Web, number of IP addresses, number of servers, average size of a page etc
- A generative model is being proposed to account for this distribution

Components of the model

- Proportional size changes (β)
  - " sites have short-term size fluctuations up or down that are proportional to the size of the site "
  - A site with 100,000 pages may gain or lose a few hundred pages in a day whereas the effect is rare for a site with only 100 pages

- There is an overall growth rate α
  - so that the size S(*t*) satisfies $S(t+1) = \alpha(1+\eta_t\beta)S(t)$ where
    - η_t is the realization of a +/-1 Bernoulli random variable at time t with probability 0.5
    - β is the absolute rate of the daily fluctuations

After T steps

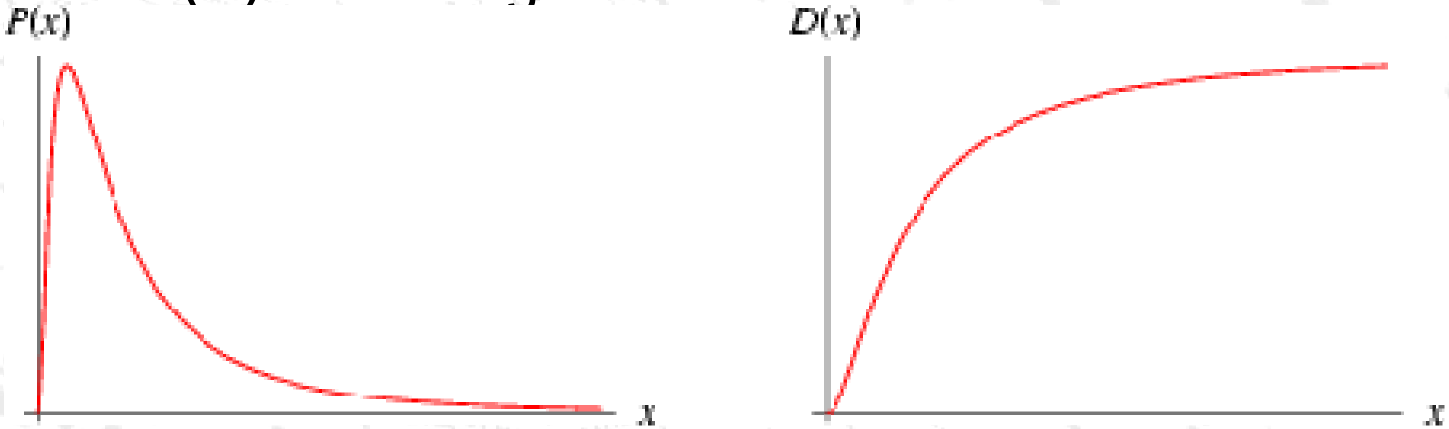
$$S(T) = \alpha^T S(0) \prod_{t=0}^{T-1} (1 + \eta_t \beta)$$

so that

$$\log S(T) = T \log \alpha + \log S(0) + \sum_{t=0}^{T-1} \log(1 + \eta_t \beta)$$

Theoretical Considerations

- Log S(T) can also be associated with a binomial distribution counting the number of times ht = +1
- Hence S(T) has a log-normal distribution



- The probability density and cumulative distribution functions for the log normal distribution

Modified Model

- Can be modified to obey power law distribution
- Model is modified to include the following inorder to obey power law distribution
  - A wide distribution of growth rates across different sites and/or
  - The fact that sites have different ages

Capturing Power Law Property

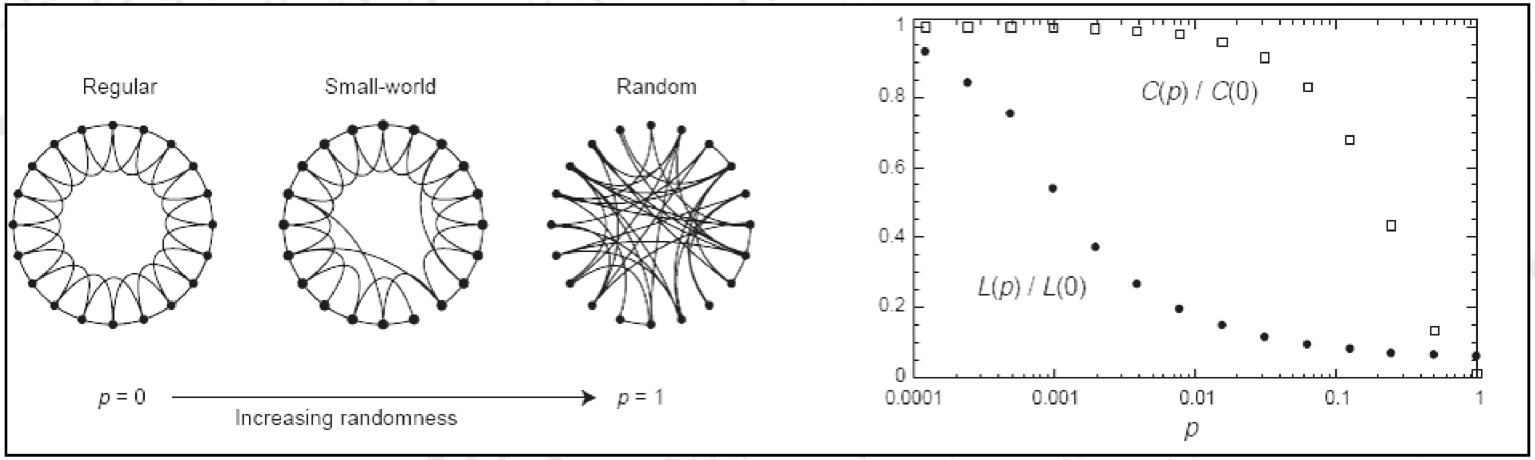
- Inorder to capture Power Law property it is sufficient to consider that
  - Web sites are being continuously created
  - Web sites grow at a constant rate α during a growth period after which their size remains approximately constant
  - The periods of growth follow an exponential distribution
- This will give a relation $\lambda = 0.8\alpha$ between the rate of exponential distribution λ and α the growth rage when power law exponent $\gamma = 1.08$

Lattice Perturbation (LP)

- Step 1:
  - Take a regular network (e.g. lattice)
- Step 2:
  - Shake it up (perturbation)
- Step 2 in detail:
  - For each vertex, pick a local edge
  - ‘Rewire’ the edge into a long-range edge with a probability (p)
  - p=0: organized, p=1: disorganized

Lattice Perturbation (LP)

- Start with a regular network, and perturb
- End up with a Semi-Organized (SO) Network



- L(p) = Average Path Length
- C(p) = Clustering coefficient (ratio of possible local edges); local property

Effect of ‘Shaking it up’

- Small shake (*p* close to zero)
  - High cliquishness AND short path lengths
- Larger shake (*p* increased further from 0)
  - *d* drops rapidly (increased small world phenomena)
  - *c* remains constant (transition to small world almost undetectable at local level)
- Effect of long-range link:
  - Addition: non-linear decrease of *d*
  - Removal: small linear decrease of *c*

Terms (Cont'd)

- Organized Networks
  - Are 'cliquish' (Subgraph that is fully connected) in local neighborhood
  - Probability of edges across neighborhoods is almost non existent (*p=0* for fully organized)
- "Disorganized" Networks
  - 'Long-range' edges exist
  - Completely Disorganized <=> Fully Random (Erdos Model) : *p=1*

Scalable Random Networks

- Evolutionary Model (grows as the first model does over timesteps)
  - Start with M_0 vertices at T = 0
  - At each step, add a new node v and m <= M_0 edges where edges connect new node to old vertices
- Vertices attach preferentially (instead of randomly); “rich get richer”

$$P(v,w) = \frac{k_w}{\sum_r k_r}$$

Summary

- Web basics
- Adversarial IR
- Web graph topology

  … and models for simulating them

Next Week

- How do search engines work?