An Introduction to Web Mining

Ricardo Baeza-Yates, Aristides Gionis

Yahoo! Labs

Barcelona, Spain

Contents of the tutorial

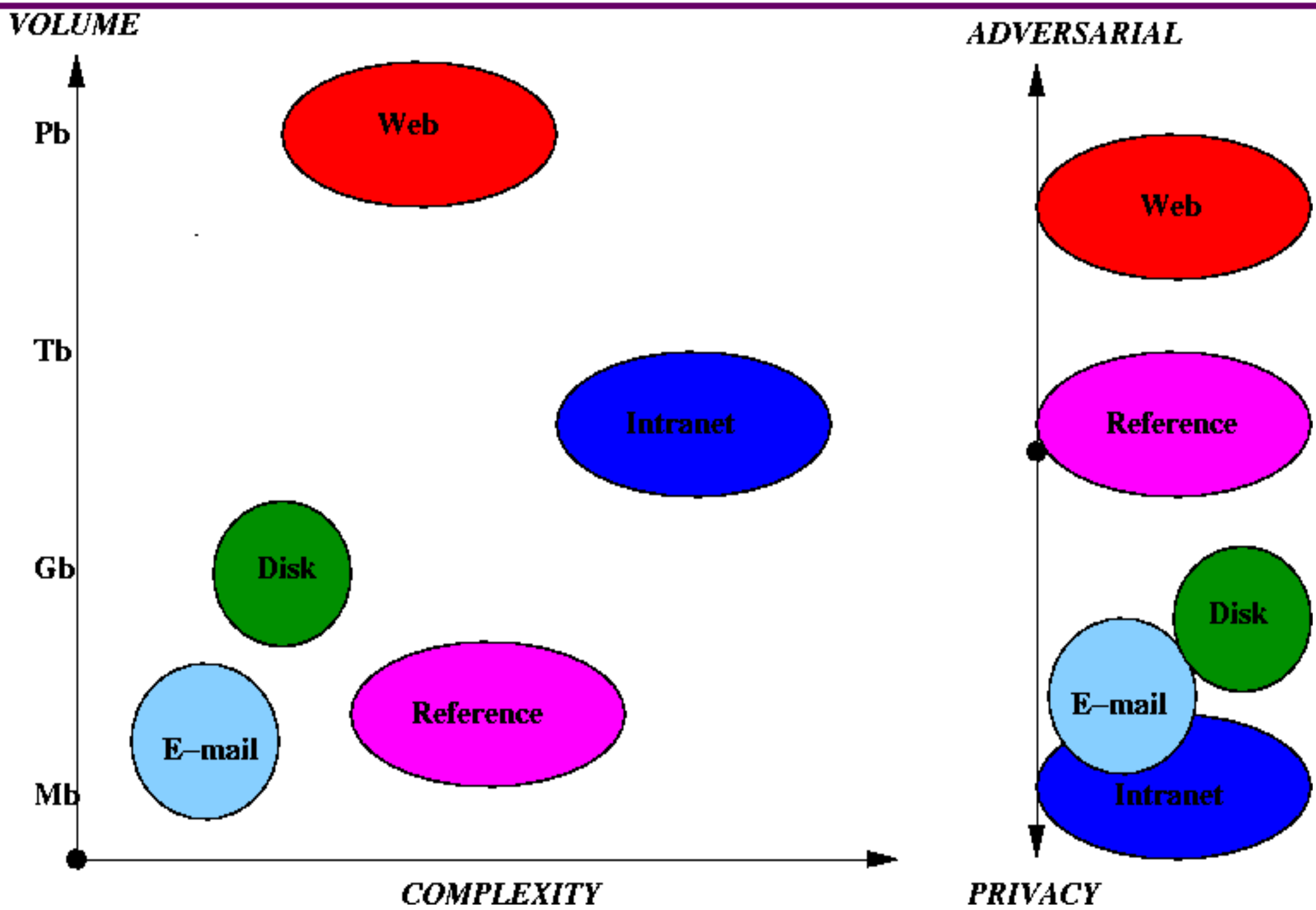
1. Motivations for Web mining
   - The Web, definitions, wisdom of crowds, the long tail, search, Web spam, advertising and social media
2. The mining process
   – Crawling, data cleaning and data anonymization
3. The basic concepts
   – Data statistics, usage mining, link mining, graph mining, finding communities
4. Detailed examples
   – Size of the web, near-duplicate detection, spam detection based on content and links, social media mining, query mining
5. Final remarks

Motivation

Internet and the Web Today

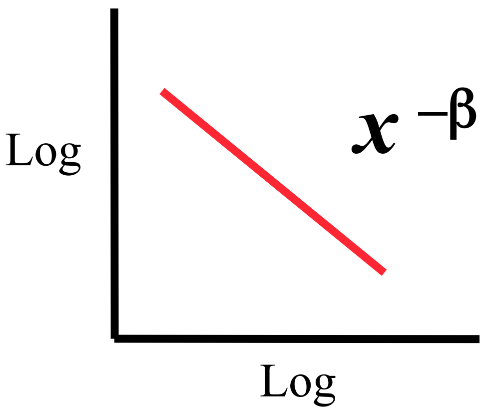
- **Between 1 and 2.5 billion people connected**
  - 5 billion estimated for 2015
- **1.8 billion mobile phones today**
  - 500 million expected to have mobile broadband during 2010
- **Internet traffic has increased 20 times in the last 5 years**
- **Today there are more than 200 million Web servers**
- **The Web is in practice unbounded**
  - Dynamic pages are unbounded
  - Static pages over 20 billion?

Different Views on Data



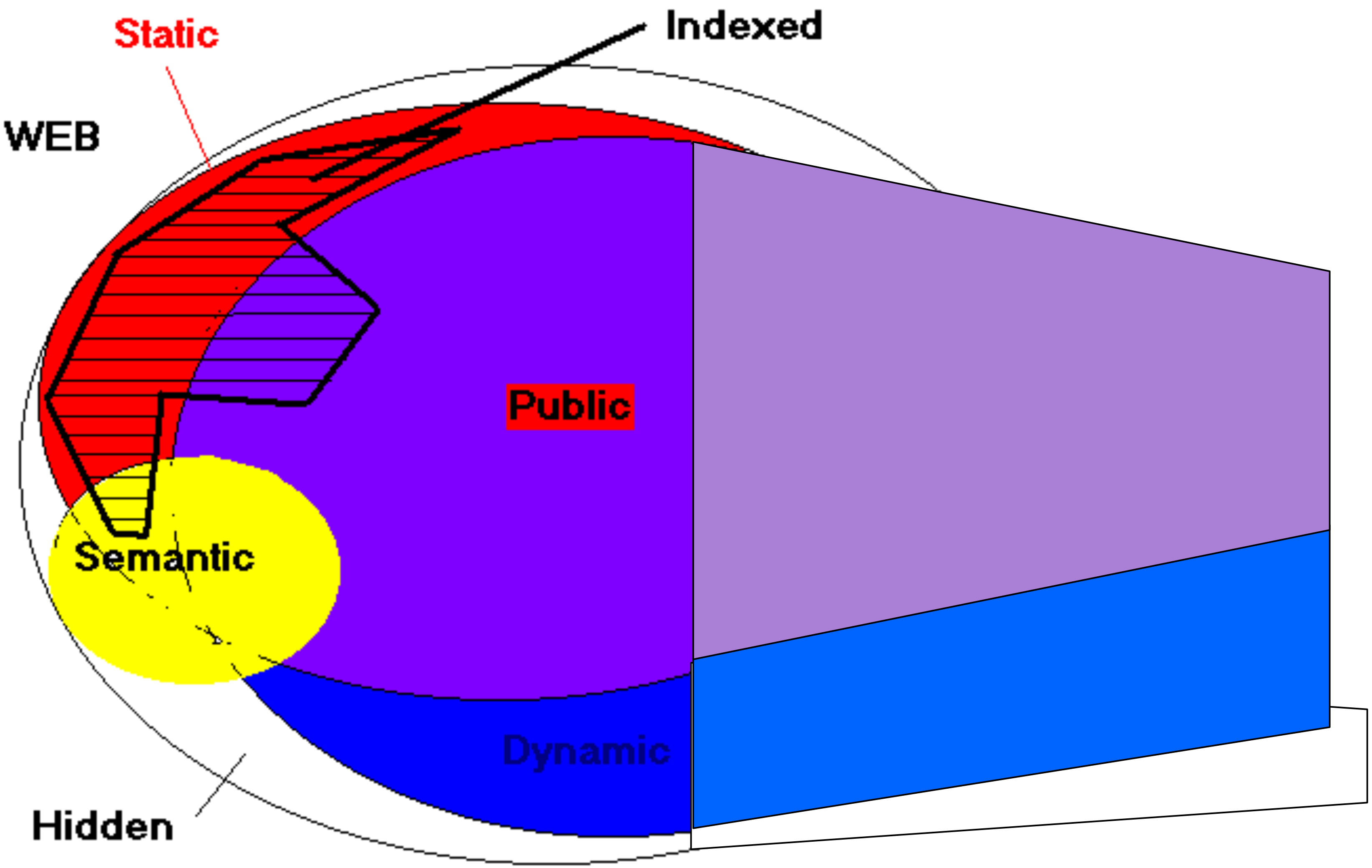
The Web

- **Largest public repository of data**
- **Today, there are more than 213 million Web servers (Aug 2010) and more than 750 million hosts (Apr 2010)**
- **Well connected graph with out-link and in-link power law distributions**

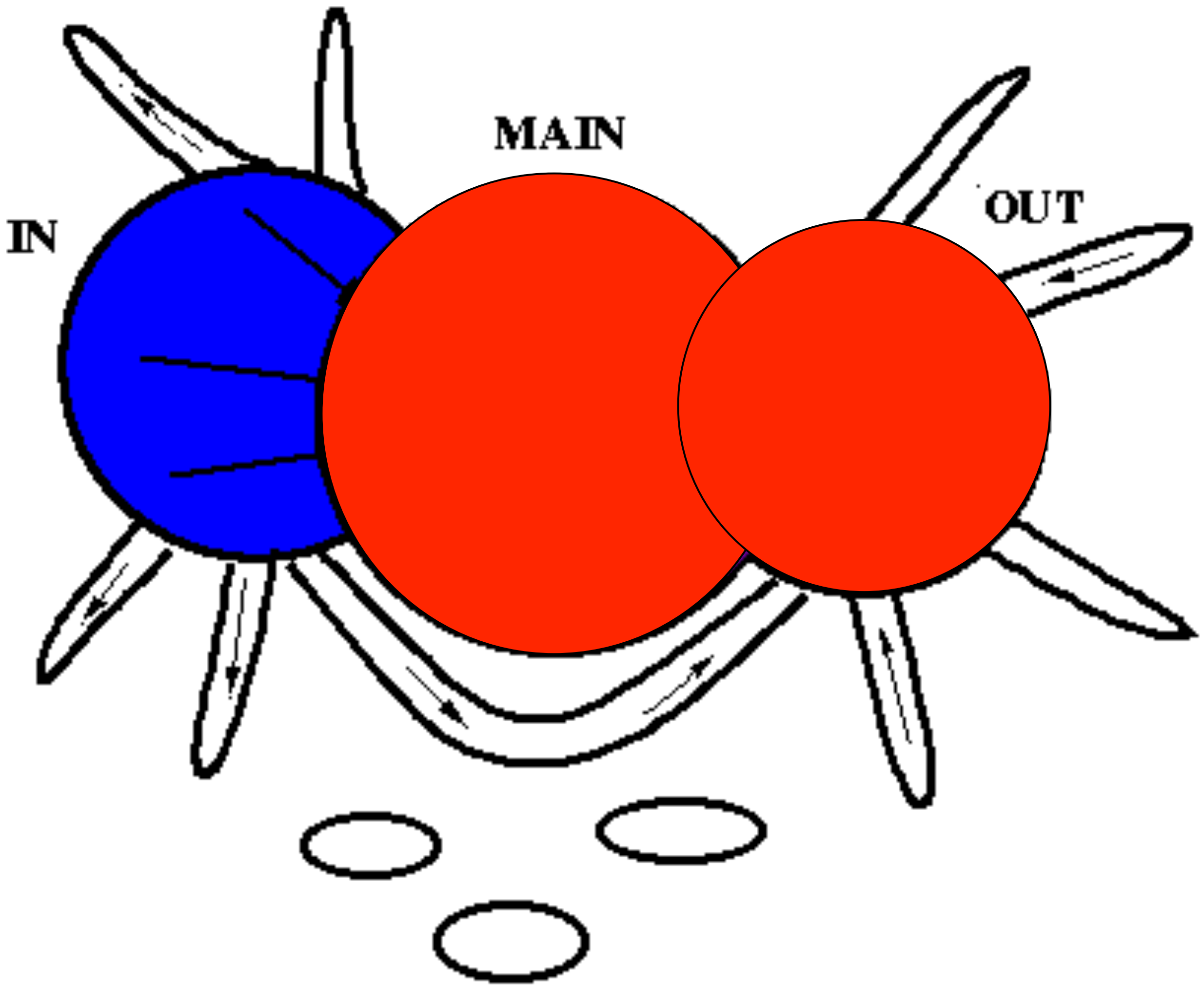


Self-similar &
Self-organizing

The Different Facets of the Web



The Structure of the Web



Motivation for Web Mining

- **The Dream of the Semantic Web**
  - Hypothesis: Explicit Semantic Information
  - Obstacle: Us
- **User Actions: Implicit Semantic Information**
  - It's free!
  - Large volume!
  - It's unbiased!
  - Can we capture it?
  - Hypothesis: Queries are the best source

Web Mining

- **Content:** text & multimedia mining
- **Structure:** link analysis, graph mining
- **Usage:** log analysis, query mining
- **Relate all of the above**
  - Web characterization
  - Particular applications

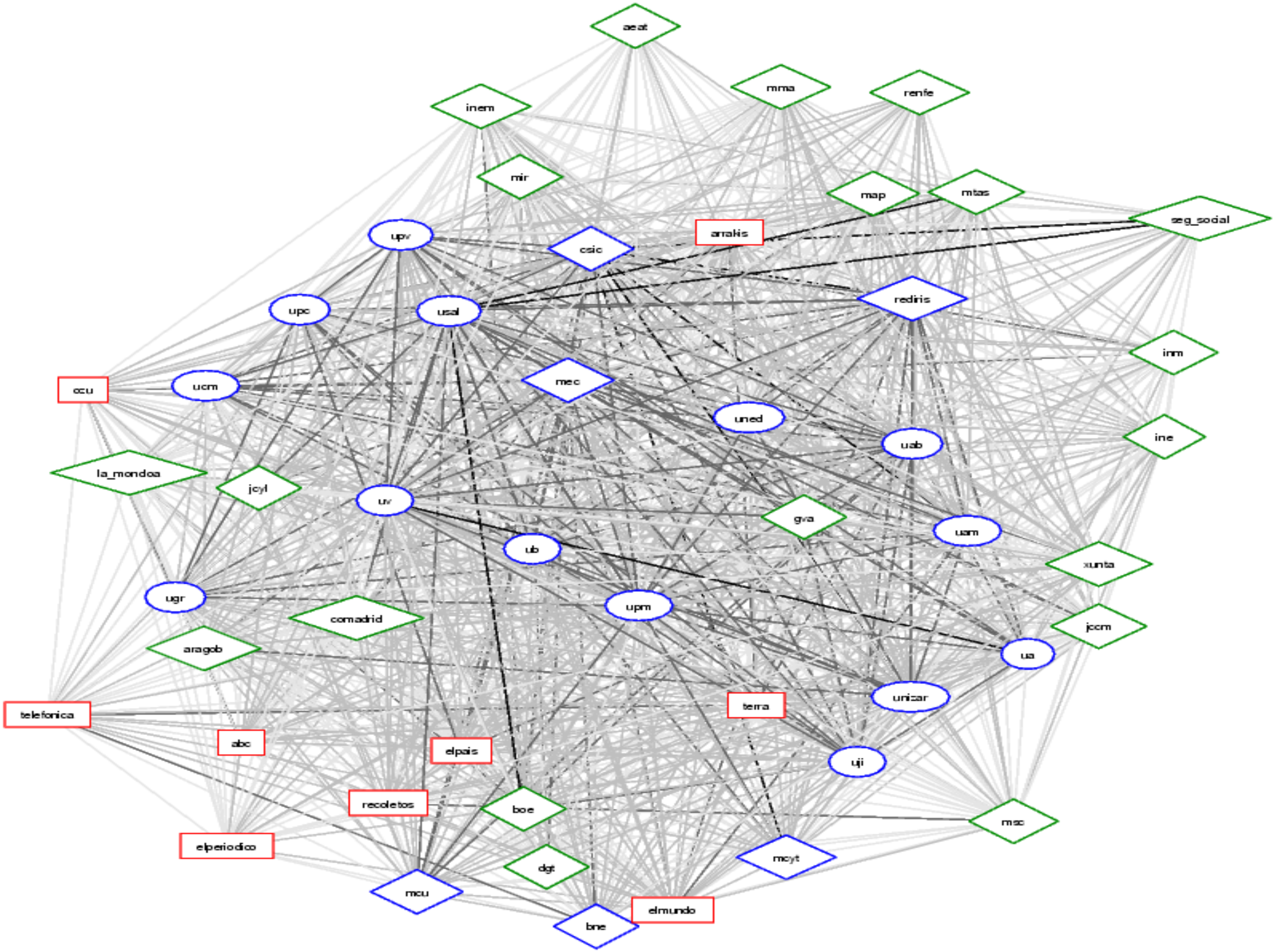
Dynamic

What for?

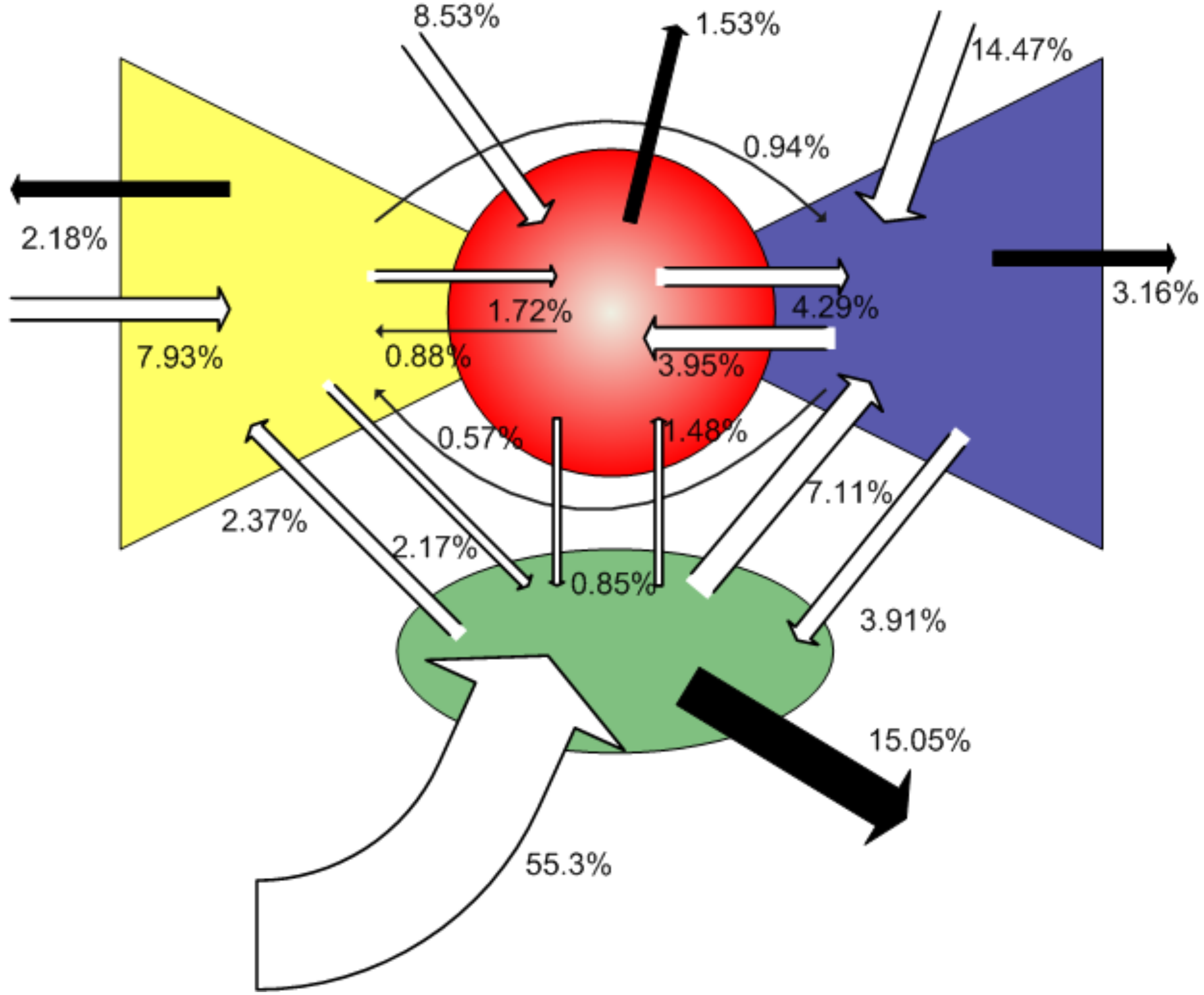
- The Web as an object
- User-driven Web design
- Improving Web applications
- Social mining
- .....

A Few Examples

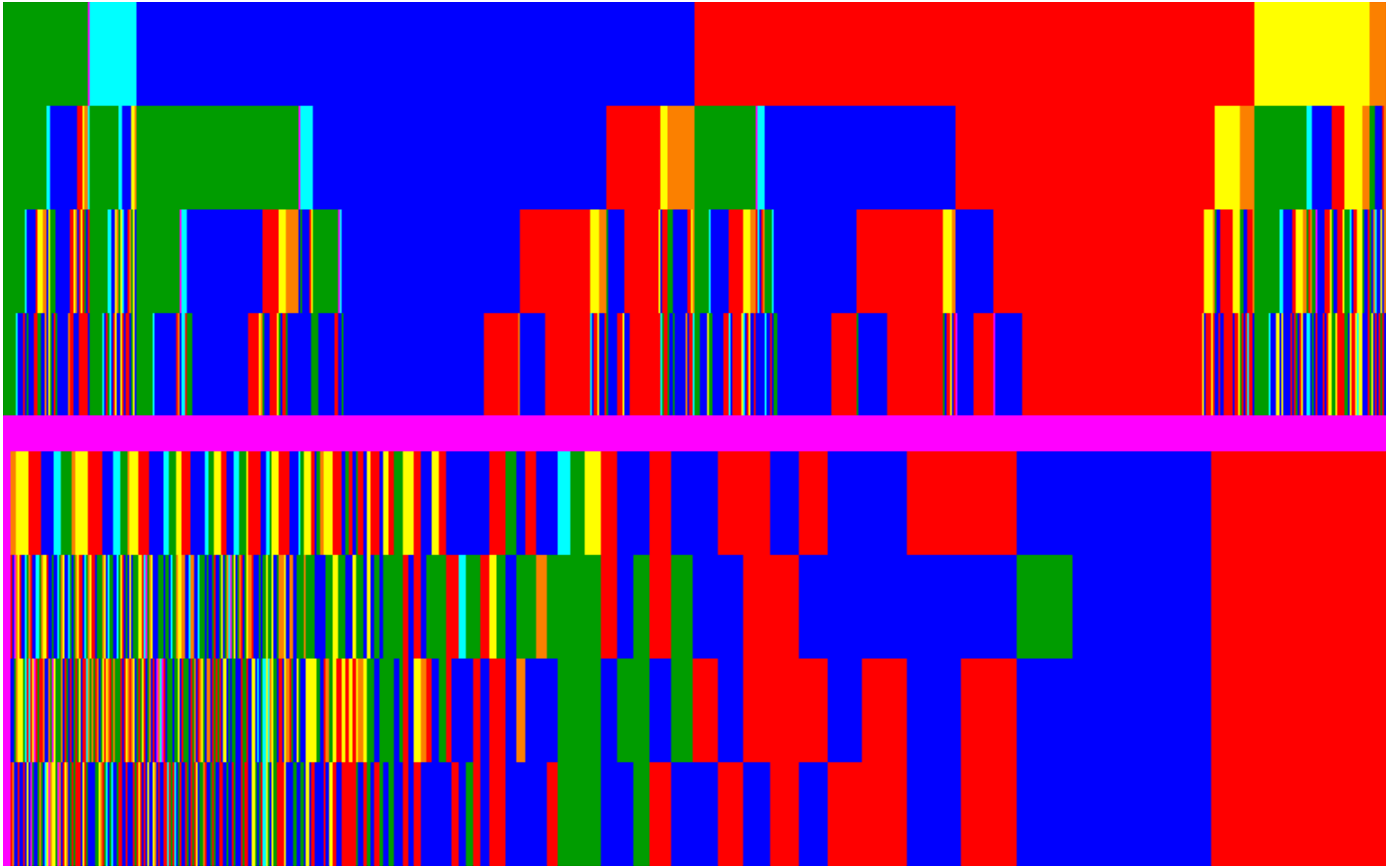
- Web Characterization of Spain
- Link Analysis
- Log Analysis
- Web Dynamics
- Social Mining


aeat
mma
renfe
inem
mir
map
mtas
seg_social
upv
csic
arrakis
upc
usal
rediris
inm
ocu
ucm
mec
uned
uab
ine
la_moncloa
jcyl
uv
gva
uam
ub
xunta
ugr
upm
comadrid
jccm
aragob
ua
unizar
telefonica
terra
abc
elpais
uji
recoletos
boe
msc
elperiodico
mcyt
dgt
mcu
bne
elmundo

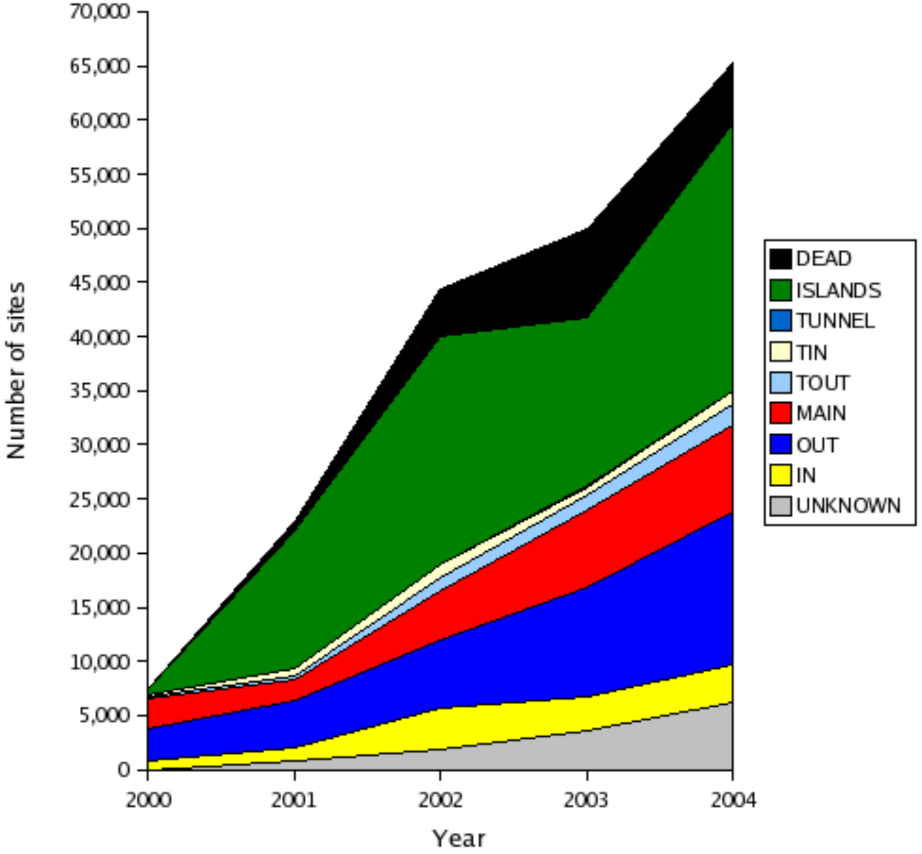
Structure Macro Dynamics

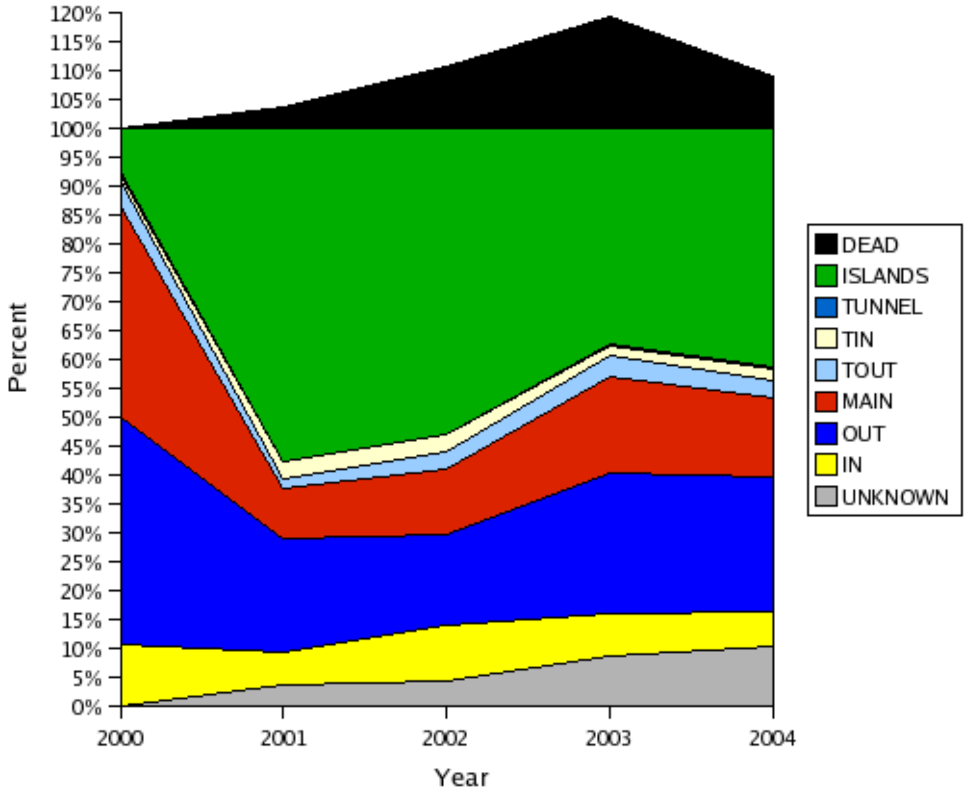


Structure Micro Dynamics

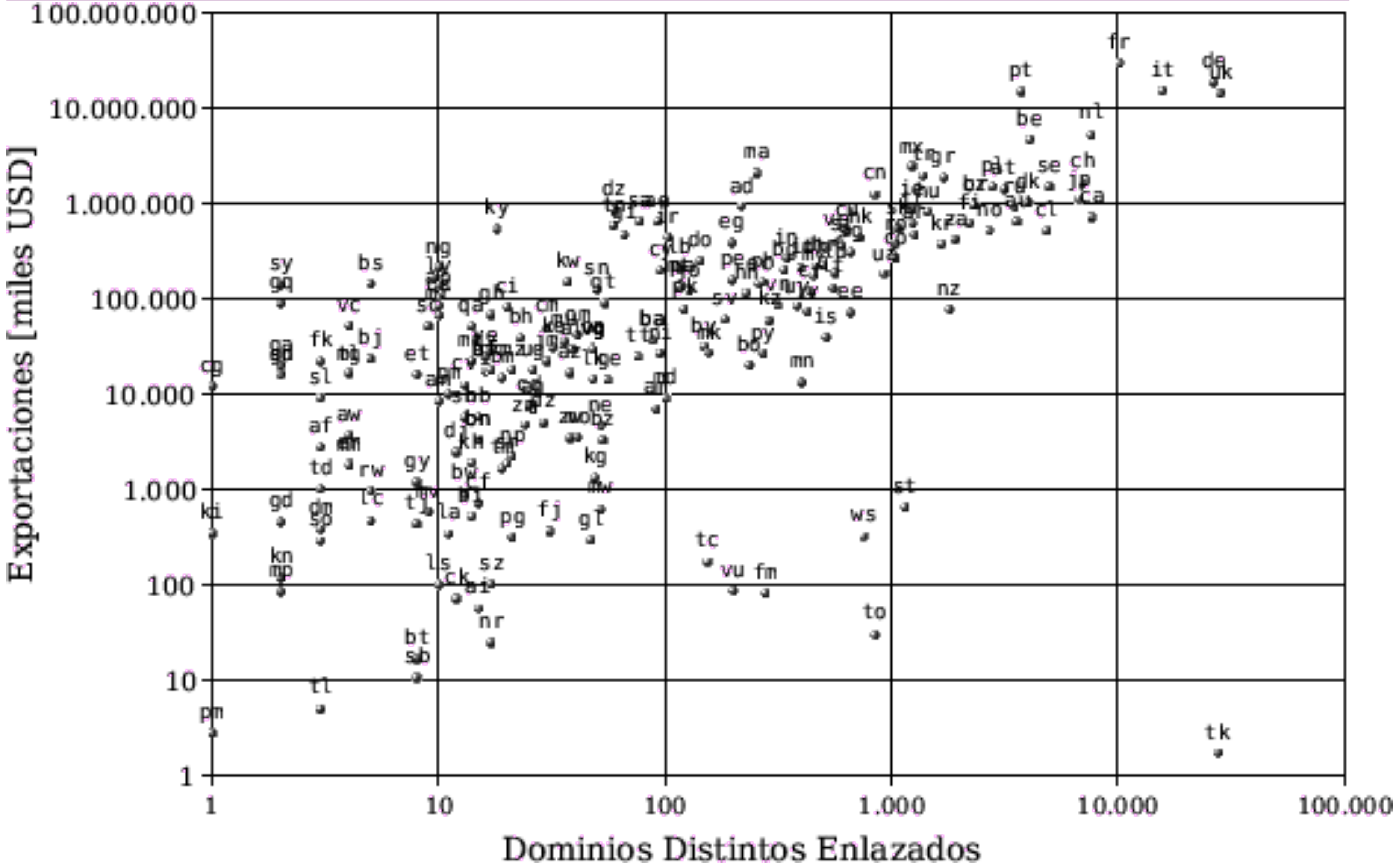


Size Evolution

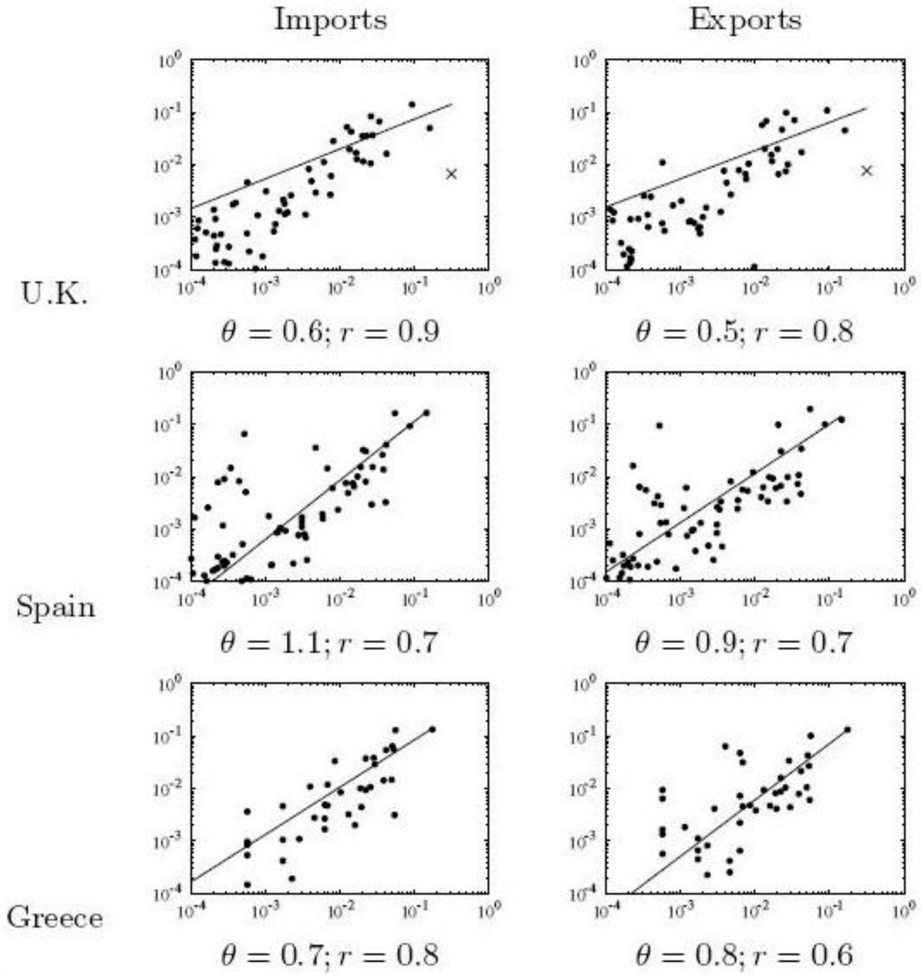


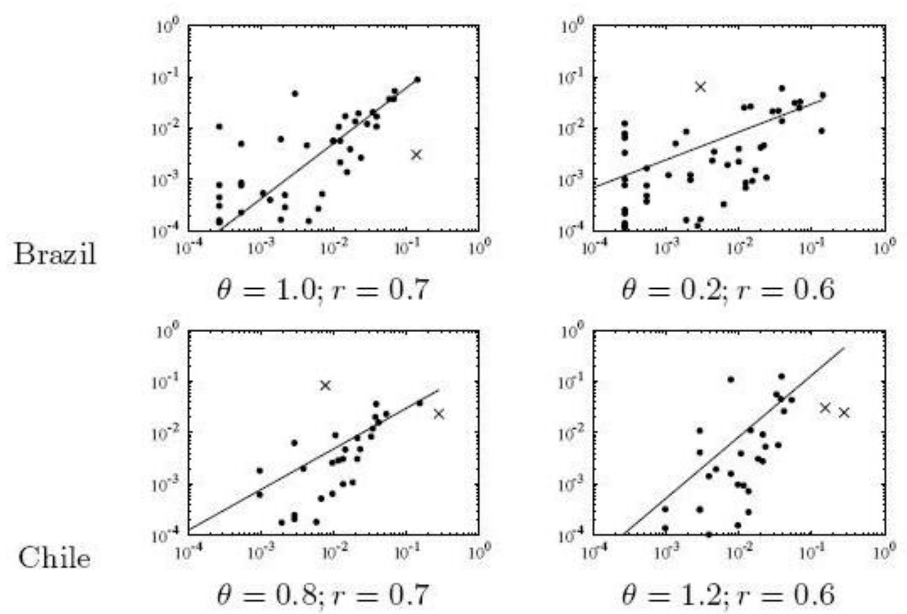


Mirror of the Society



Exports/Imports vs. Domain Links





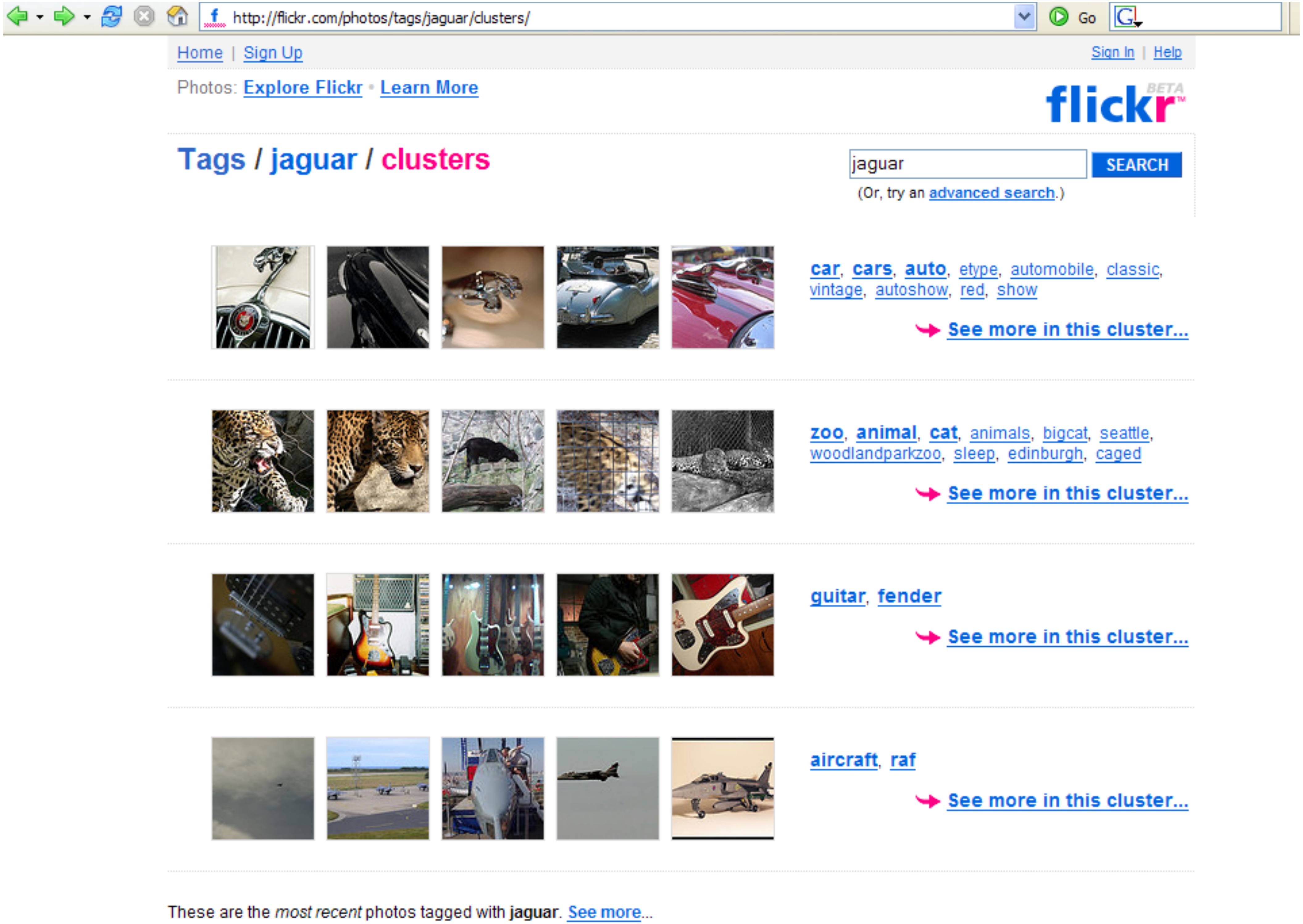
Baeza-Yates & Castillo, WWW2006

The Wisdom of Crowds

- **James Surowiecki, a *New Yorker* columnist, published this book in 2004**
  - **"Under the right circumstances, groups are remarkably intelligent"**
- **Importance of diversity, independence and decentralization**

Aggregating data

"large groups of people are smarter than an elite few, no matter how brilliant—they are better at solving problems, fostering innovation, coming to wise decisions, even predicting the future".


http://flickr.com/photos/tags/jaguar/clusters/
Go
Home | Sign Up
Sign In | Help
Photos: Explore Flickr • Learn More
flickr BETA
Tags / jaguar / clusters
jaguar
SEARCH
(Or, try an advanced search.)
car, cars, auto, etype, automobile, classic, vintage, autoshow, red, show
See more in this cluster...
zoo, animal, cat, animals, bigcat, seattle, woodlandparkzoo, sleep, edinburgh, caged
See more in this cluster...
guitar, fender
See more in this cluster...
aircraft, raf
See more in this cluster...
These are the most recent photos tagged with jaguar. See more...

Flickr: Geo-tagged pictures

The Wisdom of Crowds

- Popularity
- Diversity
- Quality
- Coverage

Long tail

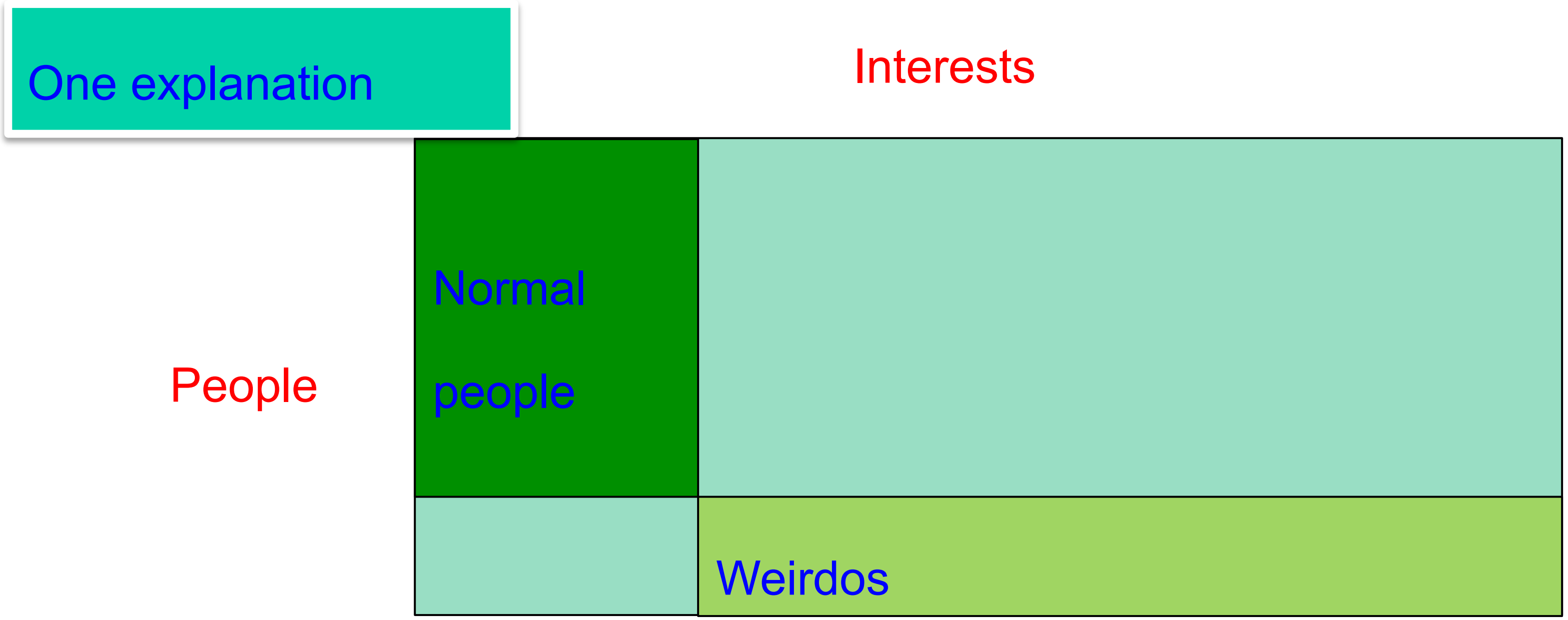
The Long Tail

Explore Flickr through tags

architecture art australia beach birthday blue bw california canada canon china christmas city concert england europe family festival flower flowers food france friends fun germany green italy japan london music nature new newyork night nikon nyc paris park party people portrait red sanfrancisco sky snow spain street summer sunset taiwan travel trip uk usa vacation water wedding white winter

Heavy tail of user interests

- **Many queries, each asked very few times, make up a large fraction of all queries**
  - Movies watched, blogs read, words used …



Heavy tail of user interests

- **Many queries, each asked very few times, make up a large fraction of all queries**
- **Applies to word usage, web page access …**
- **We are all partially eclectic**

The reality

Broder, Gabrilovich, Goel, Pang; WSDM 2009

Why the heavy tail matters

- **Not because the worst-sellers make a lot of money**
- **But because they matter to a lot of people**

The Wisdom of Crowds

- Crucial for Search Ranking
- Text: Web Writers & Editors
  - not only for the Web!
- Links: Web Publishers
- Tags: Web Taggers
- Queries: All Web Users!
  - Queries and actions (or no action!)

What is in the Web?

- Information
- Adult content

- + On-line casinos + Free movies + Cheap software + Buy a MBA diploma + Prescription - free drugs + V!-4-gra + Get rich now now now!!!

What is in the Web?

Spam is an Economic Activity

- Depending on the goal and the data spam is easier to generate
- Depending on the type & target data spam is easier to fight

- Disincentives for spammers?
  - Social
  - Economical
- Exploit the power of social networks and their work

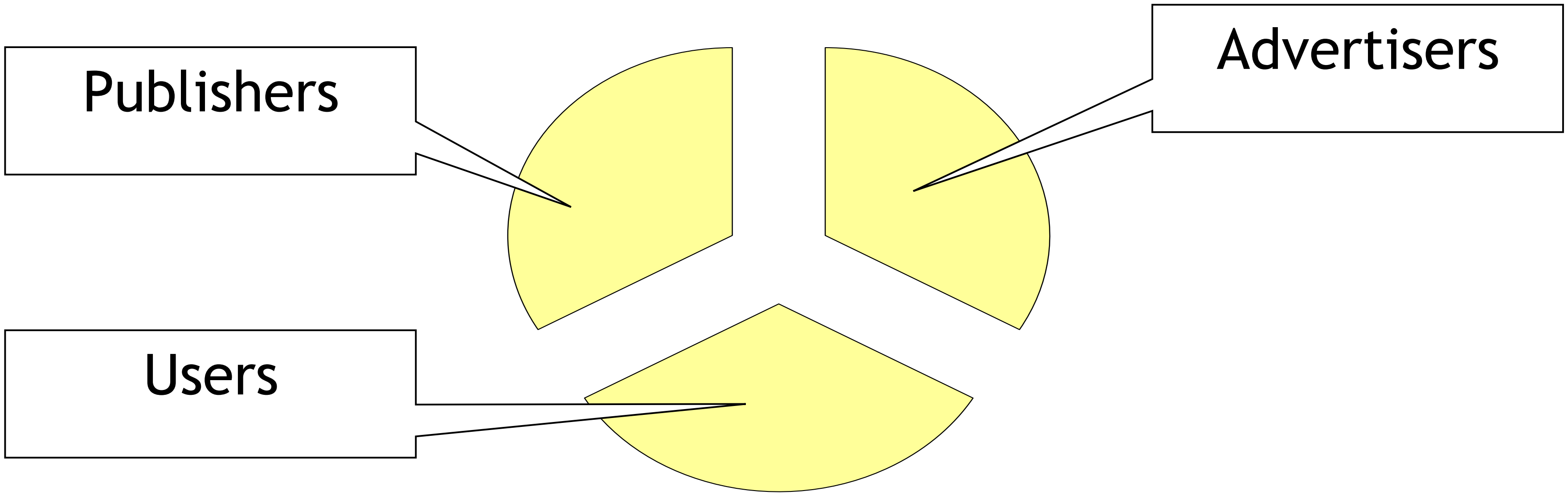
Current challenges (1)

- **Scraper spam**
  - Copies good content from other sites, adds monetization (most often Google AdSense)
  - Hard to identify at the page level (indistinguishable from original source), monetization not reliable clue (there is actually good content on the web that uses AdSense/YPN!)
- **Synthetic text**
  - Boilerplate text, randomized, built around key phrases
  - Avoids duplicate detection
- **Query-targeted spam**
  - Each page targets a single tail query (anchortext, title, body, URL). Often in large auto-constructed hosts, host-level analysis most helpful
- **DNS spam**

Current challenges (2)

- **Blog spam**
  - Continued trend toward blog “ownership” rather than comment spam
  - Orthogonal to other categories (scrapers, synthesizers). Just a hosting technique, plus exploiting blog interest
- **Example:**
  - 68,000 blogspot.com hosts all generated by the same spammer
    - 1) nursingschoolresources.blogspot.com
      2) transplantresources.blogspot.com
      ..
      67,798) beachesresourcesforyou.blogspot.com
      67,799) startrekresourcesforyou.blogspot.com

Content match = meeting of Publishers, Advertisers, Users



and Spammers! Grrr...

Contextual ads

Contextual ads

Click spam

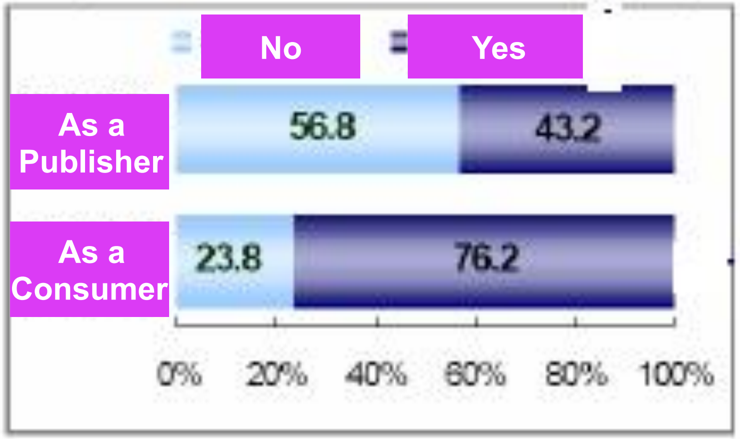
- **Rival click fraud:** Rival of advertising company employs clickers for clicking through ads to exhaust budget
- ~~**Publisher click fraud:** Publisher employs clickers to reap per click revenue from ads shown by search firm~~
- **Bidder click fraud:** Keyword bidders employ clickers to raise rate used in (click-thru-rate * bid) ranking used to allocate ad space in search engines (or to pay less!)

Other Possible Ad Spam

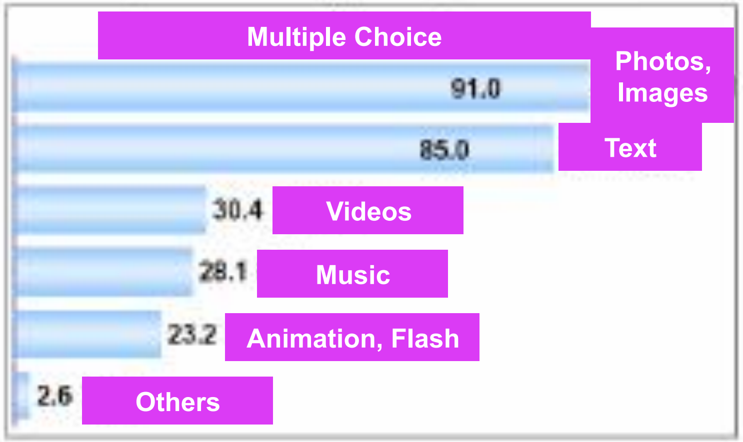
- **Rival buys misleading or fraudulent ads**
  - Queries
  - Bids
  - Ads
- **Rival submits queries that brings up competitor ad but without clicking on it**
  - *Reduces* rival's CTR and hence its ranking for ad space

Internet UGC (User Generated Content)

Have you experienced UGC?

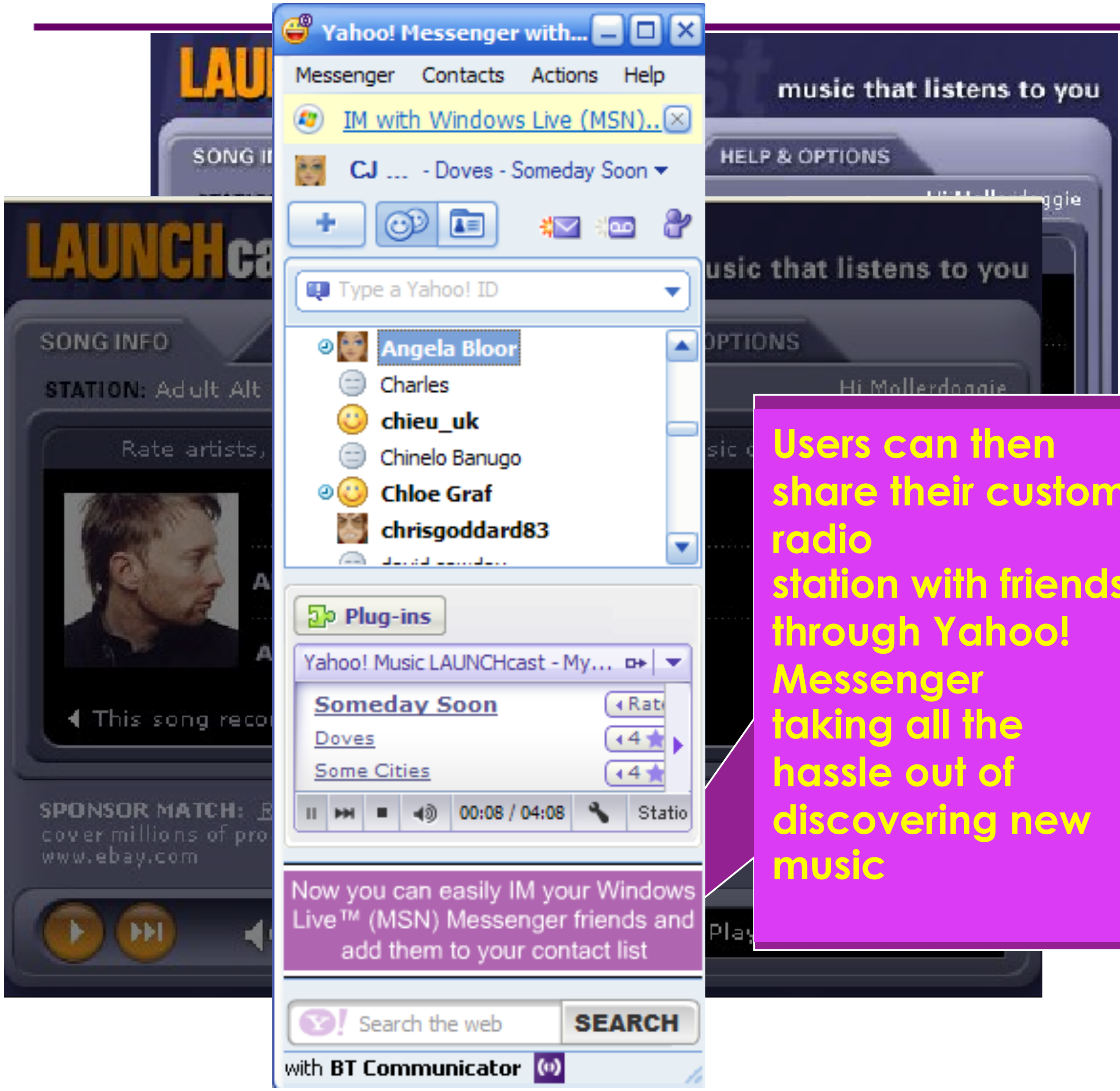


Types of Content



Source National Internet Development Agency Report in June, 2006 (South Korea)

Simple acts create value and opportunity



Using a system of user-assigned ratings, LAUNCHcast builds up a profile of preferences for each individual..

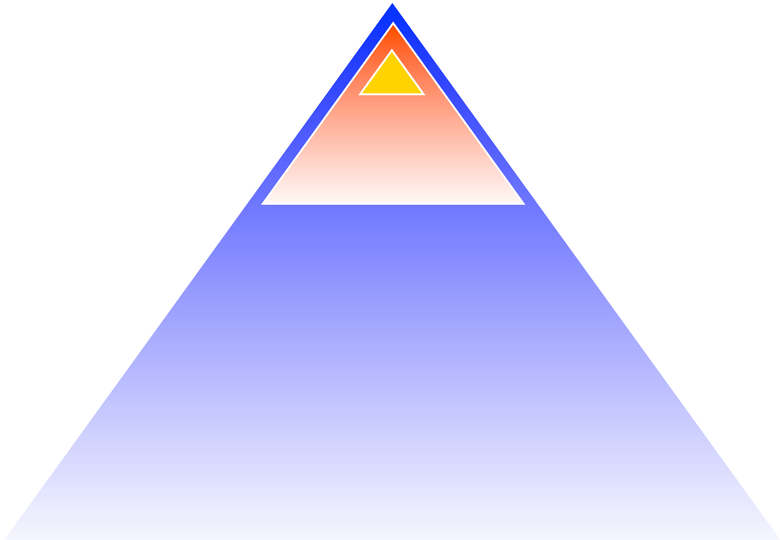
Users can then share their custom radio station with friends through Yahoo! Messenger taking all the hassle out of discovering new music

The more ratings users make, the more intelligent the radio becomes.

We have over 6 billion ratings

LAUNCHcast = music that listens to you

Community Dynamics



1 creators

10 synthesizers

100 consumers

Next generation products will blur distinctions between Creators, Synthesizers, and Consumers

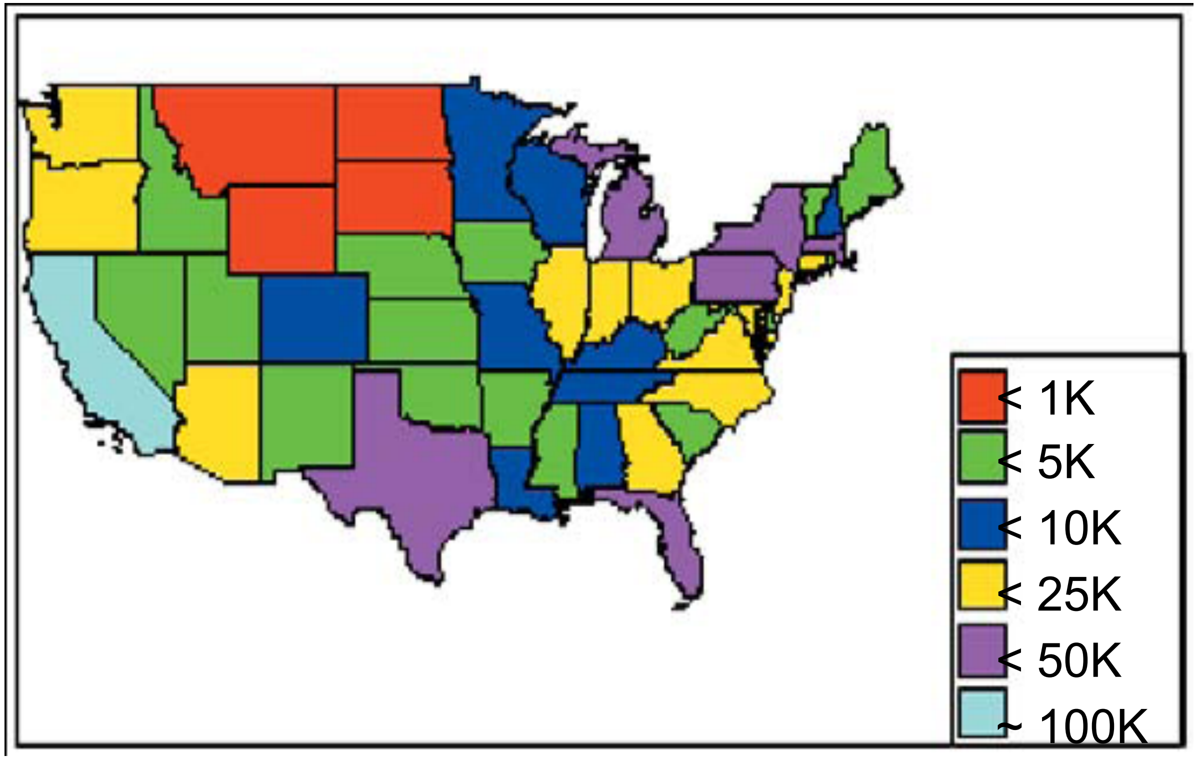
Example: Launchcast

Every act of consumption is an implicit act of production that requires no incremental effort…

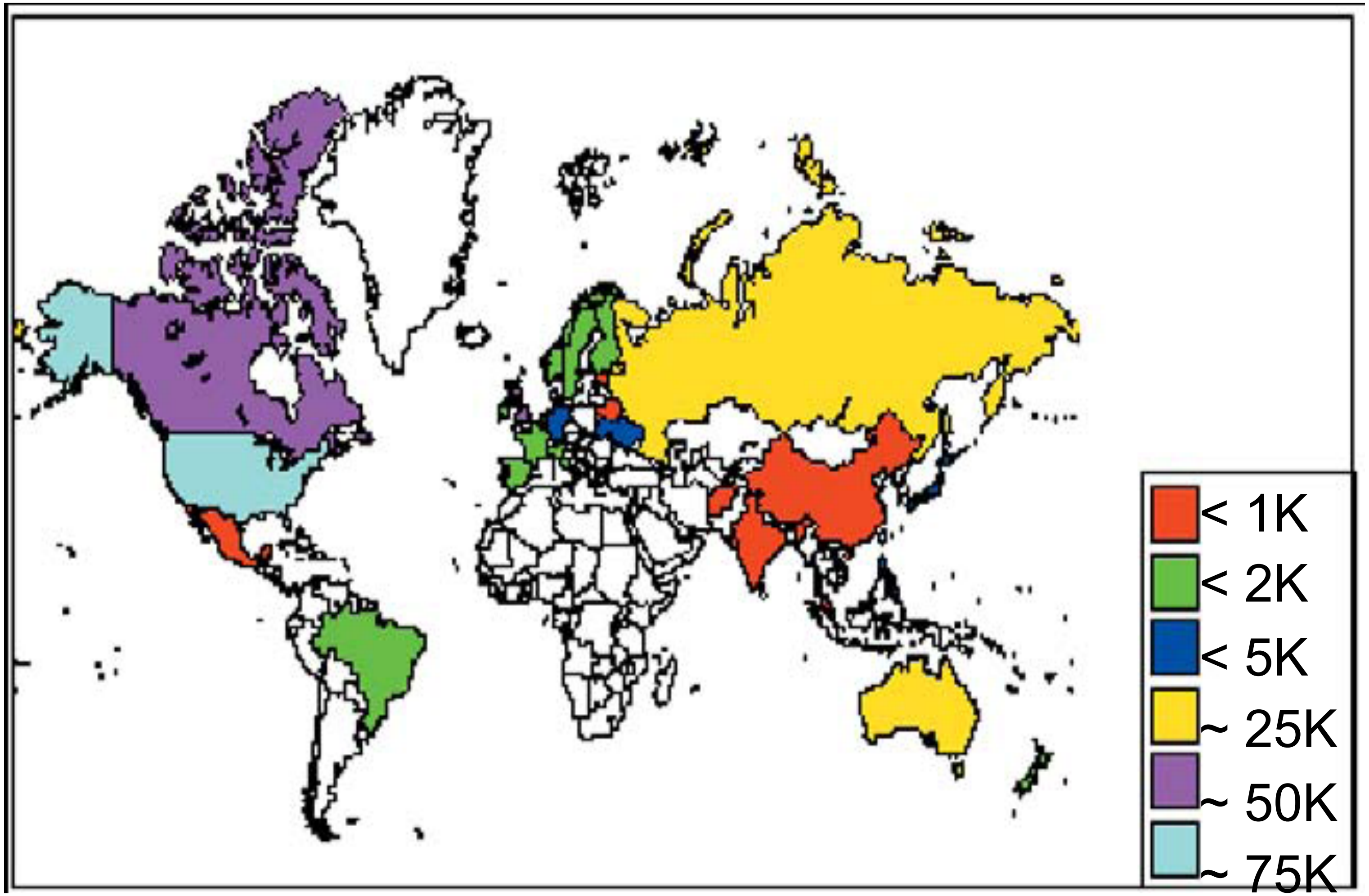
Listening itself implicitly creates a radio station…

Community Geography:

Live Journal bloggers in US



LJ bloggers world-wide



Who are they?

Age	%	Representative interests
1 to 3	0.5	treats, catnips, daddy, mommy, purring, mice, playing, napping, scratching, milk
13 to 15	3.5	webdesigning, Jeremy Sumpter, Chris Wilson, Emma Watson, T. V., Tom Felton, FUSE, Adam Carson, Guyz, Pac Sun, mall, going online
16 to 18	25.2	198{6,7,8}, class of 200{4,5}, dream street, drama club, band trips, 16, Brave New Girl, drum major, talkin on the phone, highschool, JROTC
19 to 21	32.8	198{3,5}, class of 2003, dorm life, frat parties, college life, my tattoo, pre-med
22 to 24	18.7	198{1,2}, Dumbledore's army, Midori sours, Long island iced tea, Liquid Television, bar hopping, disco house, Sam Adams, fraternity, He-Man, She-Ra
25 to 27	8.4	1979, Catherine Wheel, dive bars, grad school, preacher, Garth Ennis, good beer, public radio
28 to 30	4.4	Hal Hartley, geocaching, Camarilla, Amtgard, Tivo, Concrete Blonde, motherhood, SQL, TRON
31 to 33	2.4	my kids, parenting, my daughter, my wife, Bloom County, Doctor Who, geocaching, the prisoner, good eats, herbalism
34 to 36	1.5	Cross Stitch, Thelema, Tivo, parenting, cubs, role-playing games, bicycling, shamanism, Burning Man
37 to 45	1.6	SCA, Babylon 5, pagan, gardening, Star Trek, Hogwarts, Macintosh, Kate Bush, Zen, tarot
46 to 57	0.5	science fiction, wine, walking, travel, cooking, politics, history, poetry, jazz, writing, reading, hiking
> 57	0.2	death, cheese, photography, cats, poetry

The Process

- Data recollection: crawling, log keeping
- Data cleaning and anonymization
- Data statistics and data modeling

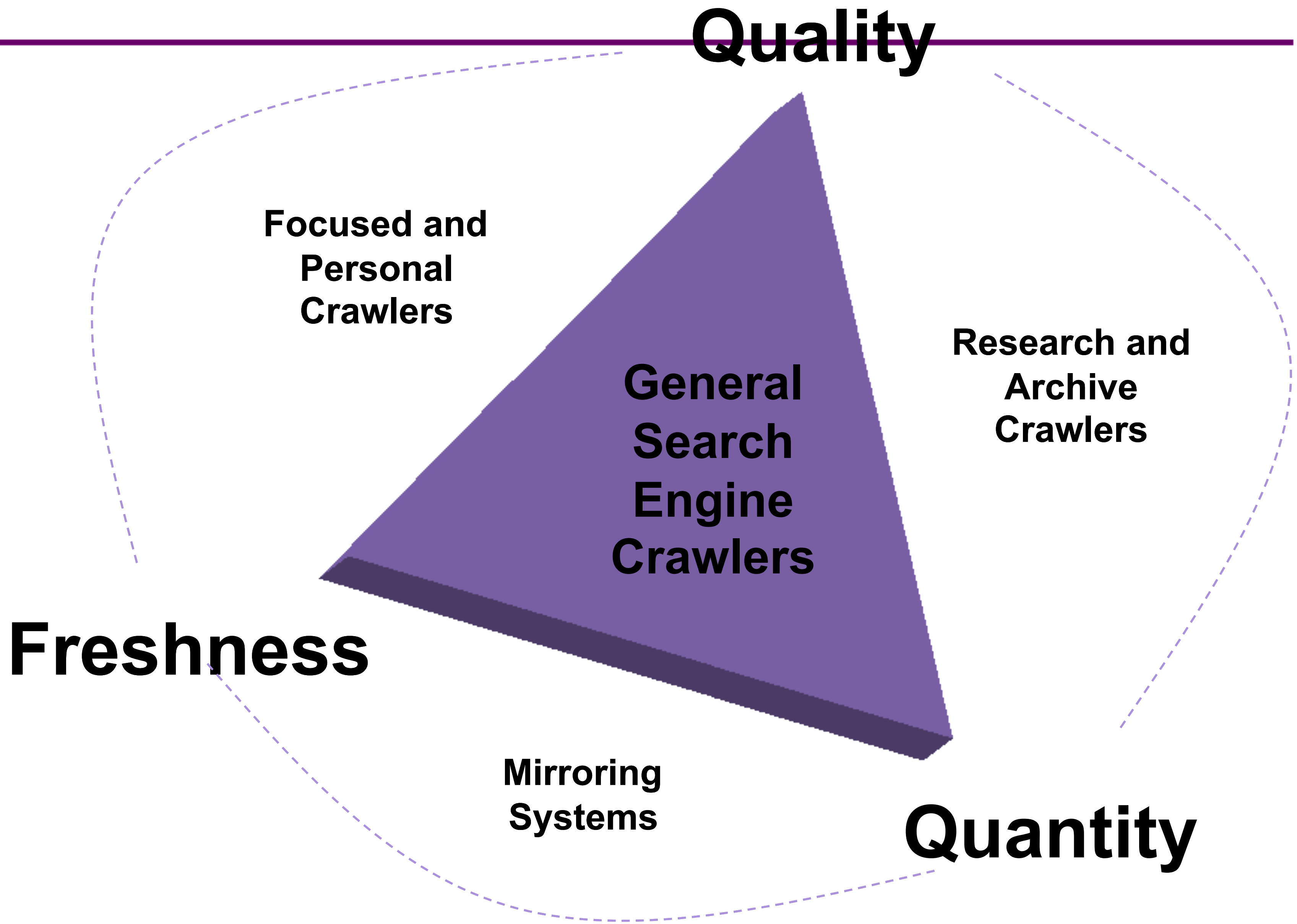
Data Recollection

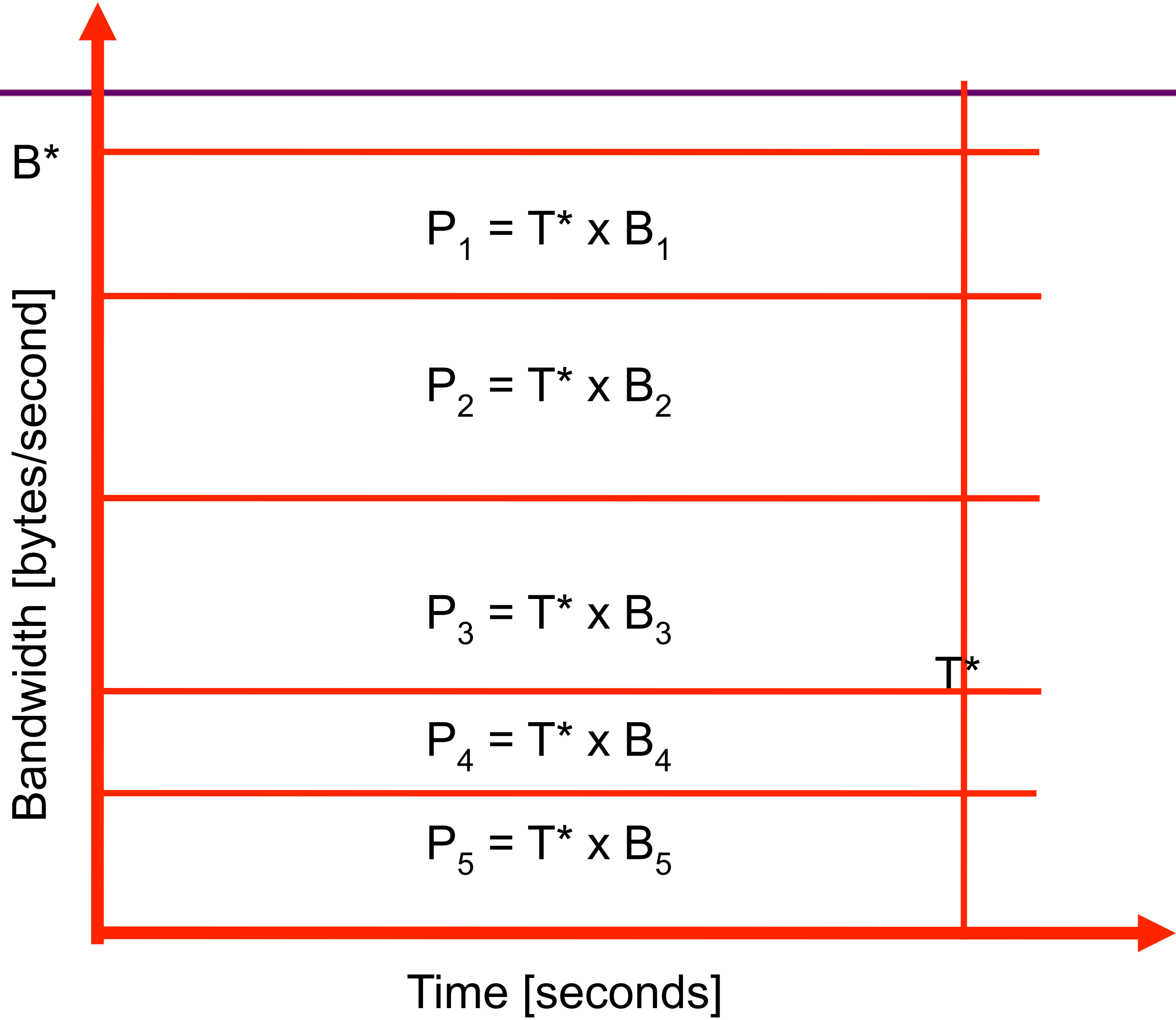
- **Content and structure: Crawling**
- **Usage: Logs**
  - Web Server logs
  - Specific Application logs

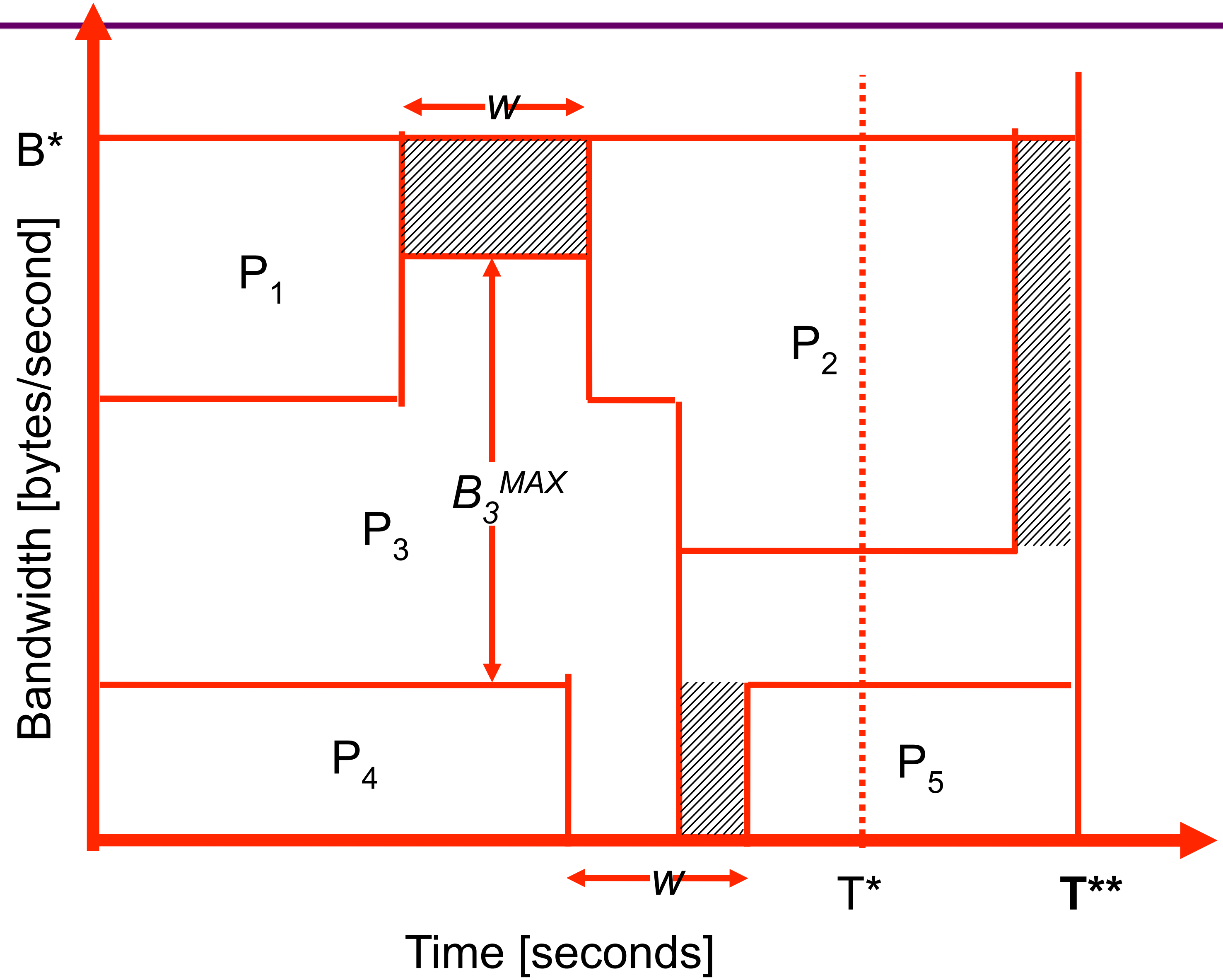
Crawling

- NP-Hard Scheduling Problem
- Different goals
- Many Restrictions
- Difficult to define optimality
- No standard benchmark

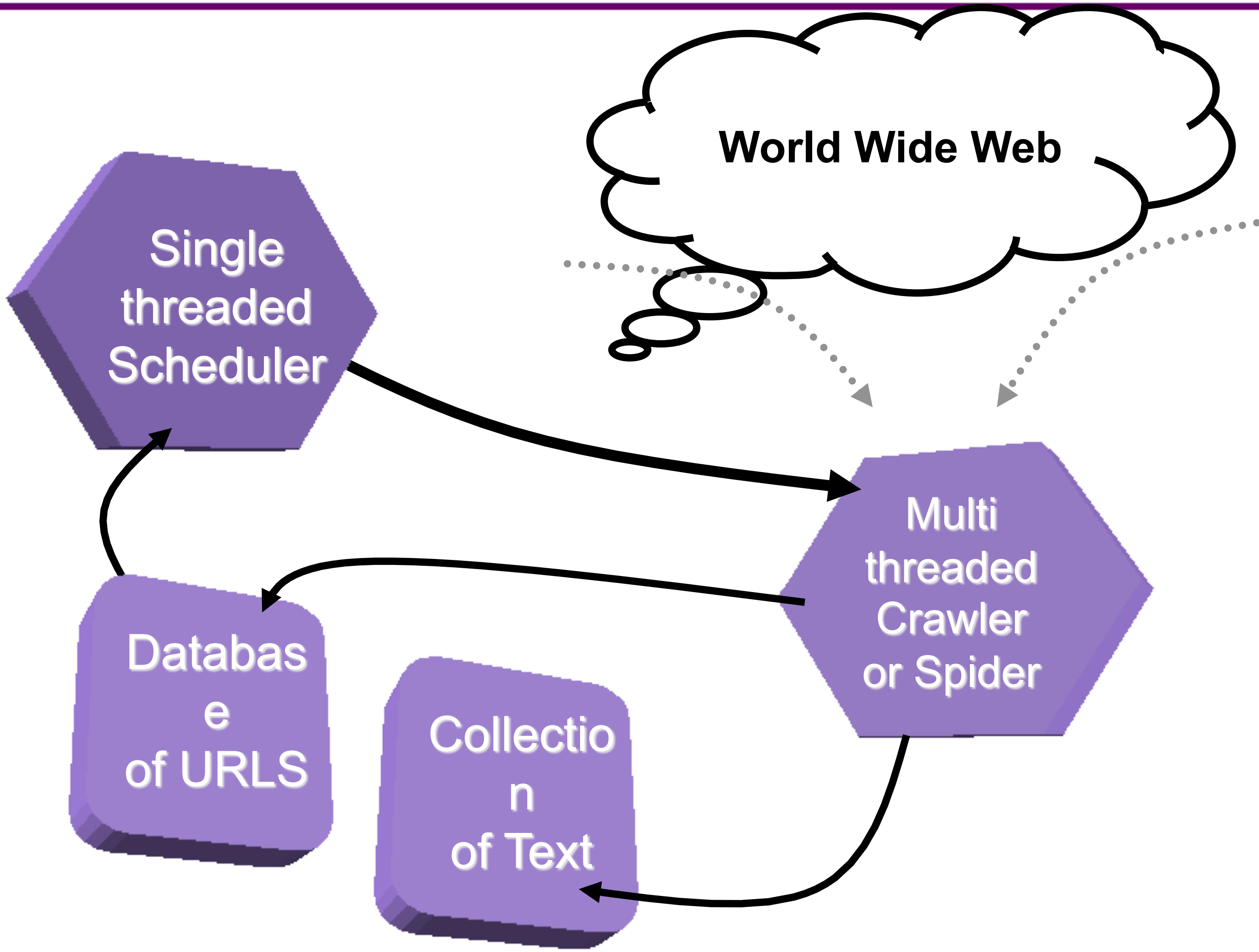
Crawling Goals

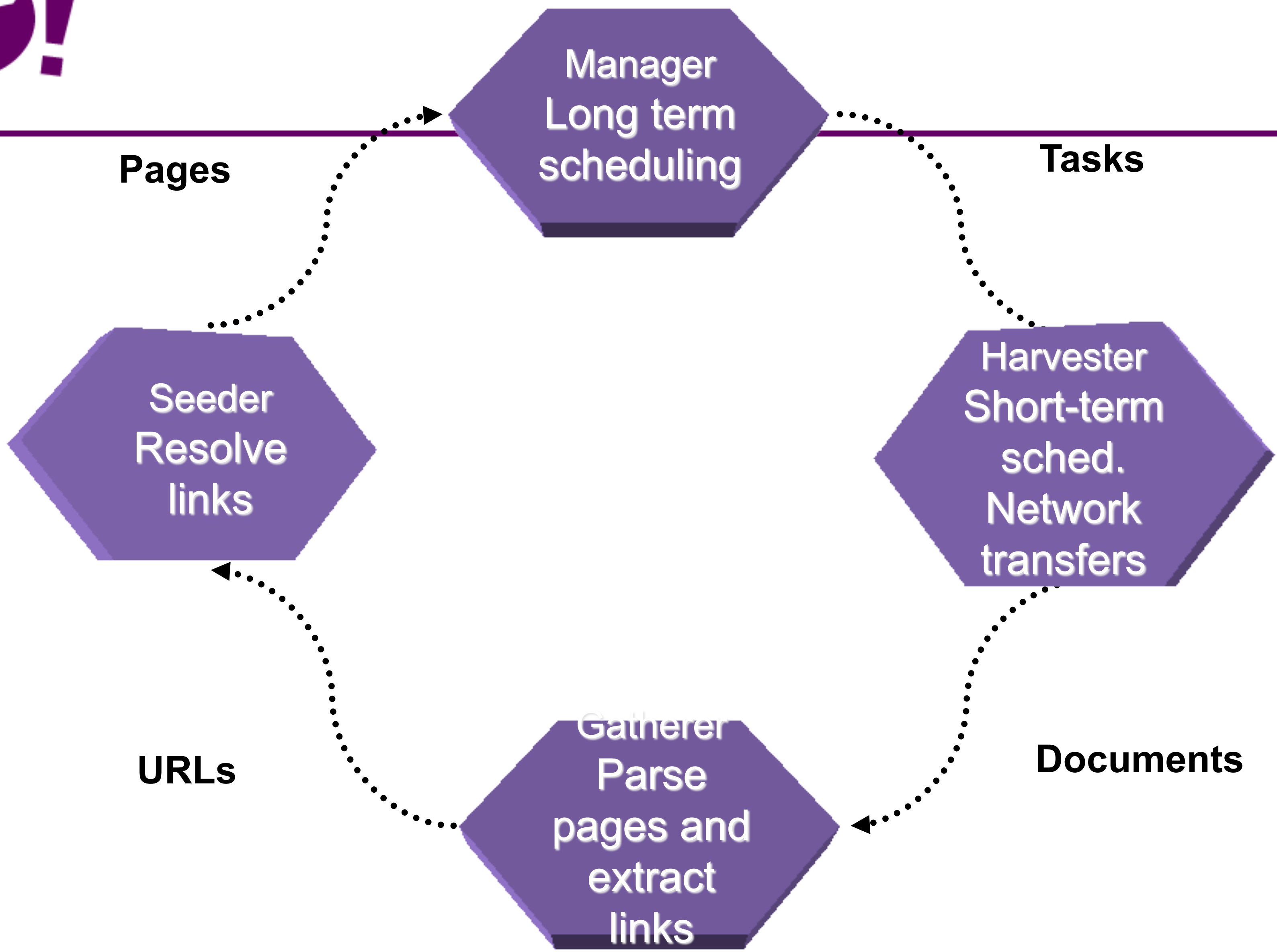


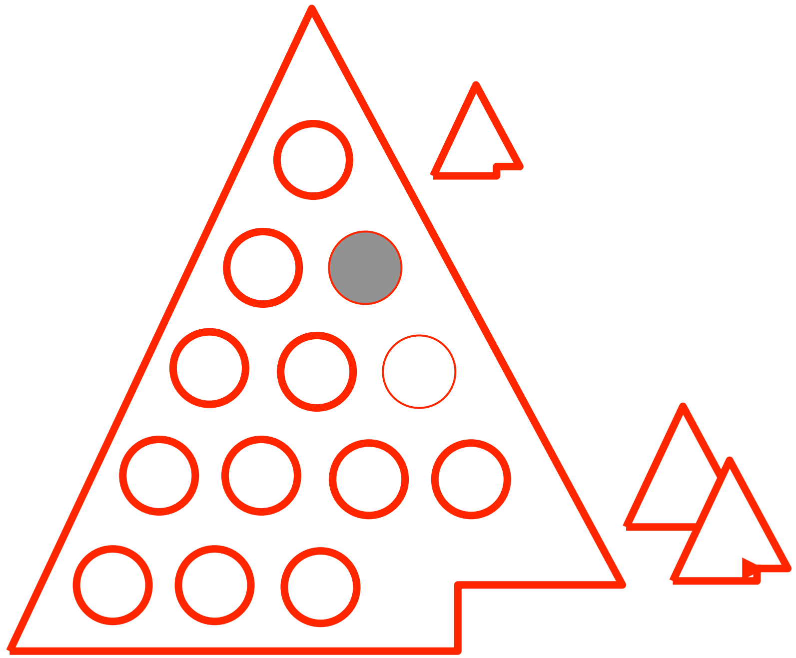

B*
$P_1 = T^* \times B_1$
$P_2 = T^* \times B_2$
$P_3 = T^* \times B_3$
T*
$P_4 = T^* \times B_4$
$P_5 = T^* \times B_5$
Bandwidth [bytes/second]
Time [seconds]

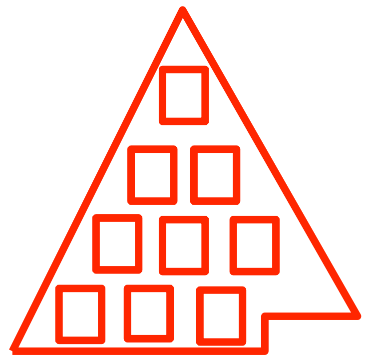

w
B*
P1
P2
P3
B_3^{MAX}
P4
P5
w
T*
T**
Bandwidth [bytes/second]
Time [seconds]

Software Architecture




Manager
Long term
scheduling
Tasks
Harvester
Short-term
sched.
Network
transfers
Documents
Gatherer
Parse
pages and
extract
links
URLs
Seeder
Resolve
links
Pages





Queue of Web pages
for each site
(short-term scheduling)

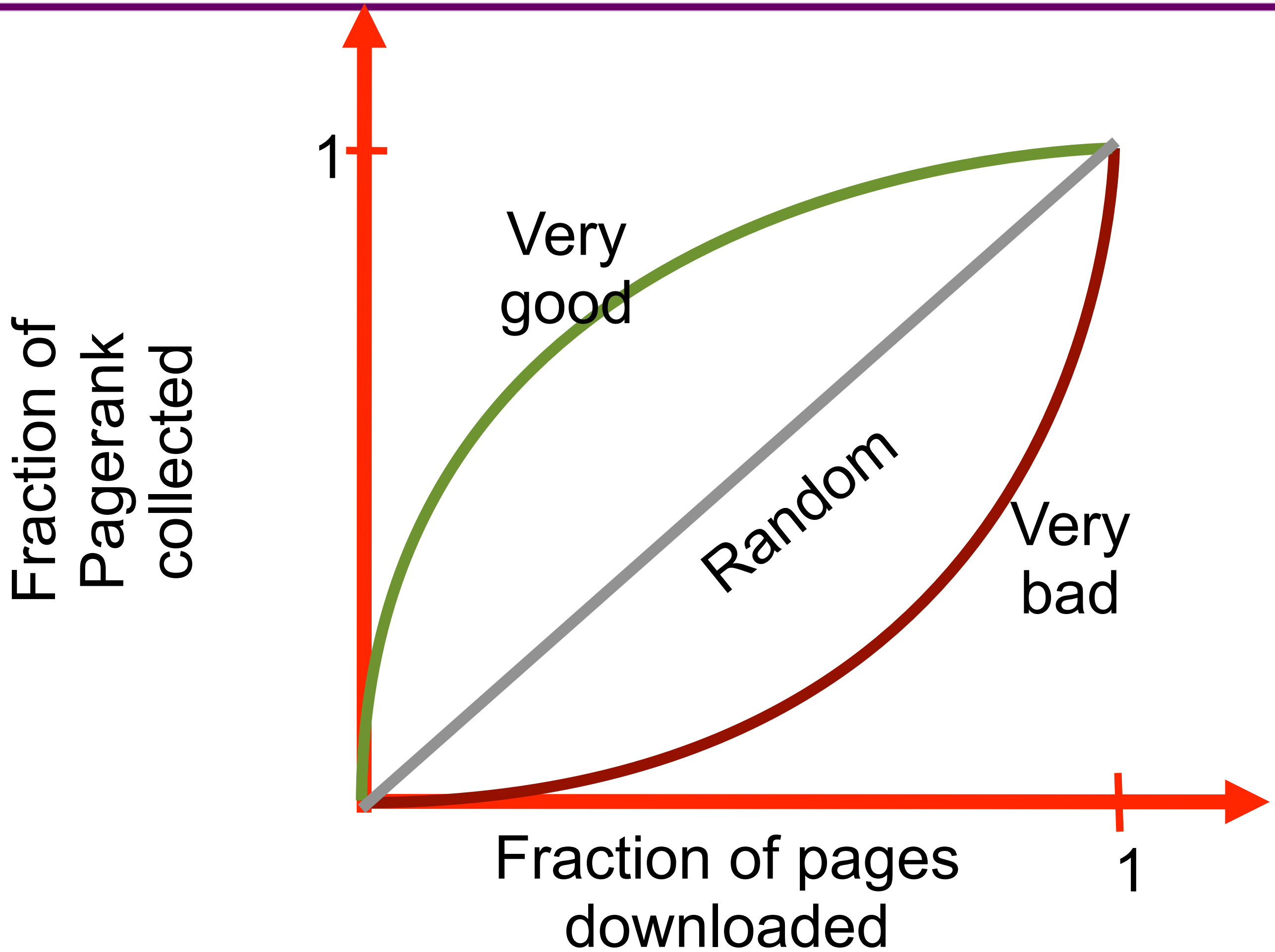
Queue of Web sites
(long-term scheduling)

Formal Problem

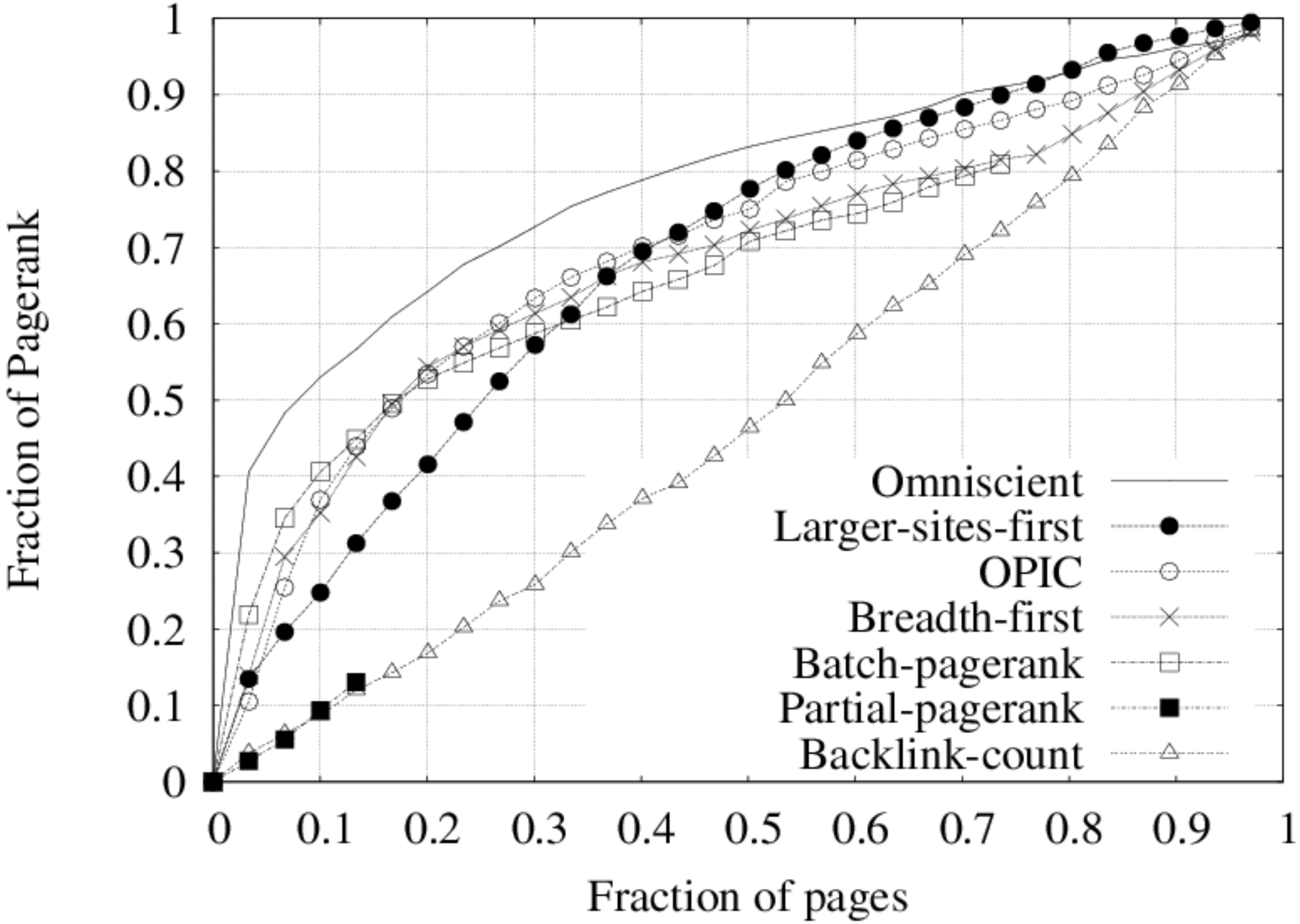
- **Find a sequence of page requests *(p,t)* that:**
  - Optimizes a function of the volume, quality and freshness of the pages
  - Has a bounded crawling time
  - Fulfils politeness
  - Maximizes the use of local bandwidth
- **Must be on-line: how much knowledge?**

Crawling Heuristics

- **Breadth-first**
- **Ranking-ordering**
  - PageRank
- **Largest Site-first**
- **Use of:**
  - Partial information
  - Historical information
- **No Benchmark for Evaluation**

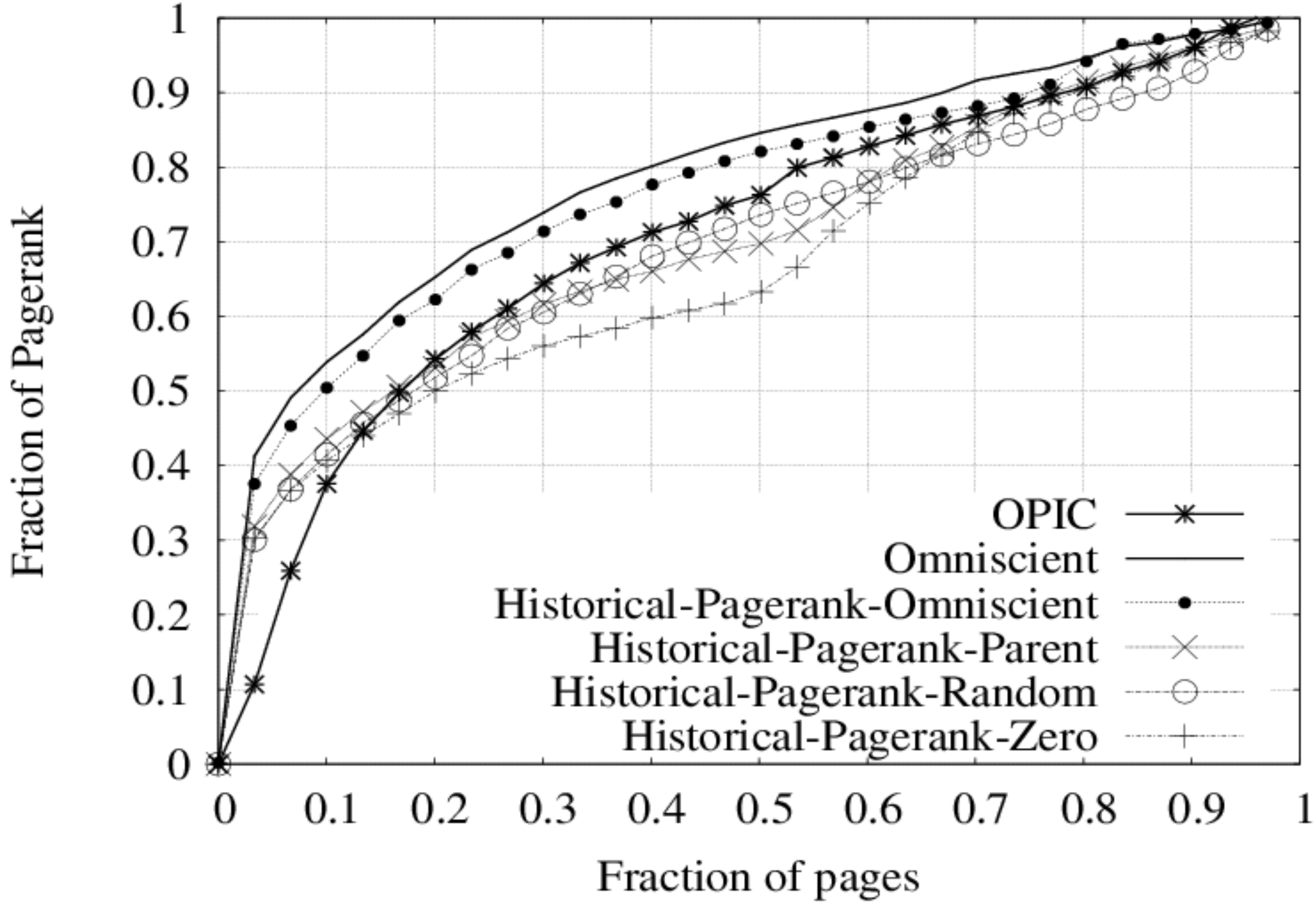

1
Very
good
Fraction of
Pagerank
collected
Random
Very
bad
Fraction of pages
downloaded
1

No Historical Information

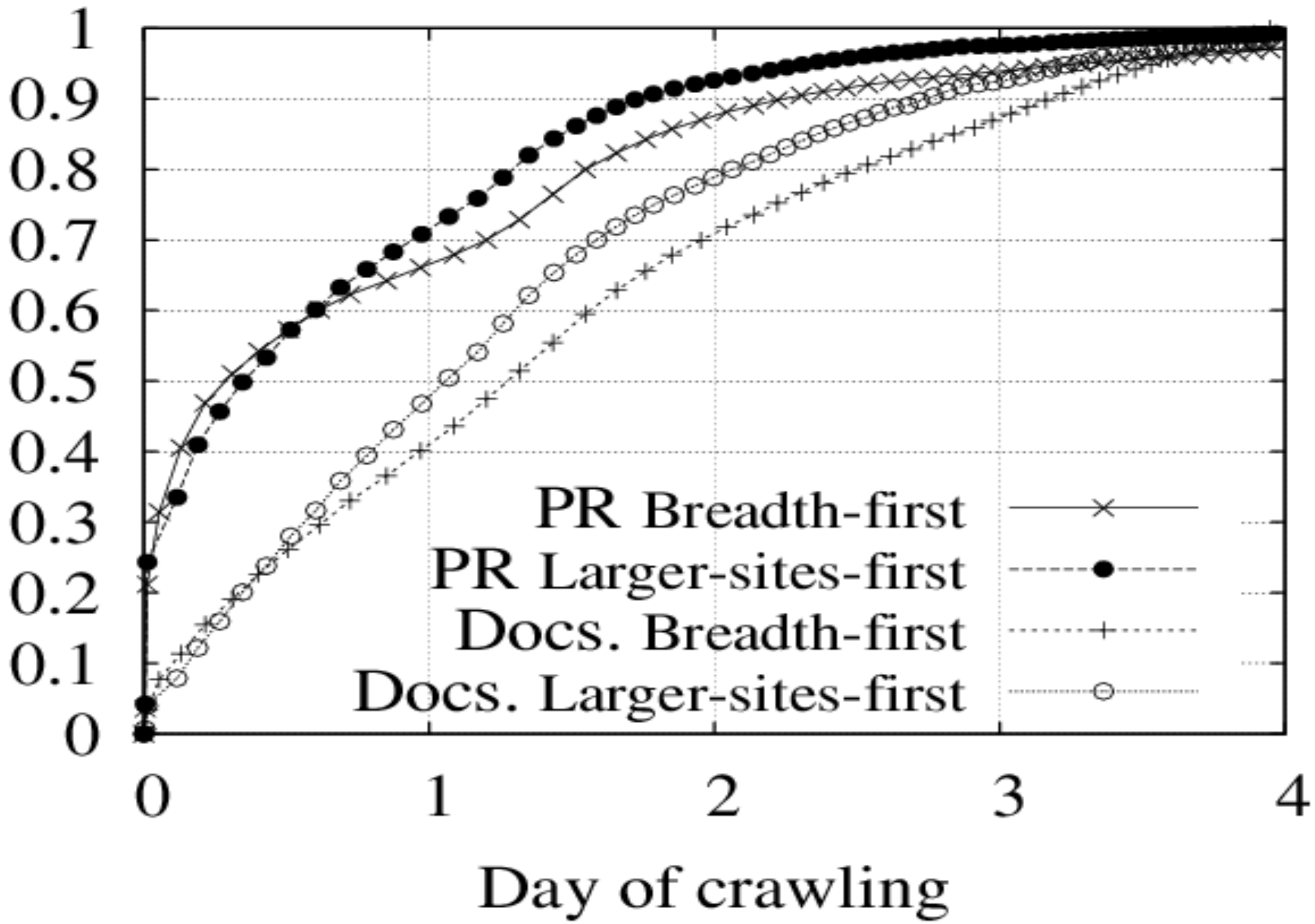


Baeza-Yates, Castillo, Marin & Rodriguez, WWW2005

Historical Information



Validation in the Greek domain



Data Cleaning

- **Problem Dependent**
- **Content: Duplicate and spam detection**
- **Links: Spam detection**
- **Logs: Spam detection**
  - Robots vs. persons

Data Processing

- **Structure: content, links and logs**
  - XML, relational database, etc.
- **Usage mining:**
  - Anonymize if needed
  - Define sessions

Data Characteristics

- **Yahoo! as a Case Study**
  - –Data Volume
  - –Data Types

Example: Yahoo! (2006)

24 languages, 20 countries

- **> 4 billion page views per day (largest in the world)**
- **> 500 million unique users each month (half the Internet users!)**
- **> 250 million mail users (1 million new accounts a day)**
- **95 million groups members**
- **7 million moderators**
- **4 billion music videos streamed in 2005**

- **20 Pb of storage (20M Gb)**
  - US Library of congress every day (28M books, 20TB)
- **12 Tb of data processed per day**
- **7 billion song ratings**
- **2 billion photos stored**
- **2 billion Mail+Messenger sent per day**

Crawled Data

- **WWW**
  - Web Pages & Links
  - Blogs
  - Dynamic Sites

heterogeneous,
large,
dangerous

- **Sales Providers (Push)**
  - Advertising
  - Items for sale: Shopping, Travel, etc.

very high quality
& structure,
expensive,
sparse,
safe

- **News Index**
  - RSS Feeds
  - Contracted information

high quality,
sparse,
redundant

Produced data

- **Yahoo's Web**
  - Ygroups
  - YCars, YHealth, Ytravel

homogeneous,
high quality,
safer,
highly structured

- **Produced Content**
  - Edited (news)
  - Purchased (news)

Trusted,
high quality,
sparse

- **Direct Interaction:**
  - Tagged Content
    - Object tagging (photos, pages, ?)
    - Social links
  - Question Answering

Ambiguous
semantics?
trust?
quality?

"Information Games"
(e..g. www.espgame.org)

Observed Data

- **Query Logs**
  - spelling, synonyms, phrases (named entities), substitutions

good quality,
sparse,
power law

- **Click-Thru**
  - relevance, intent, wording

good quality,
sparse,
mostly safe

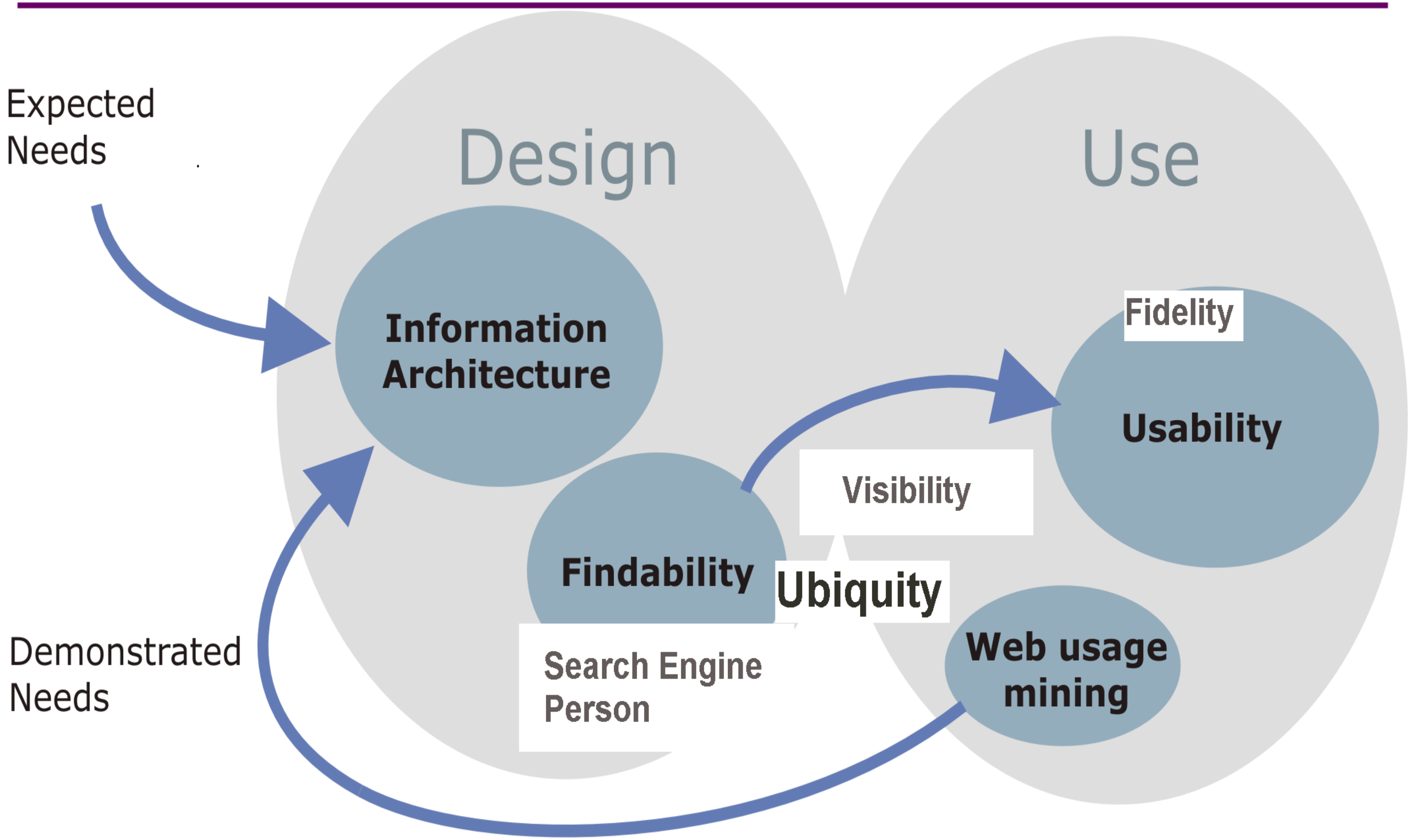
- **Advertising**
  - relevance, value, terminology

Trusted,
high quality,
homogeneous,
structured

- **Social**
  - links, communities, dialogues...

trust?
quality?

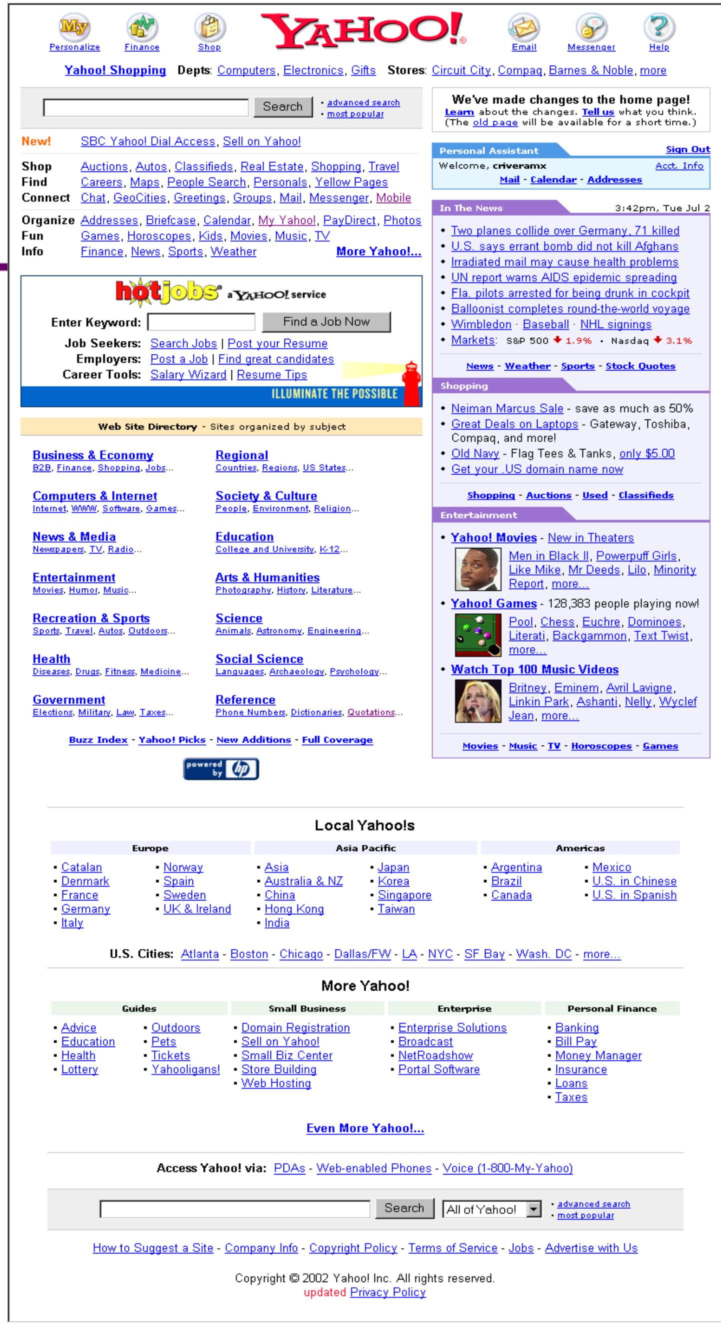
Web Design


Expected
Needs
Design
Use
Information
Architecture
Fidelity
Usability
Visibility
Findability
Ubiquity
Web usage
mining
Demonstrated
Needs
Search Engine
Person

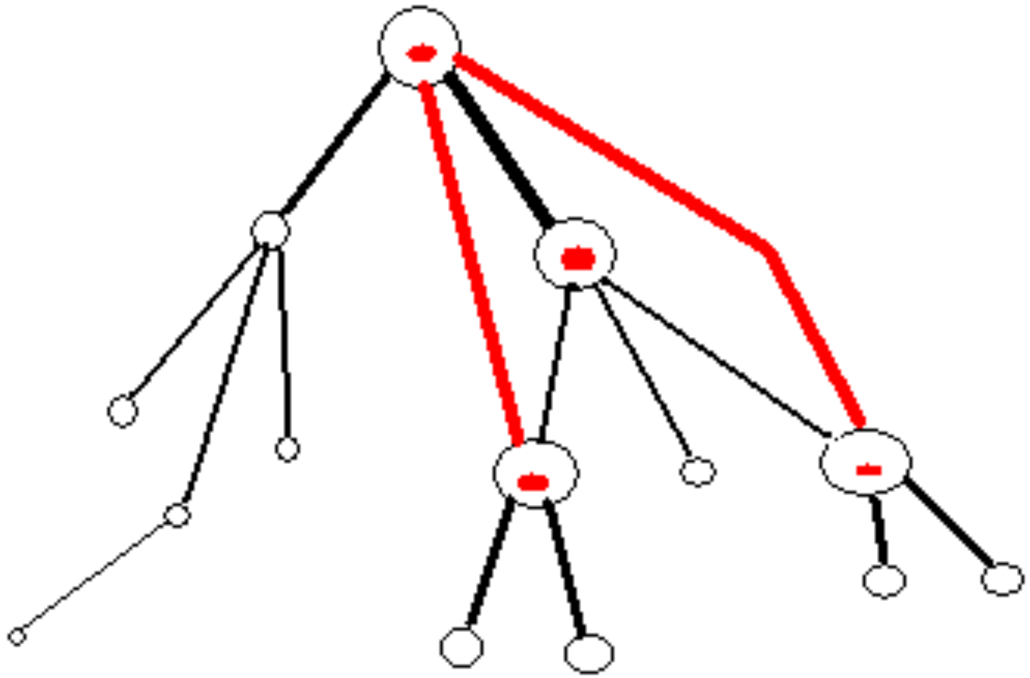
User-driven design

- ***User-driven design***
  - Best example: Yahoo!
- **Navigational log analysis**
  - Site reorganization
- **Query log analysis**
  - Information Scent
  - Content that is missing: market niches

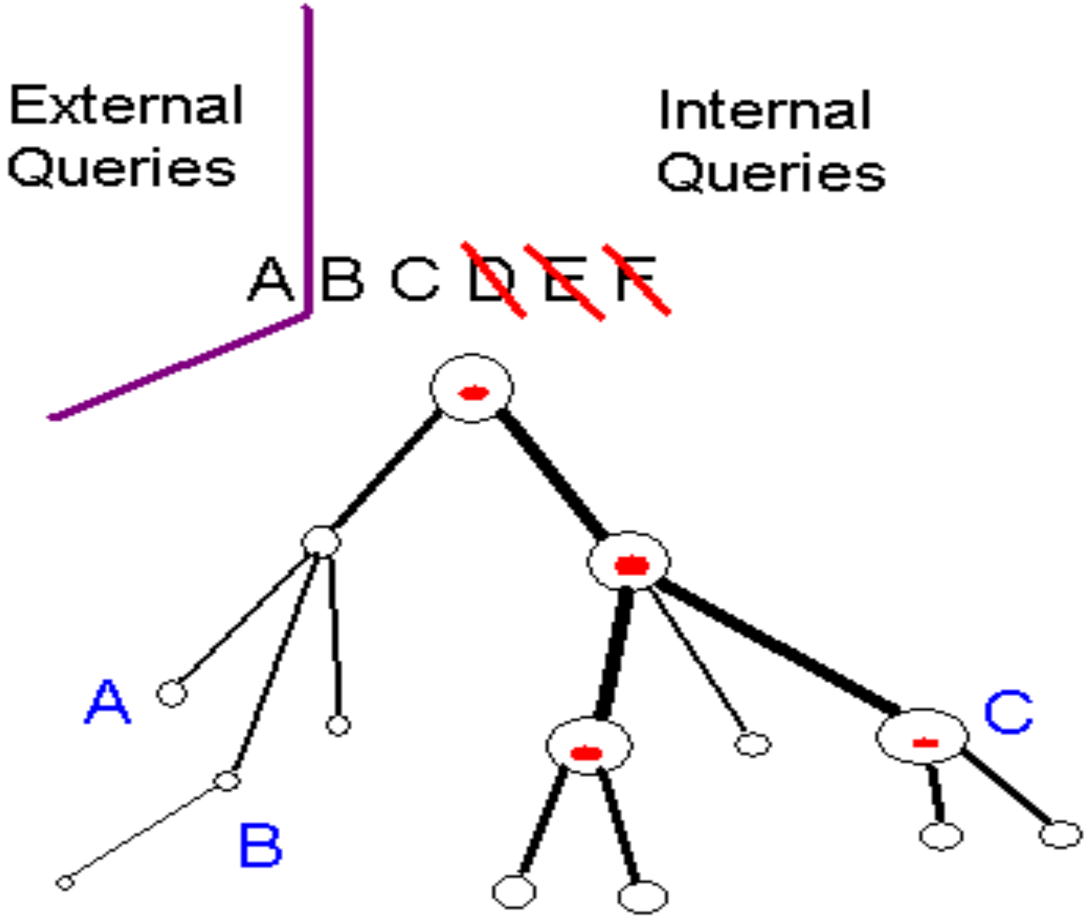

Finance Messenger Check Email
YAHOO!
What's New Personalize Help
Try the beta version of Yahoo!'s new home page
free email@yahoo.com
www.criveramx.com
Build Your Own Web Site
Search advanced search
Yahoo! Shopping Depts: Computers, Electronics, Gifts Stores: Compaq, Circuit City, Barnes & Noble, more
Shop Auctions · Autos · Classifieds · Real Estate · Shopping · Travel · Yellow Pgs · Maps Media Finance · News · Sports · Weather
Connect Careers · Chat · GeoCities · Greetings · Groups/Clubs · Mail · Members · Messenger · Mobile · Personals · People Search · Photos
Personal Addr Book · Briefcase · Calendar · My Yahoo! · PayDirect Fun Games · Horoscopes · Kids · Movies · Music · TV more...
hotjobs a Yahoo! service
Enter Keyword: Find a Job Now
Job Seekers: Search Jobs | Post your Resume
Employers: Post a Job | Find great candidates
Career Tools: Salary Wizard | Resume Tips
ILLUMINATE THE POSSIBLE
Arts & Humanities
Literature, Photography...
News & Media
Full Coverage, Newspapers, TV...
Business & Economy
B2B, Finance, Shopping, Jobs...
Recreation & Sports
Sports, Travel, Autos, Outdoors...
Computers & Internet
Internet, WWW, Software, Games...
Reference
Libraries, Dictionaries, Quotations...
Education
College and University, K-12...
Regional
Countries, Regions, US States...
Entertainment
Picks, Movies, Humor, Music...
Science
Animals, Astronomy, Engineering...
Government
Elections, Military, Law, Taxes...
Social Science
Archaeology, Economics, Languages...
Health
Medicine, Diseases, Drugs, Fitness...
Society & Culture
People, Environment, Religion...
powered by hp
In the News
· Two planes collide over Germany, 71 killed
· U.S. says errant bomb did not kill Afghans
· Irradiated mail may cause health problems
· UN report warns AIDS epidemic spreading
· Fla. pilots arrested for being drunk in cockpit
· Balloonist completes round-the-world voyage
· Wimbledon · Baseball · NHL signings
· Markets: S&P 1.7% · Nasdaq 2.9%
more...
Marketplace
· Loan Center - auto loans, mortgages, credit reports
· Yahoo! Travel - Airfare Specials
Broadcast Events
· Watch Top 100 Music Videos
Nelly, Pink, Eminem, Britney, Avril Lavigne, Linkin Park, Ashanti, more
· Watch World Cup video highlights - Finals Special
Inside Yahoo!
· SBC Yahoo! Dial Unlimited Internet Access - first month free
· Y! Games - pool, literati, spades, chess, dominoes, euchre, backgammon...
· GeoCities - build your own web site
· Make Yahoo! your home page
Local Yahoo!s
Europe : Catalan - Denmark - France - Germany - Italy - Norway - Spain - Sweden - UK & Ireland
Asia Pacific : Asia - Australia & NZ - China - HK - India - Japan - Korea - Singapore - Taiwan
Americas : Argentina - Brazil - Canada - Chinese - Mexico - Spanish
U.S. Cities : Atlanta - Boston - Chicago - Dallas/FW - LA - NYC - SF Bay - Wash. DC - more...
More Yahoo!s
Guides : Buzz Index - Education - Health - Outdoors - Pets - Real Estate - Yahooligans!
Entertainment : Horoscopes - Broadcast - Games - Movies - Music - Radio - Tickets - TV - more
Finance : Banking - Bill Pay - Money Manager - Insurance - Loans - Taxes - more
Local : Autos - Careers - Classifieds - Events - Lodging - Maps - Yellow Pages - more
News : Top Stories - Business - Entertainment - Lottery - Sports - Technology - Weather
Publishing : Advice - Briefcase - Groups - Photos - Home Pages - Message Boards
Small Business : Domain Registration - Small Biz Center - Store Building - Web Hosting
Enterprise : Enterprise Solutions - Broadcast - NetRoadshow - Portal Software
Access Yahoo! via : Pagers, PDAs, Web-enabled Phones and Voice (1-800-My-Yahoo)
Make Yahoo! your home page
How to Suggest a Site - Company Info - Copyright Policy - Terms of Service - Jobs - Advertise with Us
Copyright © 2002 Yahoo! Inc. All rights reserved.
updated Privacy Policy


Personalize Finance Shop
YAHOO!
Email Messenger Help
Yahoo! Shopping Depts: Computers, Electronics, Gifts Stores: Circuit City, Compaq, Barnes & Noble, more
Search · advanced search · most popular
We've made changes to the home page!
Learn about the changes. Tell us what you think.
(The old page will be available for a short time.)
New! SBC Yahoo! Dial Access, Sell on Yahoo!
Shop Auctions, Autos, Classifieds, Real Estate, Shopping, Travel
Find Careers, Maps, People Search, Personals, Yellow Pages
Connect Chat, GeoCities, Greetings, Groups, Mail, Messenger, Mobile
Organize Addresses, Briefcase, Calendar, My Yahoo!, PayDirect, Photos
Fun Games, Horoscopes, Kids, Movies, Music, TV
Info Finance, News, Sports, Weather More Yahoo!...
Personal Assistant Sign Out
Welcome, criveramx Acct. Info
Mail - Calendar - Addresses
In The News 3:42pm, Tue Jul 2
· Two planes collide over Germany, 71 killed
· U.S. says errant bomb did not kill Afghans
· Irradiated mail may cause health problems
· UN report warns AIDS epidemic spreading
· Fla. pilots arrested for being drunk in cockpit
· Balloonist completes round-the-world voyage
· Wimbledon · Baseball · NHL signings
· Markets: S&P 500 1.9% · Nasdaq 3.1%
News - Weather - Sports - Stock Quotes
hotjobs a Yahoo! service
Enter Keyword: Find a Job Now
Job Seekers: Search Jobs | Post your Resume
Employers: Post a Job | Find great candidates
Career Tools: Salary Wizard | Resume Tips
ILLUMINATE THE POSSIBLE
Web Site Directory - Sites organized by subject
Business & Economy
B2B, Finance, Shopping, Jobs...
Regional
Countries, Regions, US States...
Computers & Internet
Internet, WWW, Software, Games...
Society & Culture
People, Environment, Religion...
News & Media
Newspapers, TV, Radio...
Education
College and University, K-12...
Entertainment
Movies, Humor, Music...
Arts & Humanities
Photography, History, Literature...
Recreation & Sports
Sports, Travel, Autos, Outdoors...
Science
Animals, Astronomy, Engineering...
Health
Diseases, Drugs, Fitness, Medicine...
Social Science
Languages, Archaeology, Psychology...
Government
Elections, Military, Law, Taxes...
Reference
Phone Numbers, Dictionaries, Quotations...
Buzz Index - Yahoo! Picks - New Additions - Full Coverage
powered by hp
Shopping
· Neiman Marcus Sale - save as much as 50%
· Great Deals on Laptops - Gateway, Toshiba, Compaq, and more!
· Old Navy - Flag Tees & Tanks, only $5.00
· Get your .US domain name now
Shopping - Auctions - Used - Classifieds
Entertainment
· Yahoo! Movies - New in Theaters
Men in Black II, Powerpuff Girls, Like Mike, Mr Deeds, Lilo, Minority Report, more...
· Yahoo! Games - 128,383 people playing now!
Pool, Chess, Euchre, Dominoes, Literati, Backgammon, Text Twist, more...
· Watch Top 100 Music Videos
Britney, Eminem, Avril Lavigne, Linkin Park, Ashanti, Nelly, Wyclef Jean, more...
Movies - Music - TV - Horoscopes - Games
Local Yahoo!s
Europe
· Catalan · Denmark · France · Germany · Italy · Norway · Spain · Sweden · UK & Ireland
Asia Pacific
· Asia · Australia & NZ · China · Hong Kong · India · Japan · Korea · Singapore · Taiwan
Americas
· Argentina · Brazil · Canada · Mexico · U.S. in Chinese · U.S. in Spanish
U.S. Cities: Atlanta - Boston - Chicago - Dallas/FW - LA - NYC - SF Bay - Wash. DC - more...
More Yahoo!
Guides
· Advice · Education · Health · Lottery · Outdoors · Pets · Tickets · Yahooligans!
Small Business
· Domain Registration · Sell on Yahoo! · Small Biz Center · Store Building · Web Hosting
Enterprise
· Enterprise Solutions · Broadcast · NetRoadshow · Portal Software
Personal Finance
· Banking · Bill Pay · Money Manager · Insurance · Loans · Taxes
Even More Yahoo!...
Access Yahoo! via: PDAs - Web-enabled Phones - Voice (1-800-My-Yahoo)
Search All of Yahoo! · advanced search · most popular
How to Suggest a Site - Company Info - Copyright Policy - Terms of Service - Jobs - Advertise with Us
Copyright © 2002 Yahoo! Inc. All rights reserved.
updated Privacy Policy

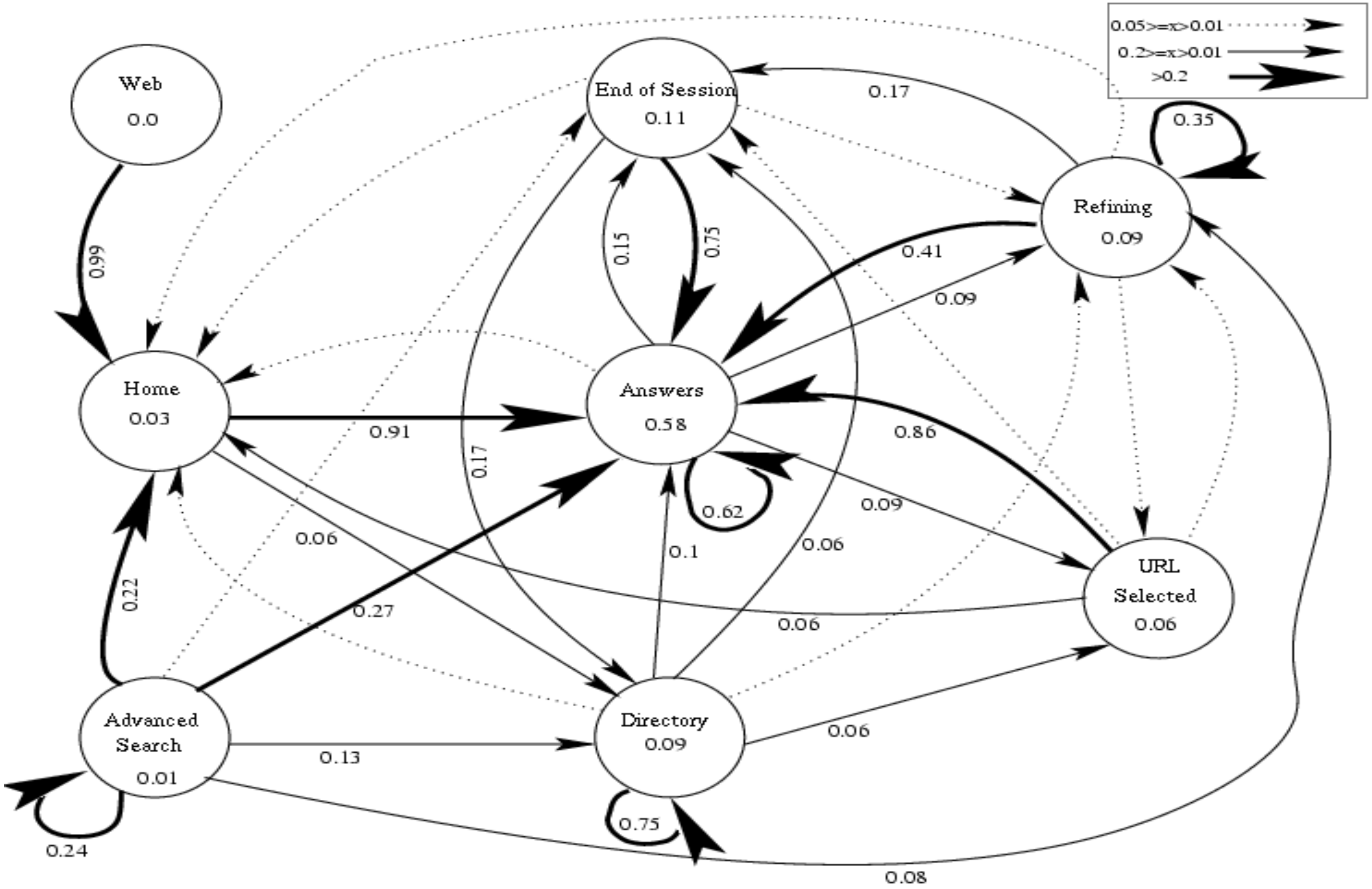
Navigation Mining



Web Site Query Mining



User Modeling



Applications

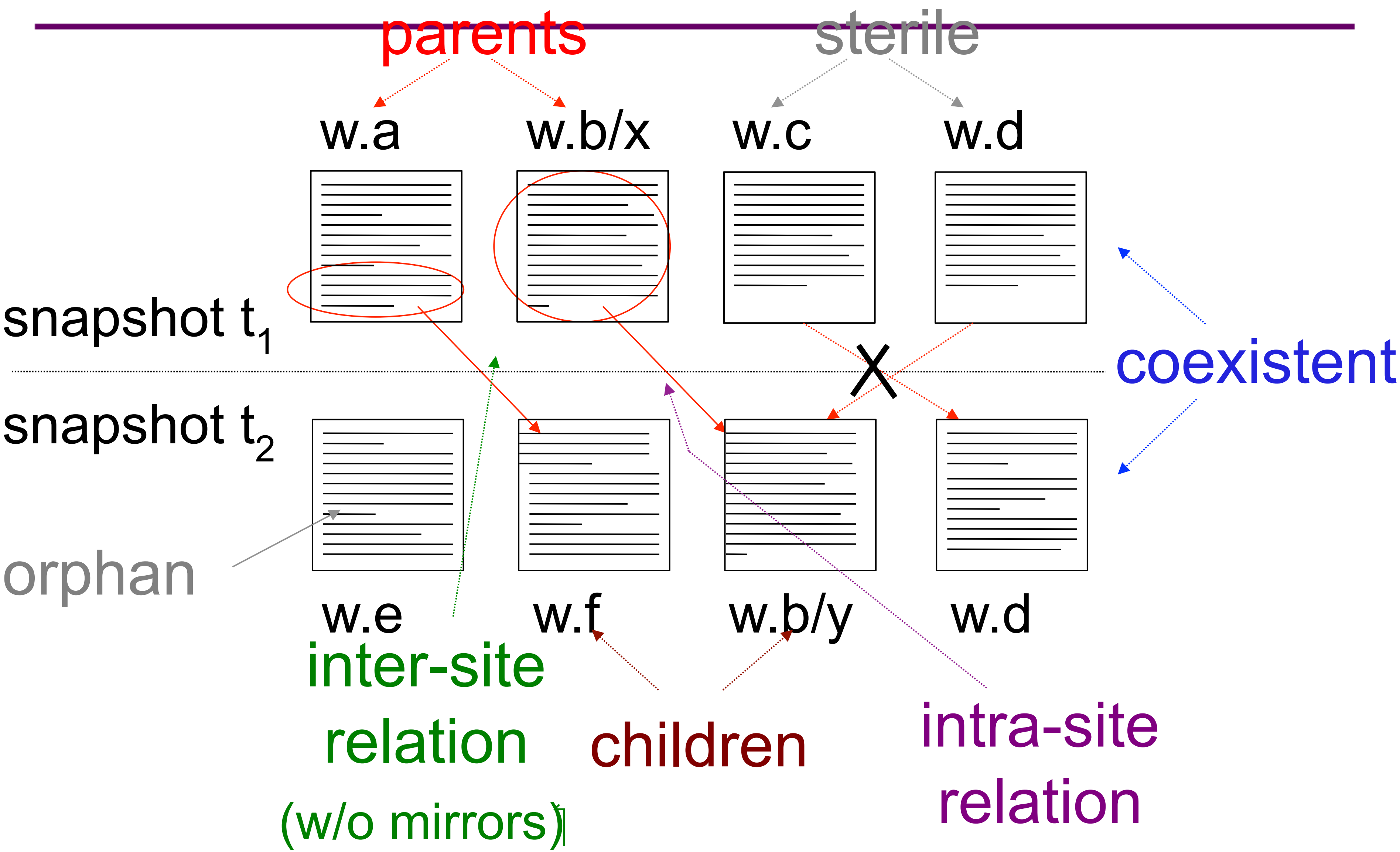
Content mining

- Web genealogy
- Content-based Web spam detection
- Finding high-quality content in social media

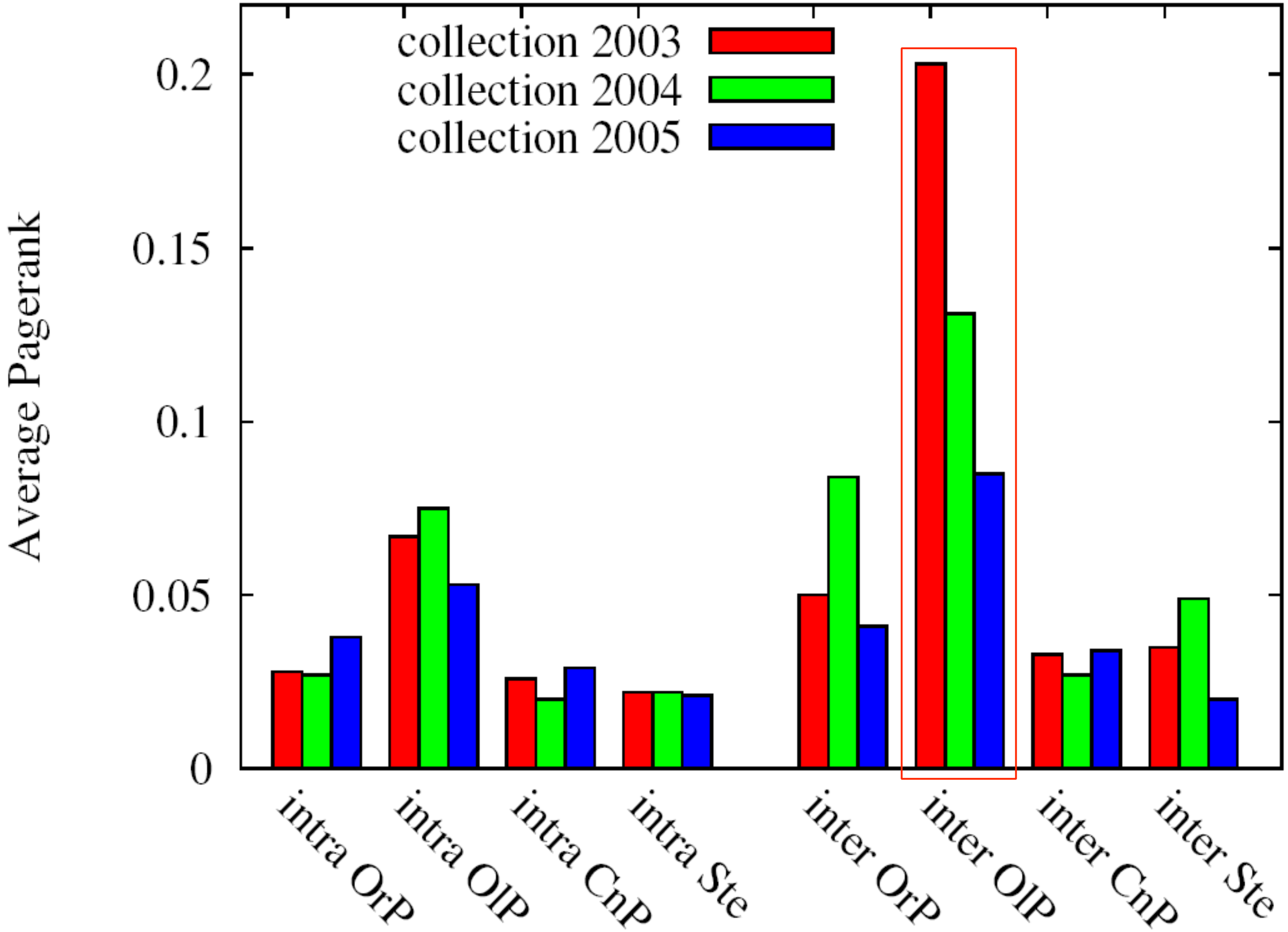
Study genealogy of the Web

- **[Baeza-Yates et al., 2008]**
- **New pages copy content from existing pages**
- **Web genealogy study:**
  - How textual content of source pages (parents) are reused to compose part of new Web pages (children)
  - Not near-duplicates, as similarities of short passages are also identified
- **How can search engines benefit?**
  - By associating more relevance to a parent page?
  - By trying to decrease the bias?

Web Genealogy



Pagerank for each component



The wisdom of spammers

- Many world-class athletes, from all sports, have the ability to get in the right state of mind and when looking for **women looking for love** the state of mind is most important. [..] You should have the same attitude in looking for **women looking for love** and we make it easy for you.

- Many world-class athletes, from all sports, have the ability to get in the right state of mind and when looking for **texas boxer dog breeders** the state of mind is most important. [..] You should be thinking the same when you are looking for **texas boxer dog breeders** and we make it easy for you.

Bookmark Page Home Home

SOFT SEARCH

lava soft | php script | top soft | java script | MP3

Top Searches:

- Acne
- Weight Loss Pills
- Debt Consolidation
- Loan
- Domain Names
- Advertising
- Online Pharmacy
- Home Loan
- Dedicated Server
- Car Rental
- Adipex
- Levitra
- Online Poker
- Work At Home
- Propecia
- Consolidate Debt
- Mortgage Rates
- Online Craps
- Vegas Casinos
- Buy Ionamin

Top Web Results

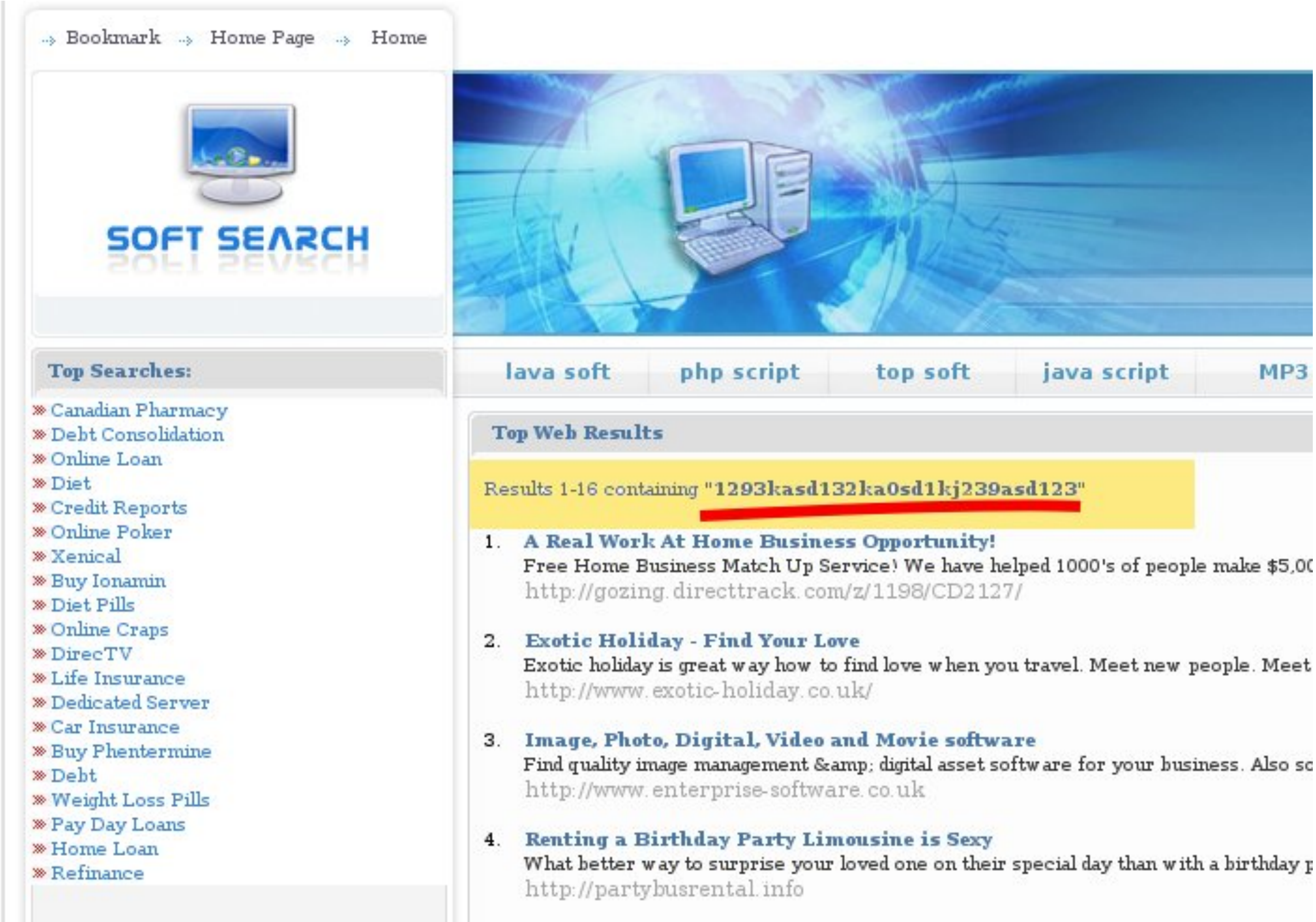
Results 1-16 containing "**sports book**"

1. **Place Your Bet with #1 Sports Betting Site Online**
   Kentucky Derby, NBA, MLB, NHL and all other sports betting and odds. Place a full ran
   sportsbook in North America
   http://www.sportsinteraction.com

2. **AnteUp GamblingLinks.com - Safe Online Casinos**
   Links to safe and secure online casino gambling and sports betting including reviews, ne
   http://gamblinglinks.com

3. **Free Casino Bonuses. Links To the Best Casinos**
   Get $20 - $500 in Free Chips. Most popular casino games with great graphics. Play for f
   rules and strategy. Links to the Best Casinos
   http://www.fastfreecash.net

4. **AnteUp GamblingLinks.com - Safe Online Casinos**


Bookmark Home Page Home
SOFT SEARCH
Top Searches:
Canadian Pharmacy
Debt Consolidation
Online Loan
Diet
Credit Reports
Online Poker
Xenical
Buy Ionamin
Diet Pills
Online Craps
DirecTV
Life Insurance
Dedicated Server
Car Insurance
Buy Phentermine
Debt
Weight Loss Pills
Pay Day Loans
Home Loan
Refinance
lava soft
php script
top soft
java script
MP3
Top Web Results
Results 1-16 containing "1293kasd132ka0sd1kj239asd123"
1. A Real Work At Home Business Opportunity!
Free Home Business Match Up Service! We have helped 1000's of people make $5,0
http://gozing.directtrack.com/z/1198/CD2127/
2. Exotic Holiday - Find Your Love
Exotic holiday is great way how to find love when you travel. Meet new people. Meet
http://www.exotic-holiday.co.uk/
3. Image, Photo, Digital, Video and Movie software
Find quality image management & digital asset software for your business. Also sc
http://www.enterprise-software.co.uk
4. Renting a Birthday Party Limousine is Sexy
What better way to surprise your loved one on their special day than with a birthday p
http://partybusrental.info

Sample query-targeted outlinks

- spam blocker
  free spam blocker
  outlook express spam blocker
  outlook spam blocker
  email spam blocker
  yahoo spam blocker
  free spam blocker outlook express
  spam blocker utility
  anti spam blocker
  microsoft spam blocker
  pop up spam blocker
  download free spam blocker
  free yahoo spam blocker
  bay area spam blocker
  blocking exchange server

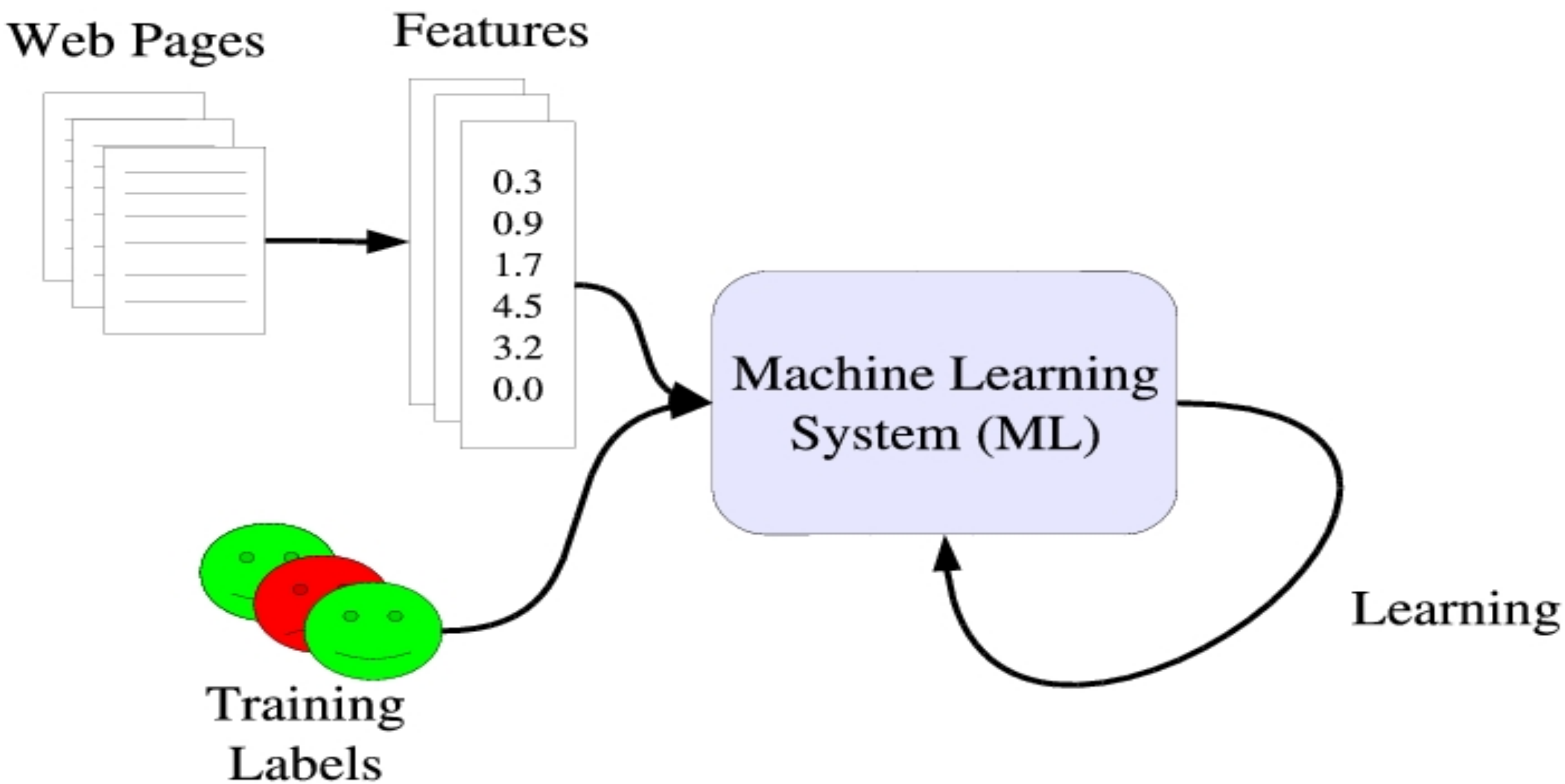
spam
spam e mail
mcafee anti spam
best anti spam
catch configuring email filter spam
blocker spam
send spam email
free junk spam filter outlook
adaptive filtering spam
anit software spam xp
blocker free spam
best spam block
free spam blocker and filter

The Power of Social Networks

- Spammers many times are (or look like) social networks
  - But the Web has larger social networks
- Examples
  - Any statistical deviation is suspicious
  - Any bounded amount of work is suspicious
    - Truncated PageRank
      - Spammers link support have shorter incoming paths

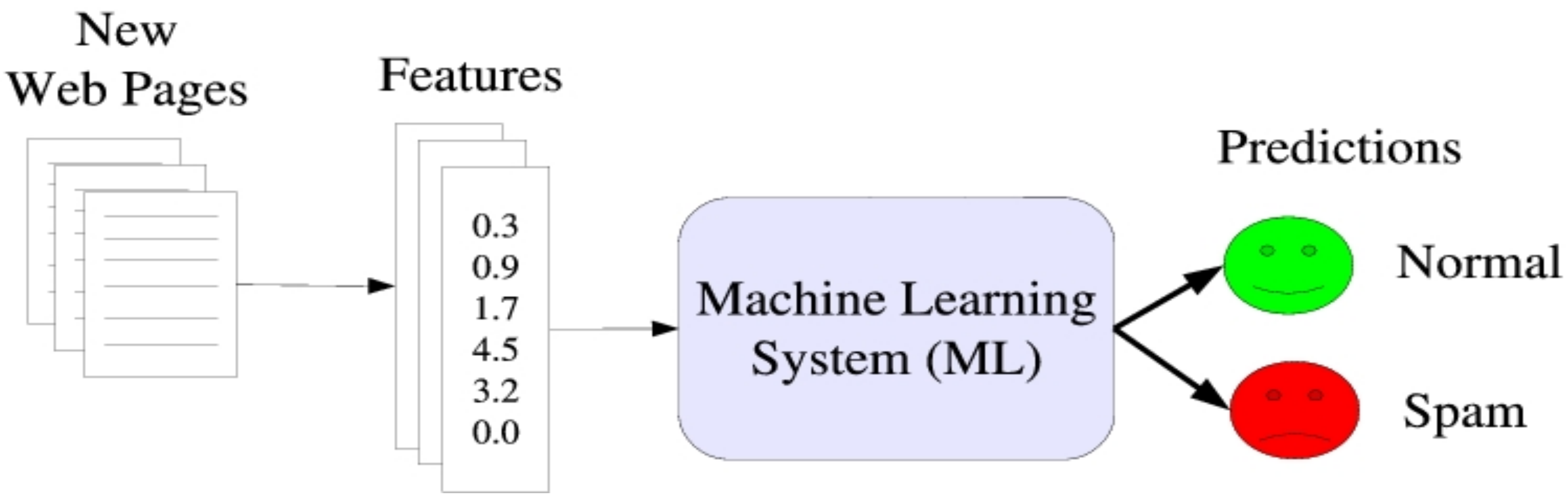
Content-based spam detection

- **Machine-learning approach --- training**



Content-based spam detection

- **Machine-learning approach --- prediction**



The dataset

- **Label "spam" nodes on the host level**
  - agrees with existing granularity of Web spam
- **Based on a crawl of .uk domain from May 2006**
- **77.9 million pages**
- **3 billion links**
- **11,400 hosts**

The dataset

- 20+ volunteers tagged a subset of host
- Labels are “spam”, “normal”, “borderline”
- Hosts such as .gov.uk are considered “normal”
- In total 2,725 hosts were labelled by at least two judges
- hosts in which both judges agreed, and “borderline” removed
- Dataset available at

http://www.yr-bcn.es/webspam/

Content-based features

- Number of words in the page
- Number of words in the title
- Average word length
- Fraction of anchor text
- Fraction of visible text

See also [Ntoulas et al., 06]

Content-based features Entropy related

- Let $T = \{ (w_1, p_1), \ldots, (w_k, p_k) \}$ the set of trigrams in a page, where trigram w_i has frequency p_i
- Features:
- ✓ Entropy of trigrams: $H = - \sum_i p_i \log(p_i)$
- ✓ Independent trigram likelihood: $- (1/k) \sum_i \log(p_i)$
- ✓ Also, compression rate, as measured by bzip

Content-based features related to popular keywords

- *F* set of most frequent terms in the collection
- *Q* set of most frequent terms in a query log
- *P* set of terms in a page
- Features:

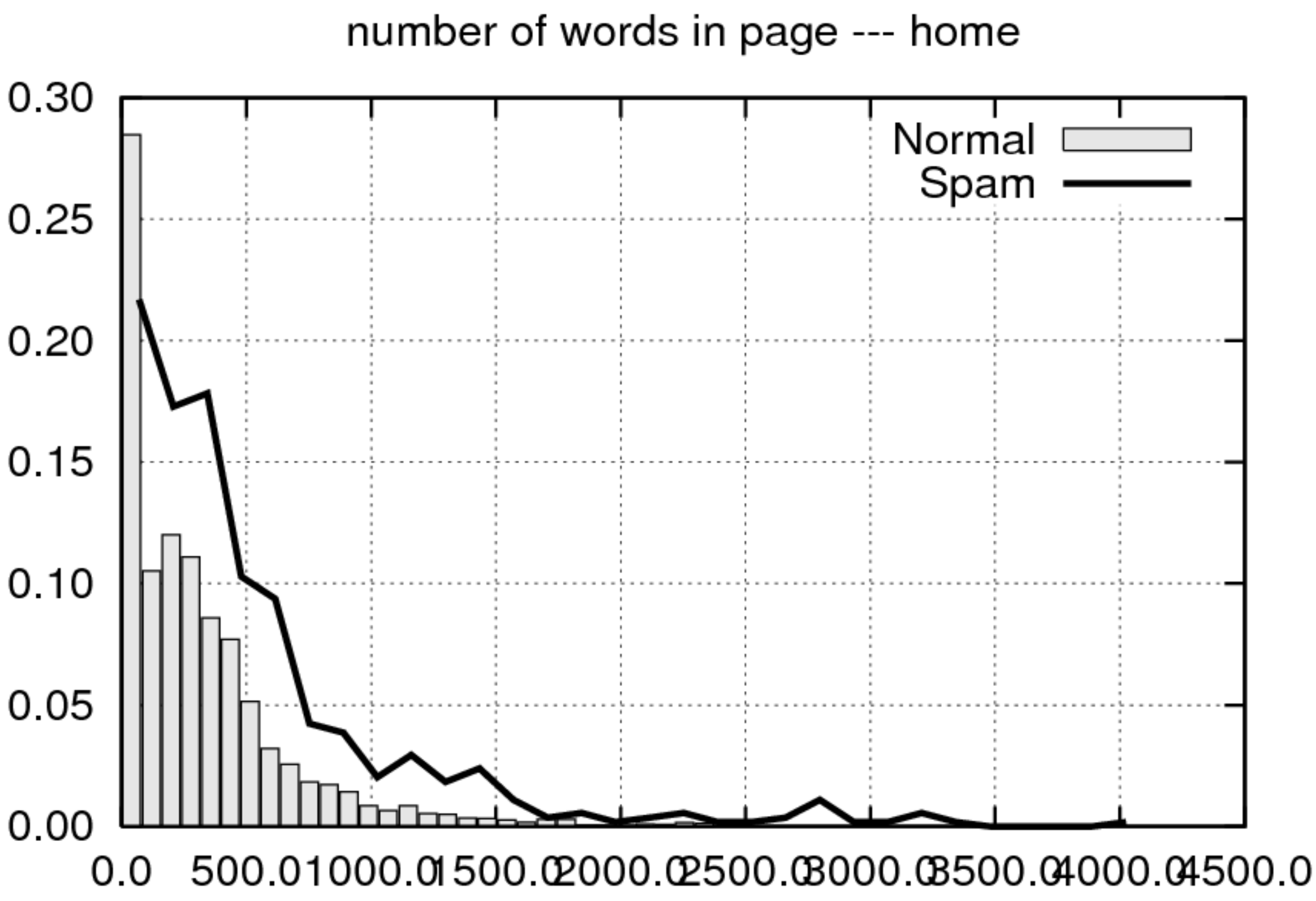
✓ Corpus "precision" $|P \cap F| / |P|$

✓ Corpus "recall" $|P \cap F| / |F|$

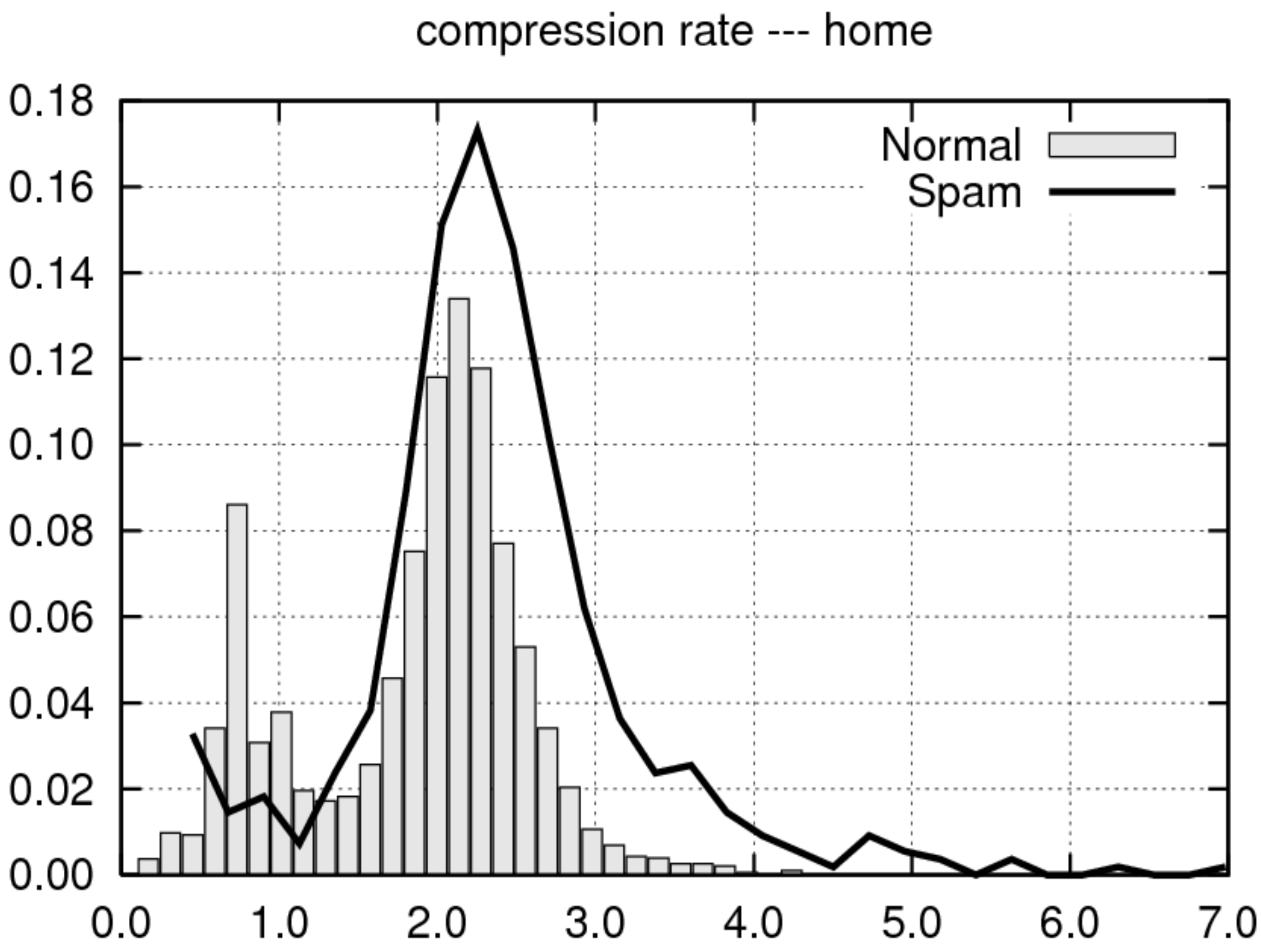
✓ Query "precision" $|P \cap Q| / |P|$

✓ Query "recall" $|P \cap Q| / |Q|$

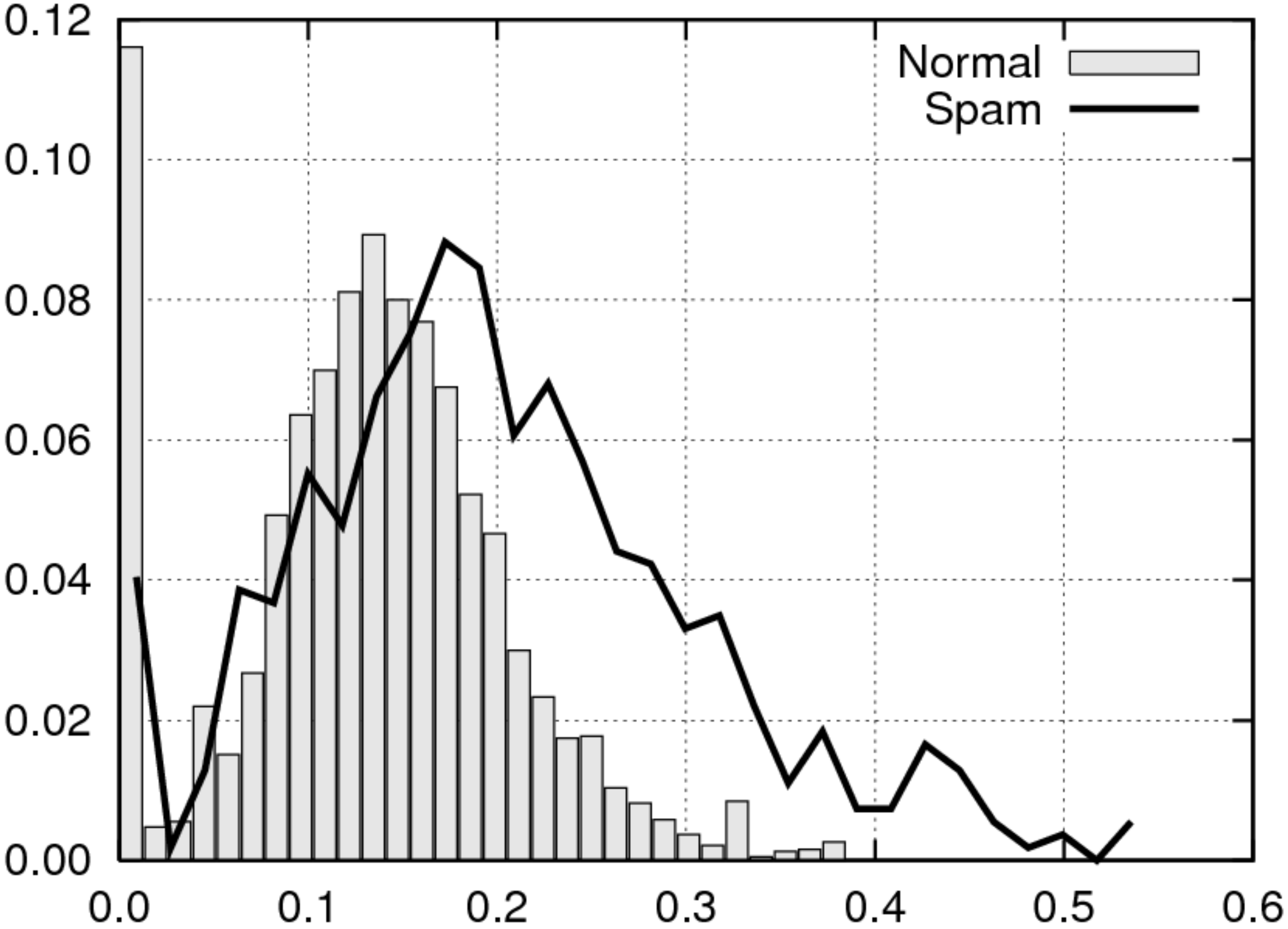
Content-based features number of words in home page



Content-based features compression rate



Content-based features
Query precision



The classifier

- C4.5 decision tree with bagging and cost weighting for class imbalance

- With content-based features achieves:
  - True positive rate: 64.9%
  - False positive rate: 3.7%
  - F-Measure: 0.683

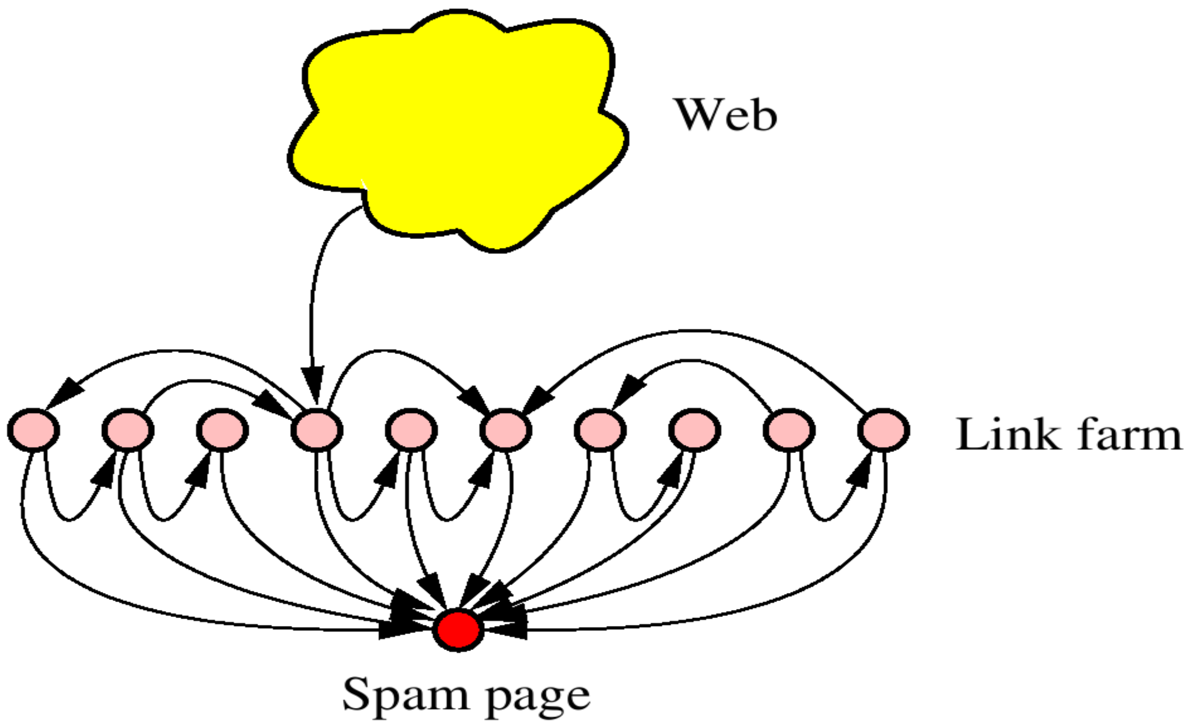
Structure and link analysis

- Link-based spam detection
- Finding high-quality content in social media

Link-based spam detection

- Link farms used by spammers to raise popularity of spam pages
- Link farms and other spam strategies leave traces on the structure of the web graph
- Dependencies between neighbouring nodes of the web graph are created
- Naturally, spammers try to remove traces and dependencies

Link farms



- **Single-level link farms can be detected by searching for nodes sharing their out-links**
- **In practice more sophisticated techniques are used**

Link-based features
Degree related

- **in-degree**
- **out-degree**
- **edge reciprocity**
  – number of reciprocal links
- **assortativity**
  – degree over average degree of neighbors

Link-based features PageRank related

- **PageRank**
- **indegree/PageRank**
- **outdegree/PageRank**
- **...**
- **Truncated PageRank [Becchetti et al., 2006]**
  - A variant of PageRank that diminishes the influence of a page the PageRank score of its neighbors
- **TrustRank [Gyongyi et al., 2004]**
  - As PageRank but with teleportation at Open Directory pages

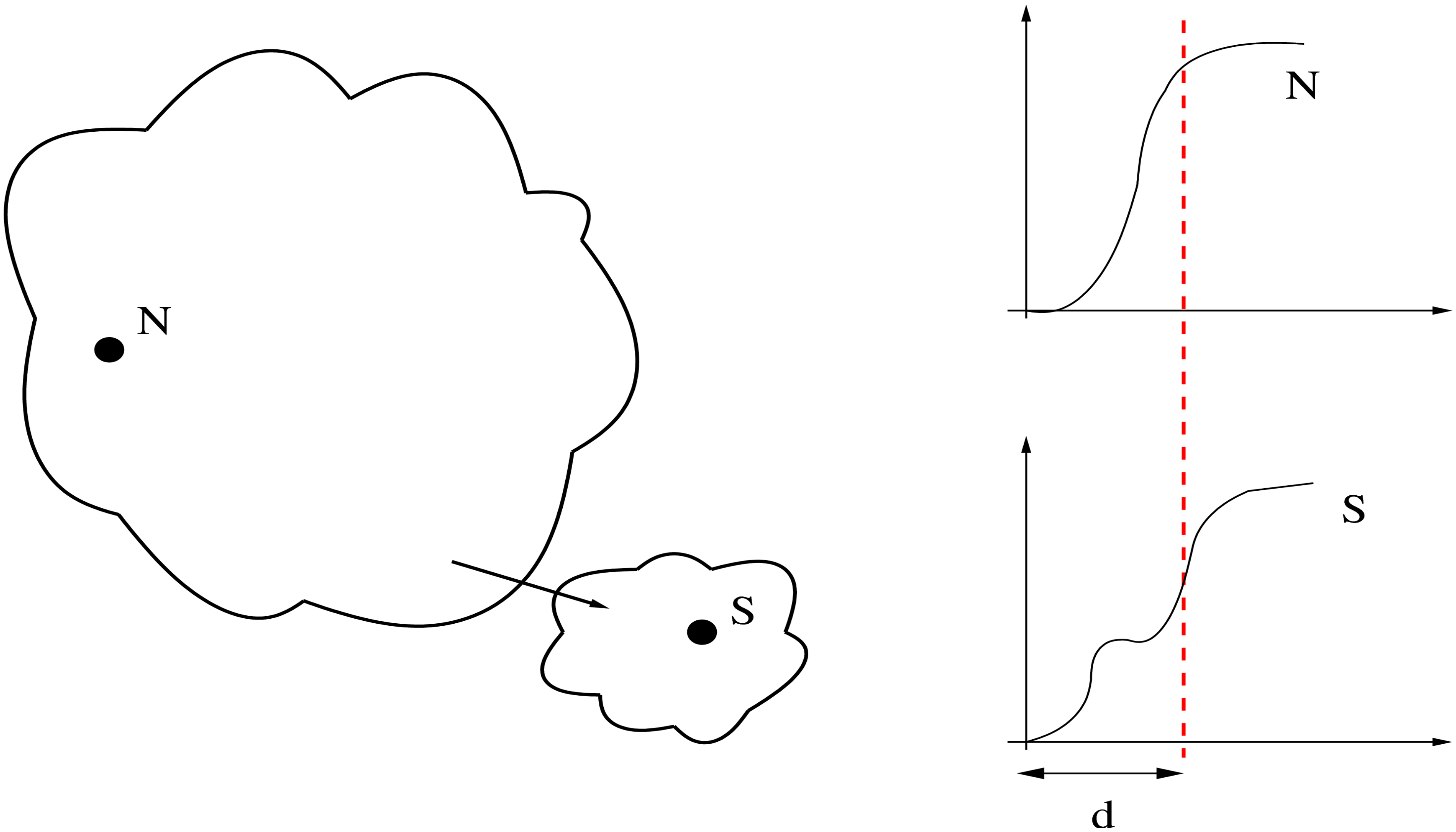
Link-based features
Supporters

- **Let *x* and *y* be two nodes in the graph**
- **Say that *y* is a *d*-supporter of *x*, if the shortest path from *y* to *x* has length at most *d***
- **Let $N_d(x)$ be the set of the *d*-supporters of *x***
- **Define bottleneck number of *x*, up to distance *d* as**

$$b_d(x) = \min_{j <= d} \; N_j(x)/N_{j-1}(x)$$

- **minimum rate of growth of the neighbors of *x* up to a certain distance**

Link-based features
Supporters



Link-based features
Supporters

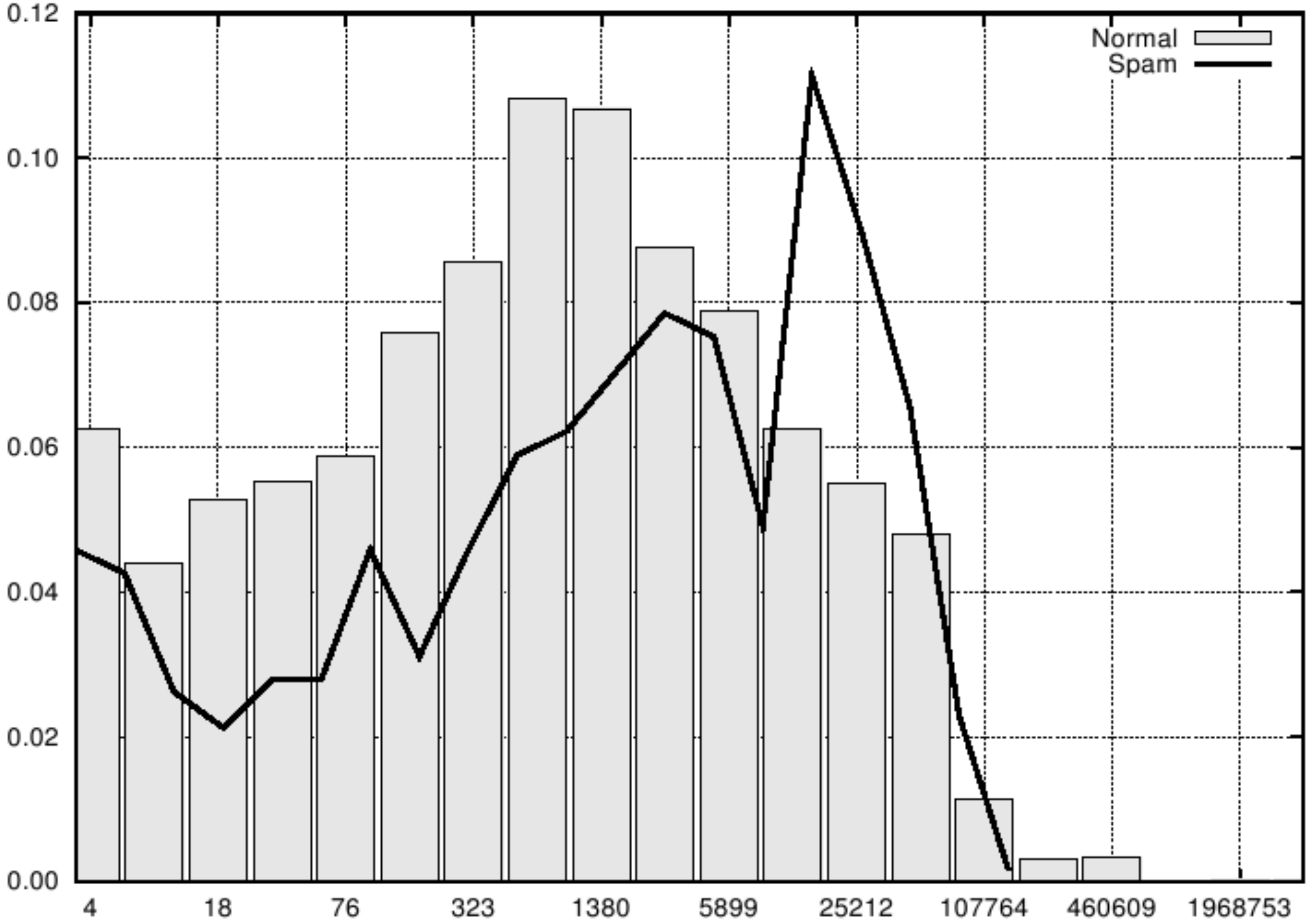
- **How to compute the supporters?**
- **Utilize *neighborhood function***

$$N(h) = |\{ (u,v) \mid d(u,v) <= h \}| = \sum_u N(u,h)$$

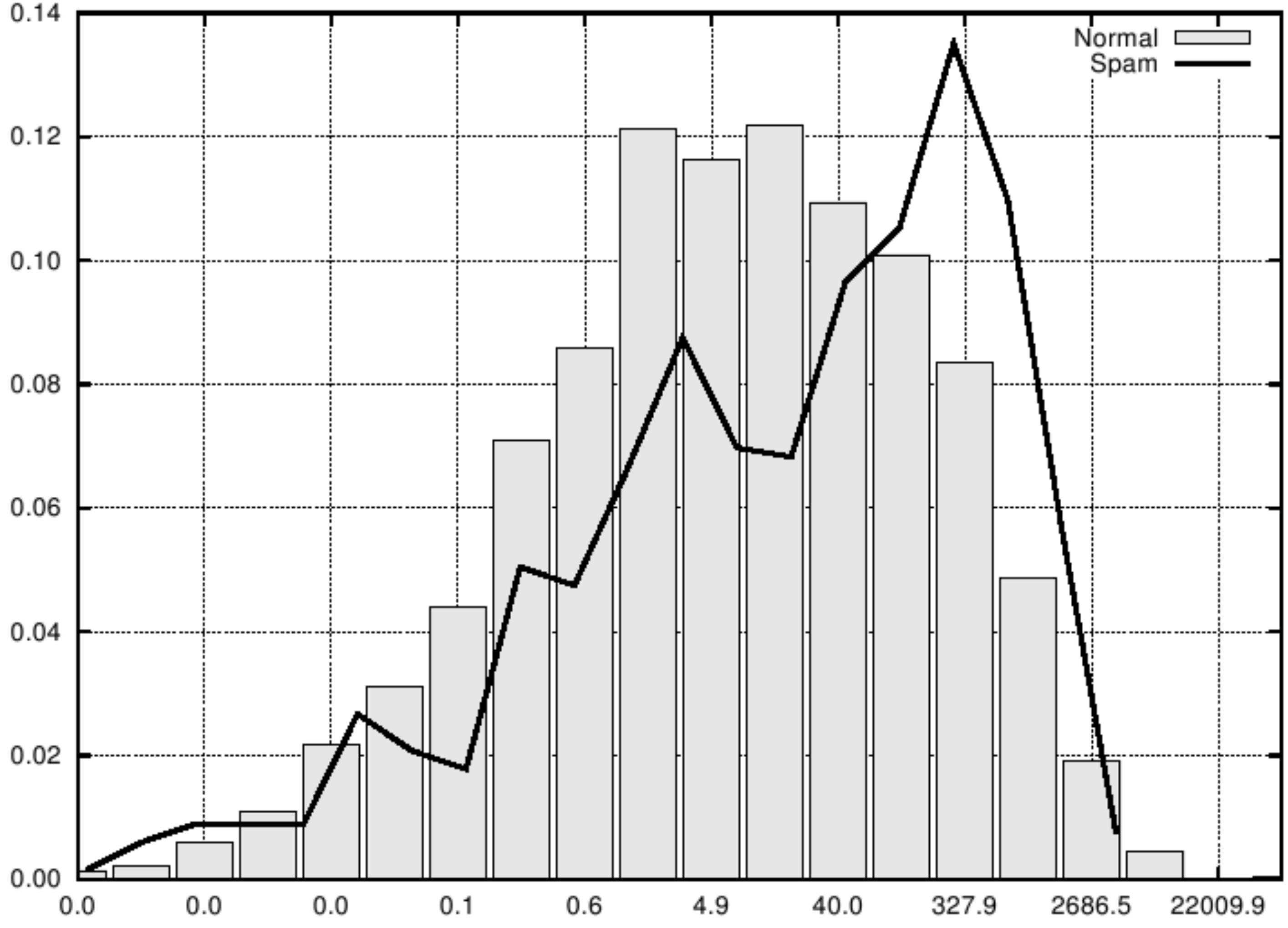
Palmer

- **and ANF algorithm [Palmer et al., 2002]**
- **Probabilistic counting using Flajolet-Martin sketches or other data-stream technology**
- **Can be done with a few passes and exchange of sketches, instead of executing BFS from each node**

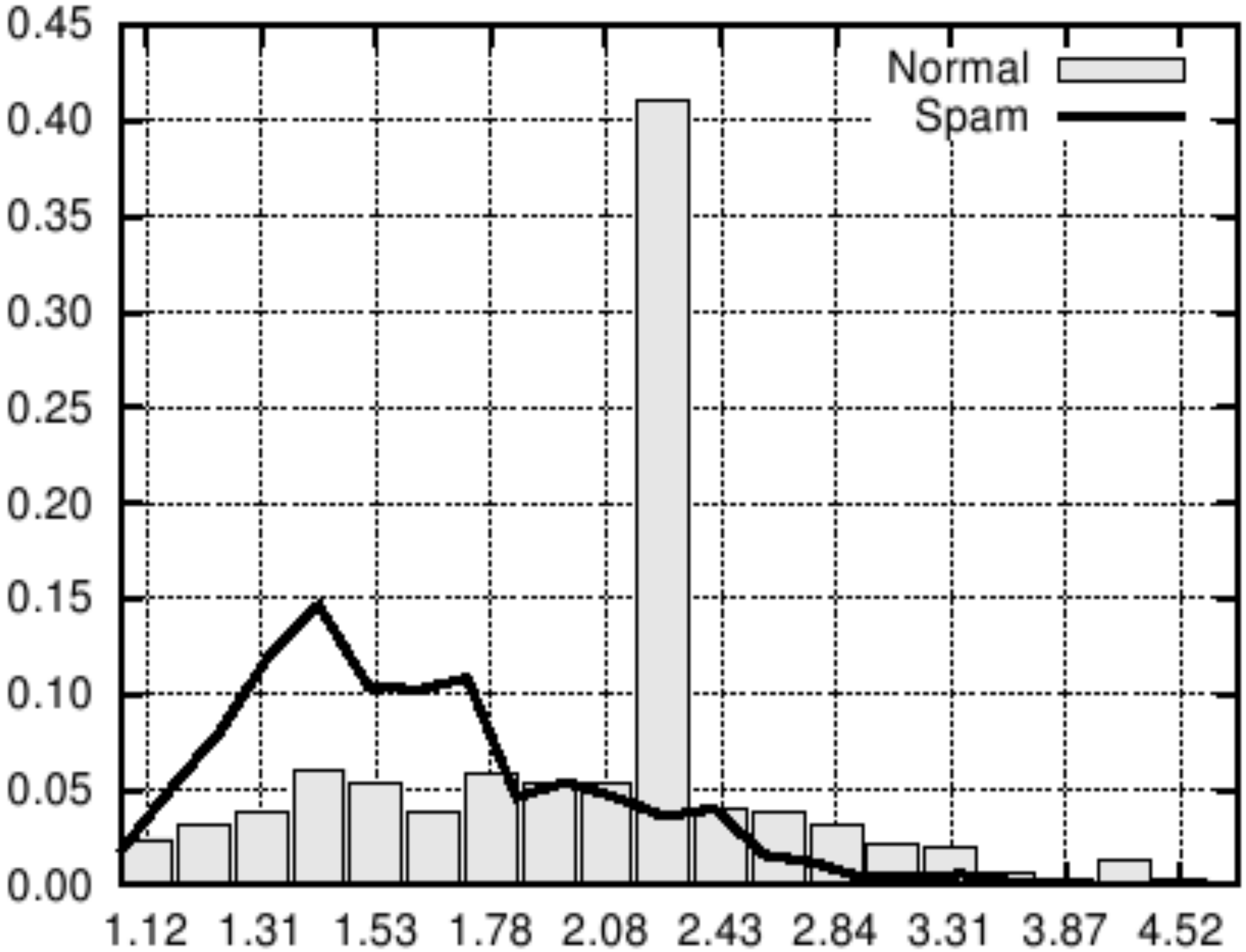
Link-based features - In-degree



Link-based features - Assortativity



Link-based features - Supporters

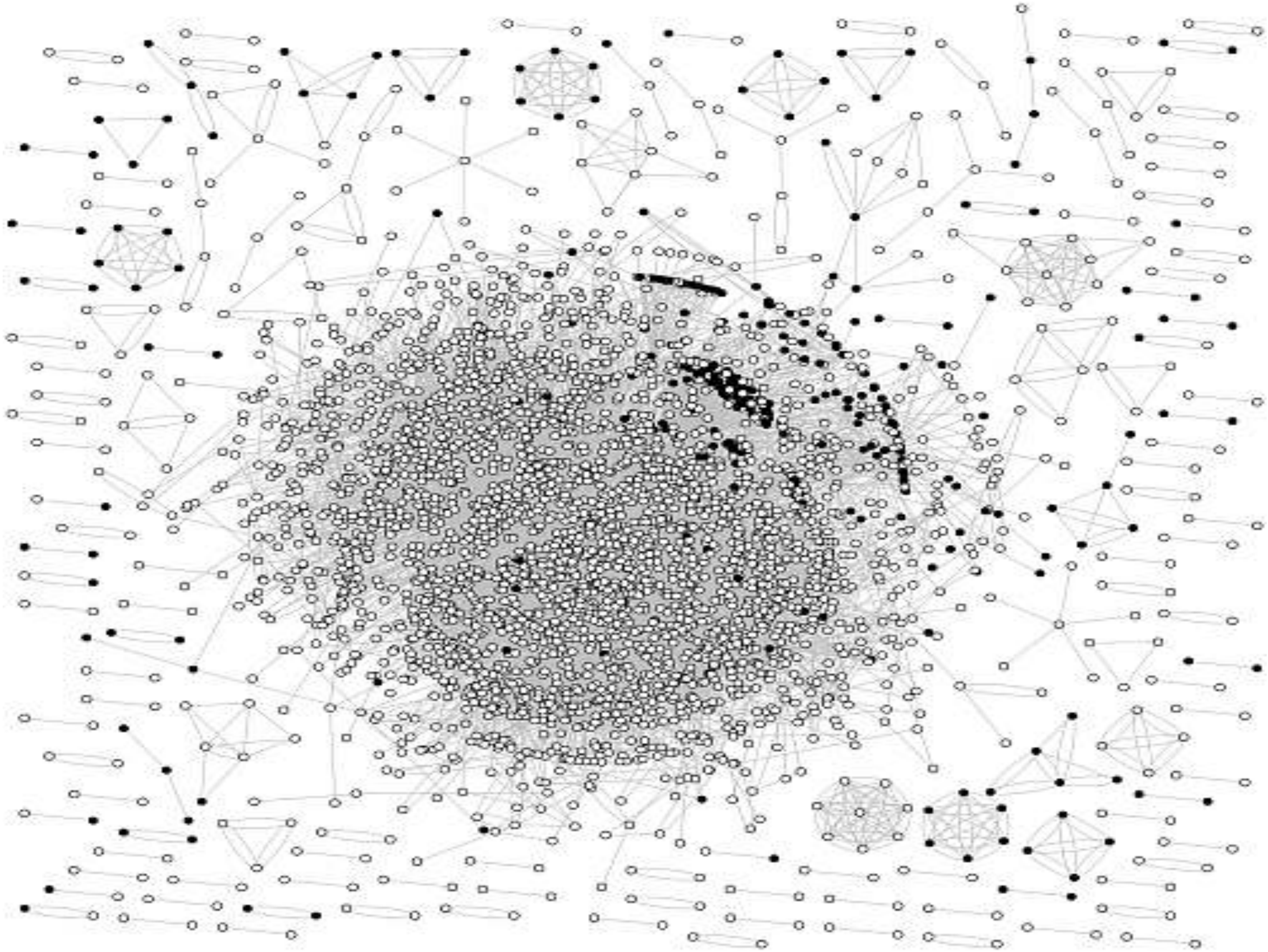


The classifier
Combining features

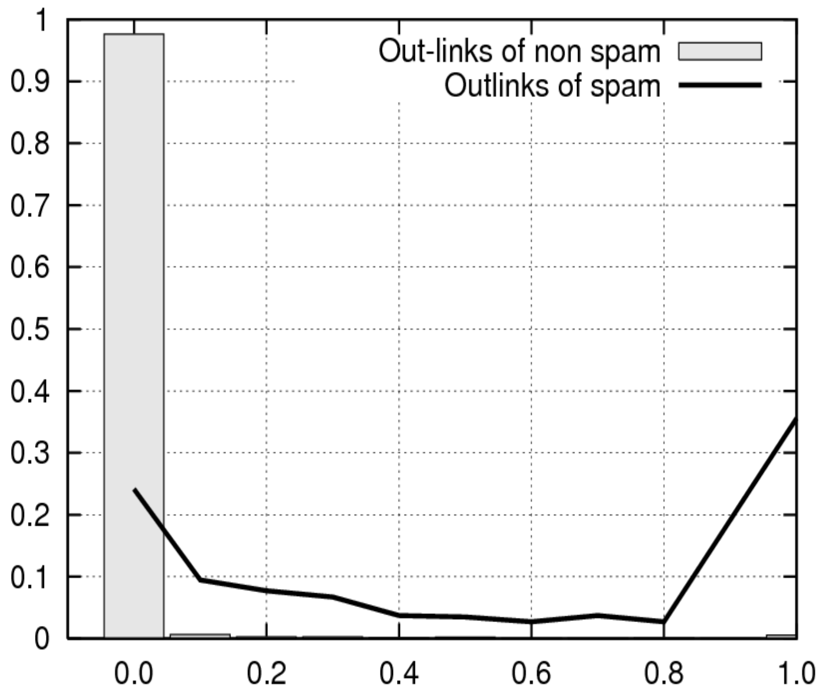
- C4.5 decision tree with bagging and cost weighting for class imbalance

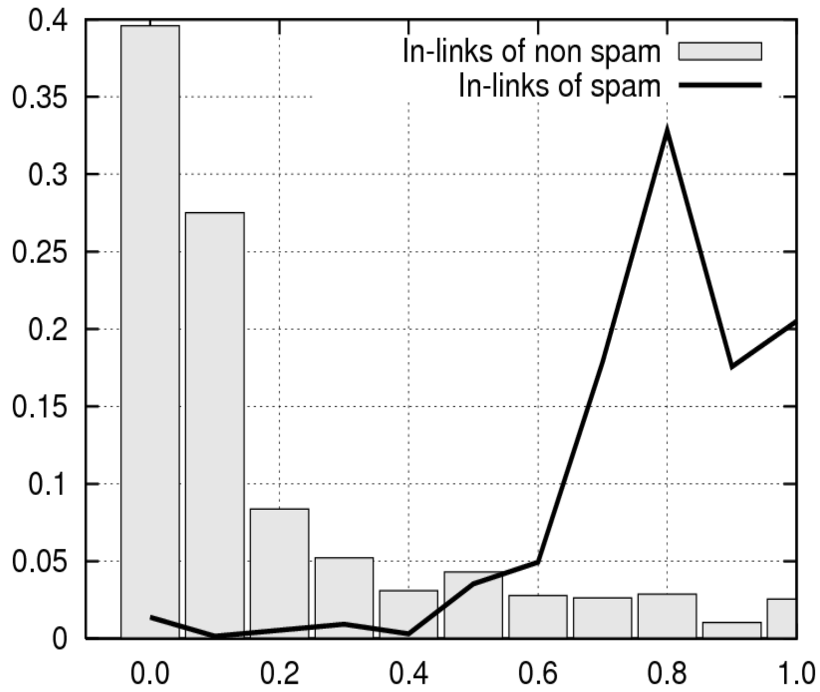
features:	**Content**	**Link**	**Both**
True positive rate:	64.9%	79.4%	78.7%
False positive rate:	3.7%	9.0%	5.7%
F-Measure:	0.683	0.659	**0.723**

Dependencies among spam nodes



Dependencies among spam nodes





- **Spam nodes in out-links**
- **Spam nodes from in-links**

Exploiting dependencies

- **Use a dataset with labeled nodes**
- **Extract content-based and link-based features**
- **Learn a classifier for predicting spam nodes independently**
- **Exploit the graph topology to improve classification**
  - Clustering
  - Propagation
  - Stacked learning

Exploiting dependencies Clustering

- **Let *$G=(V,E,w)$* be the host graph**
- **Cluster *G* into *m* disjoint clusters *$C_1,\ldots,C_m$***
- **Compute *$p(C_i)$*, the fraction of nodes classified as spam in cluster *C_i***
  - if $p(C_i) > t_u$ label all as spam
  - if $p(C_i) < t_l$ label all as non-spam
- **A small improvement:**

	Baseline	**Clustering**
True positive rate:	78.7%	76.9%
False positive rate:	5.7%	5.0%
F-Measure:	0.723	**0.728**

Exploiting dependencies Propagation

- **Perform a random walk on thegraph**
- **With probability α follow a link**
- **With prob *1-*α jump to a random node labeled spam**
- **Relabel as spam every node whose stationary distribution component is higher than a threshold**

- **Improvement:**

	Baseline	**Propagation (backwds)**
True positive rate:	78.7%	75.0%
False positive rate:	5.7%	4.3%
F-Measure:	0.723	**0.733**

Exploiting dependencies
Stacked learning

- Meta-learning scheme [Cohen and Kou, 2006]
- Derive initial predictions
- Generate an additional attribute for each object by combining predictions on neighbors in the graph
- Append additional attribute in the data and retrain

- Let *p(h)* be the prediction of a classification algorithm for *h*
- Let *N(h)* be the set of pages related to *h*
- Compute:

$$f(h) = \sum_{g \in N(h)} p(g) \,/\, |N(h)|$$

- Add *f(h)* as an extra feature for instance *h* and retrain

Exploiting dependencies
Stacked learning

- **First pass:**

	Baseline	**in**	**out**	**both**
True positive rate:	78.7%	84.4%	78.3%	85.2%
False positive rate:	5.7%	6.7%	4.8%	6.1%
F-Measure:	0.723	0.733	0.742	**0.750**

- **Second pass:**

	Baseline	**1st pass**	**2nd pass**
True positive rate:	78.7%	85.2%	88.2%
False positive rate:	5.7%	6.1%	6.3%
F-Measure:	0.723	0.750	**0.763**

Current goals for Web spam effort

- Prevent spam from distorting ranking, but preserve:
  - Relevance
    - “Perfect spam” is a sensible category
  - Freshness
    - Can’t slow down discovery just because spammers exploit it
  - Comprehensiveness
    - Navigational queries for spam should succeed
- Focus on two kinds of spam only:
  - 1) Spam that is succeeding in ranking inappropriately highly
  - 2) Spam that consumes huge amounts of system resources

    (Everything else is “dark matter”)

The power of social media

- Flickr – community phenomenon
- Millions of users share and tag each others' photographs (why???)
- The *wisdom of the crowds* can be used to search
  - – Ranking features to Yahoo! Answers
- The principle is not new – anchor text used in "standard" search
- What about generating pseudo-semantic resources?

Yahoo! Answers

Yahoo! My Yahoo! Mail Make Y! your home page Search: Web Search

YAHOO! ANSWERS

Welcome, **chato**
[Sign Out, My Account]

Answers Home - Forum - Blog - Help

ask. answer. discover.

Search for questions: Search Advanced My Profile

Home > Consumer Electronics > Land Phones > Resolved Question

ndyou

Resolved Question

Show me another »

What's the best way to get telemarketers off my back?

i have caller id and usually don't answer. how can i get them to stop calling (i hear the donotcall registry doesn't work) and if i do pick up the phone aside from immediately hanging up what can i say to deter additional calls?

1 year ago

Report It

hrh_grac...

Best Answer - Chosen by Asker

Register at the online do not call registry. Cell phones, business and home phones can be registered... You will still get some calls for about 30 days. Just tell anyone who calls in that time period that you are registered with the do not call registry and to please remove you from their calling list. If they give you any hassle advise them that you will file a report.

I had to do this too and every solicitor I spoke to was immediately ready to get off the phone and apologized quickly. Keep a log next to your phone for the first 30 days and file it with your phone bill after that. (You will then have a

Hello **ChaTo**
Total Points 340
Level 2

Categories

- All Categories
- **Consumer Electronics**
- Camcorders
- Cameras
- Cell Phones & Plans
- Games & Gear
- Home Theater
- **Land Phones**
- Music & Music Players
- PDAs & Handhelds
- TiVO & DVRs
- TVs
- Other - Electronics

Finding high-quality content in social media

- A lot of social-media sites in which users publish their own content
- Various types of activities and information: links, social ties, comments, feedback, views, votes, stars, user status, etc.
- Quality of published items can vary greatly
- Highly relevant information might be present
- But, how do we find it?

jumpcut™
Make Amazing Movies Online

askville BETA
by amazon

GARAGEBAND

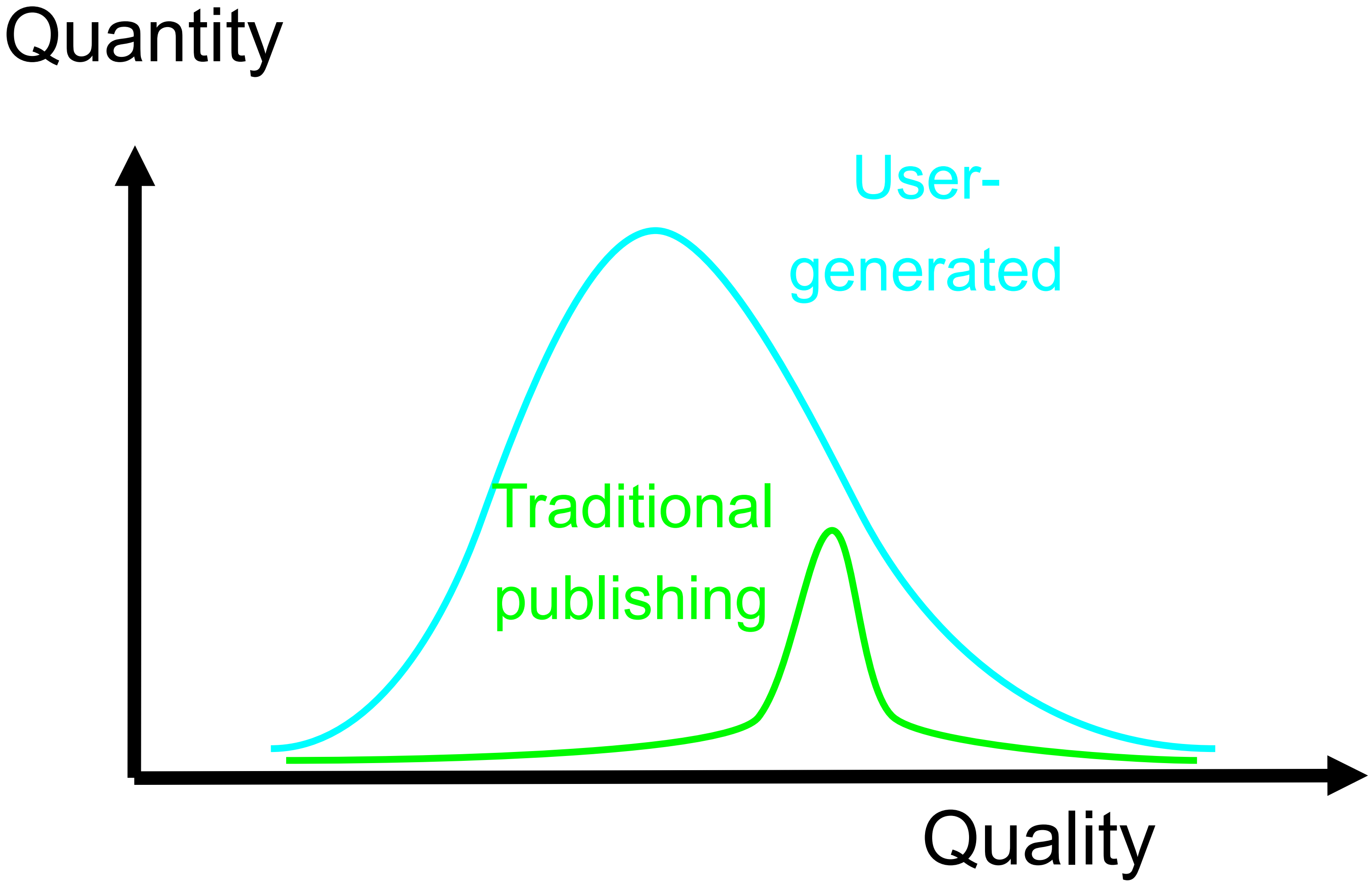
blinkx

flickr BETA™

YouTube

WIKIPEDIA

Mahalo
Beta


Quantity
User-
generated
Traditional
publishing
Quality

kieran.b...

Resolved Question

Show me another »

Do girls like computer geeks / nerds?

2 weeks ago

Report It

tabitha c

not really

2 weeks ago

0 | 1 | Report It

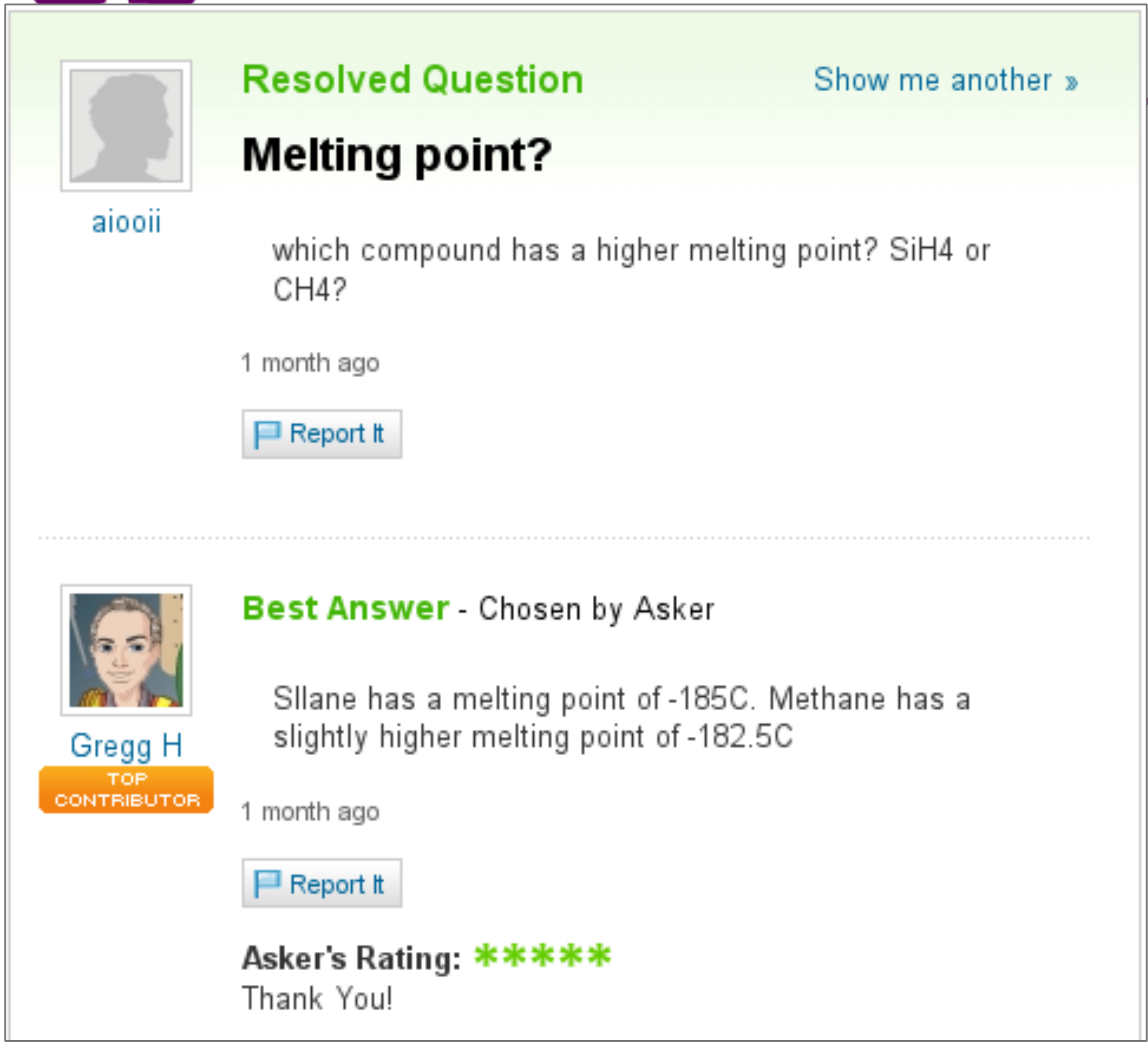
Ella G

a little geekiness is endearing, as long as they still have social skills and good personal hygiene!

2 weeks ago

1 | 0 | Report It

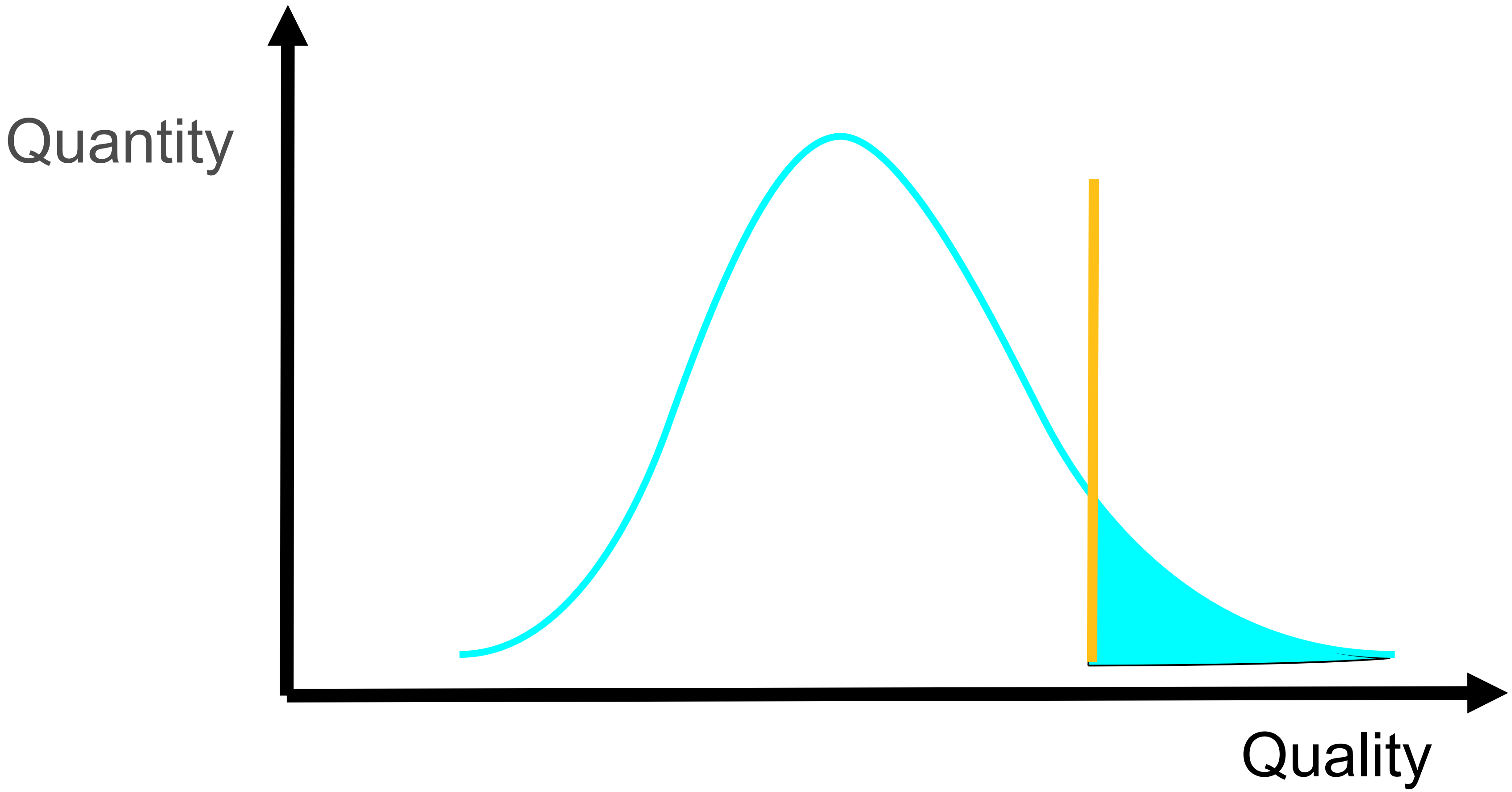
Q. Su, D. Pavlov, J.-H. Chow, W. C. Baker. “Internet-scale collection of human-reviewed data”.WWW'07.



17%-45% of answers were correct

65%-90% of questions had at least one correct answer

Task: find high-quality items



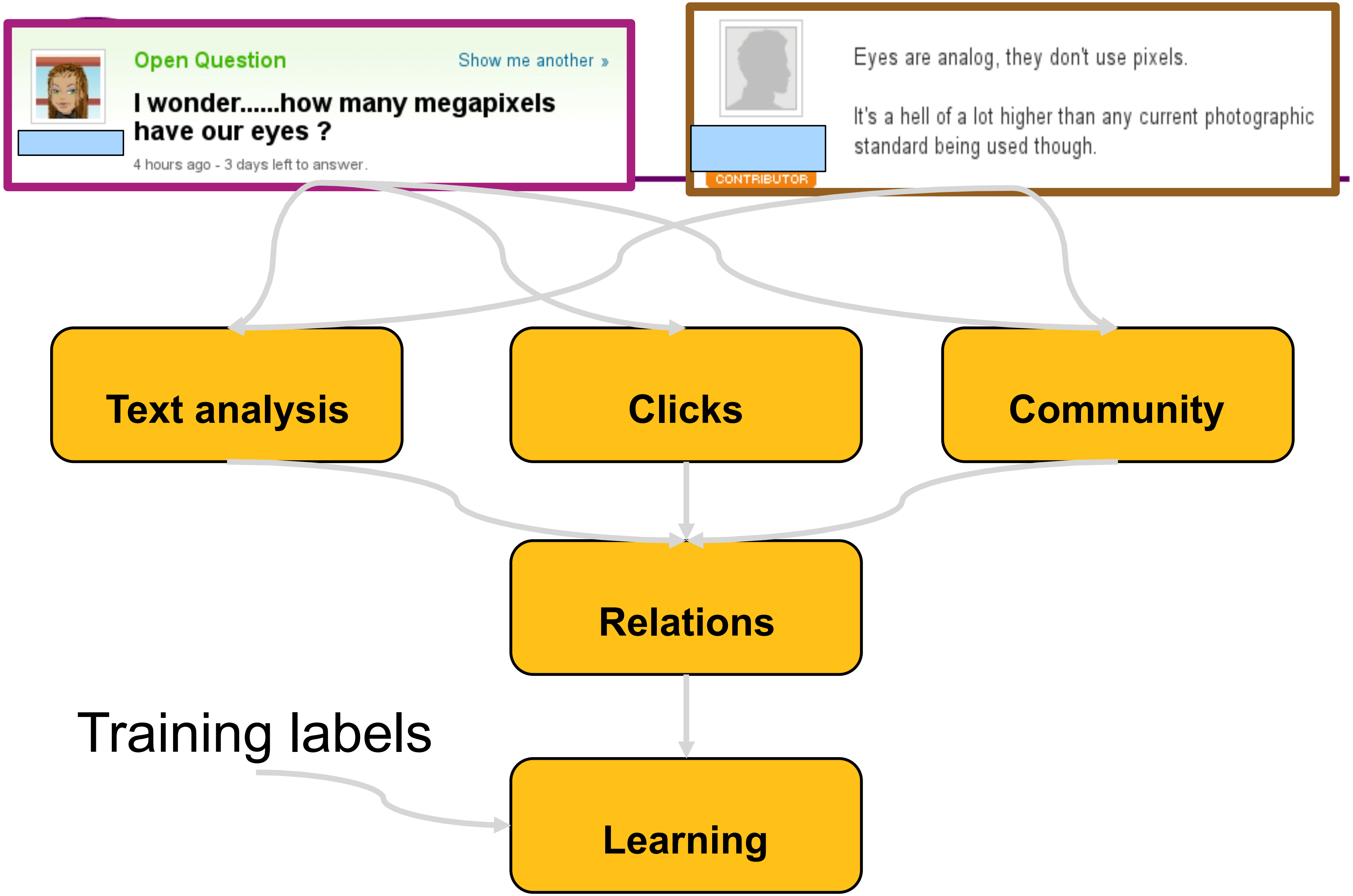
Existing techniques

- Information retrieval methods
- Automatic text analysis
- Link-based ranking methods
- Propagation of trust/distrust
- Usage mining

Sources of information

- Content
- Usage data (clicks)
- Community ratings

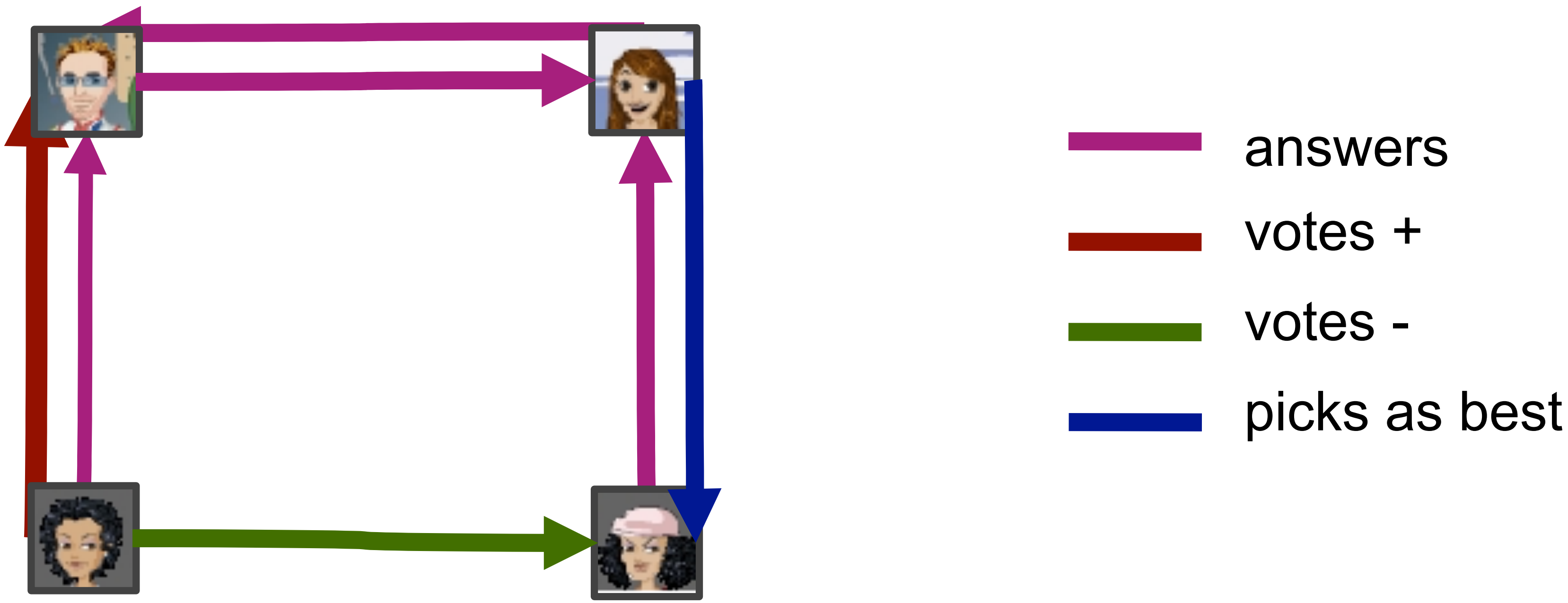
- ...but sparse, noisy, and with spam...


Open Question
Show me another »
I wonder......how many megapixels have our eyes ?
4 hours ago - 3 days left to answer.
Eyes are analog, they don't use pixels.
It's a hell of a lot higher than any current photographic standard being used though.
CONTRIBUTOR
Text analysis
Clicks
Community
Relations
Training labels
Learning

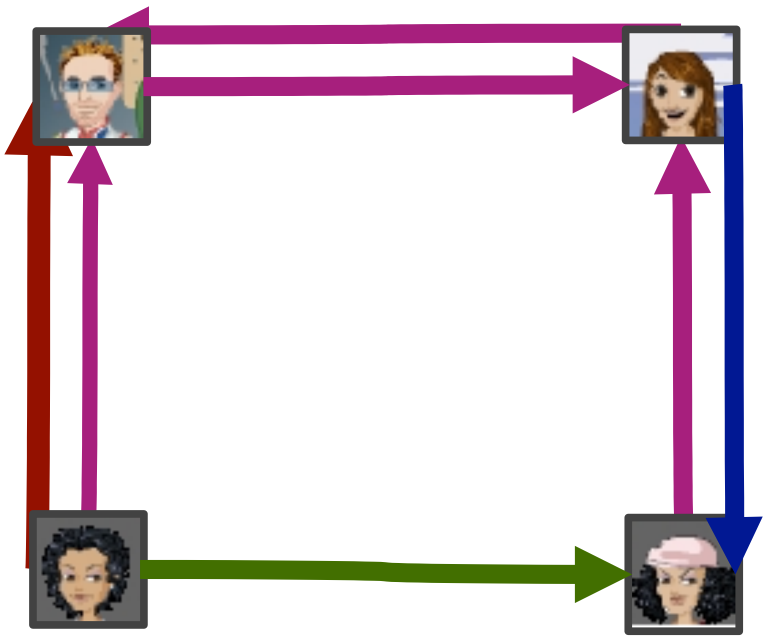
Combining the existing information

- **Text features**
  - Distribution of n-grams
- **Linguistic features**
  - Punctuation, syntactic, case, part-of-speech tags
- **Social features**
  - Consider user-interaction graphs:
    - G1: user A answers a question of user B
    - G2: user A votes for an answer of user B
  - Apply HITS and PageRank
- **Usage features**
  - Number of clicks
  - Deviation of number of clicks from mean of category

Community



Community



Propagation-based metrics

1. Pagerank score
2. HITS hub score
3. HITS authority score

Computed on each graph

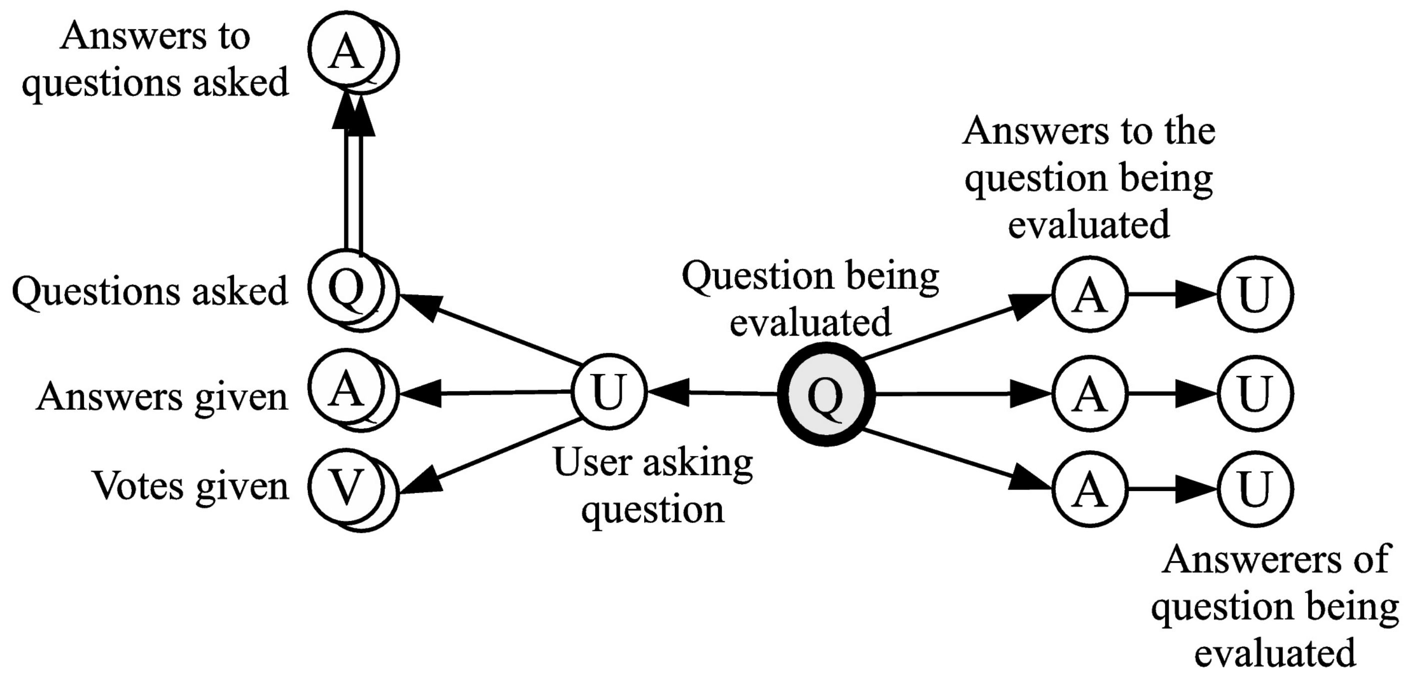
Relations

Question quality

Answer quality		High	Medium	Low
	High	41%	15%	8%
	Medium	53%	76%	74%
	Low	6%	9%	18%
		100%	100%	100%

Question quality and answer quality are not independent

Propagation of features



Task: high-quality questions

	Precision	Recall	AUC
N-grams (N)	65%	48%	0.52
N+ text analysis	76%	65%	0.65
N+ clicks	68%	57%	0.58
N+ relations	74%	65%	0.66
All	79%	77%	0.76

Challenges in social media

- What's the ratings and reputation system?
- How do you cope with spam?
  - The wisdom of the crowd can be used against spammers
- The bigger challenge: where else can you exploit the power of the people?
- What are the incentive mechanisms?
  - Example: ESP game

Discussion

- Relevant content is available in social media, but the variance of the quality is very high
- Classifying questions/answers is different than document classification
- Combine many orthogonal features and heterogeneous information

Overall summary

- **Open problems and challenges:**
  - Manage and integrate highly heterogeneous information:
  - Content, links, social links, tags, feedback, usage logs, wisdom of crowd, etc.
  - Model and benefit from evolution
  - Battle adversarial attempts and collusions

Web Search Queries

- **Cultural and educational diversity**
- **Short queries & impatient interaction**
  - few queries posed & few answers seen
- **Smaller & different vocabulary**
- **Different user goals [Broder, 2000]:**
  - Information need
  - Navigational need
  - Transactional need
- **Refined by Rose & Levinson, WWW 2004**

User Needs

- Need (Broder 2002)
  - **Informational** – want to learn about something (~40% / 65%)

    `Low hemoglobin`

  - **Navigational** – want to go to that page (~25% / 15%)

    `United Airlines`

  - **Transactional** – want to do something (web-mediated) (~35% / 20%)
    - Access a service
    - Downloads
    - Shop

    `Edinburgh weather`

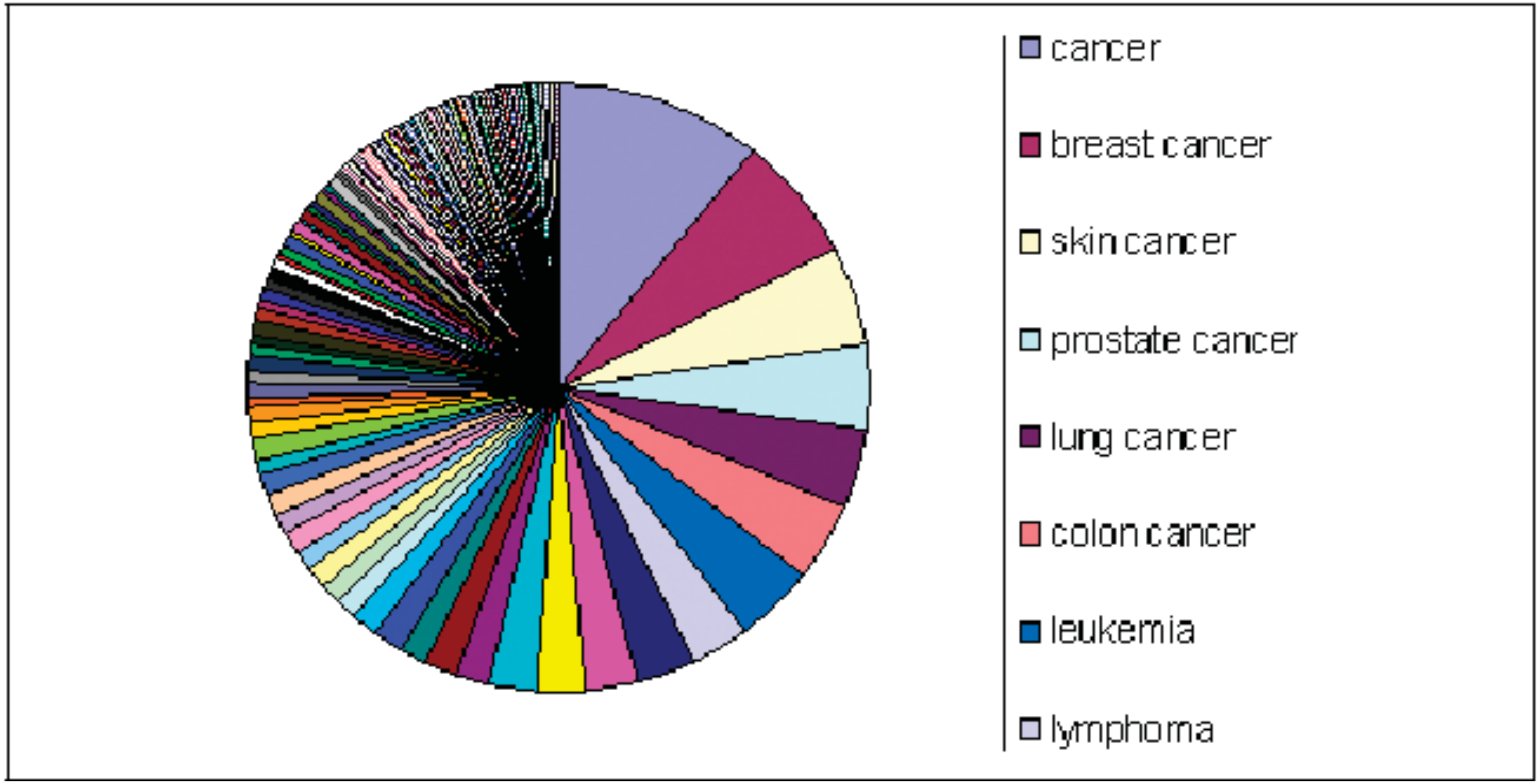
    `Mars surface images`

    `Canon S410`

  - Gray areas
    - Find a good hub
    - Exploratory search "see what's there"

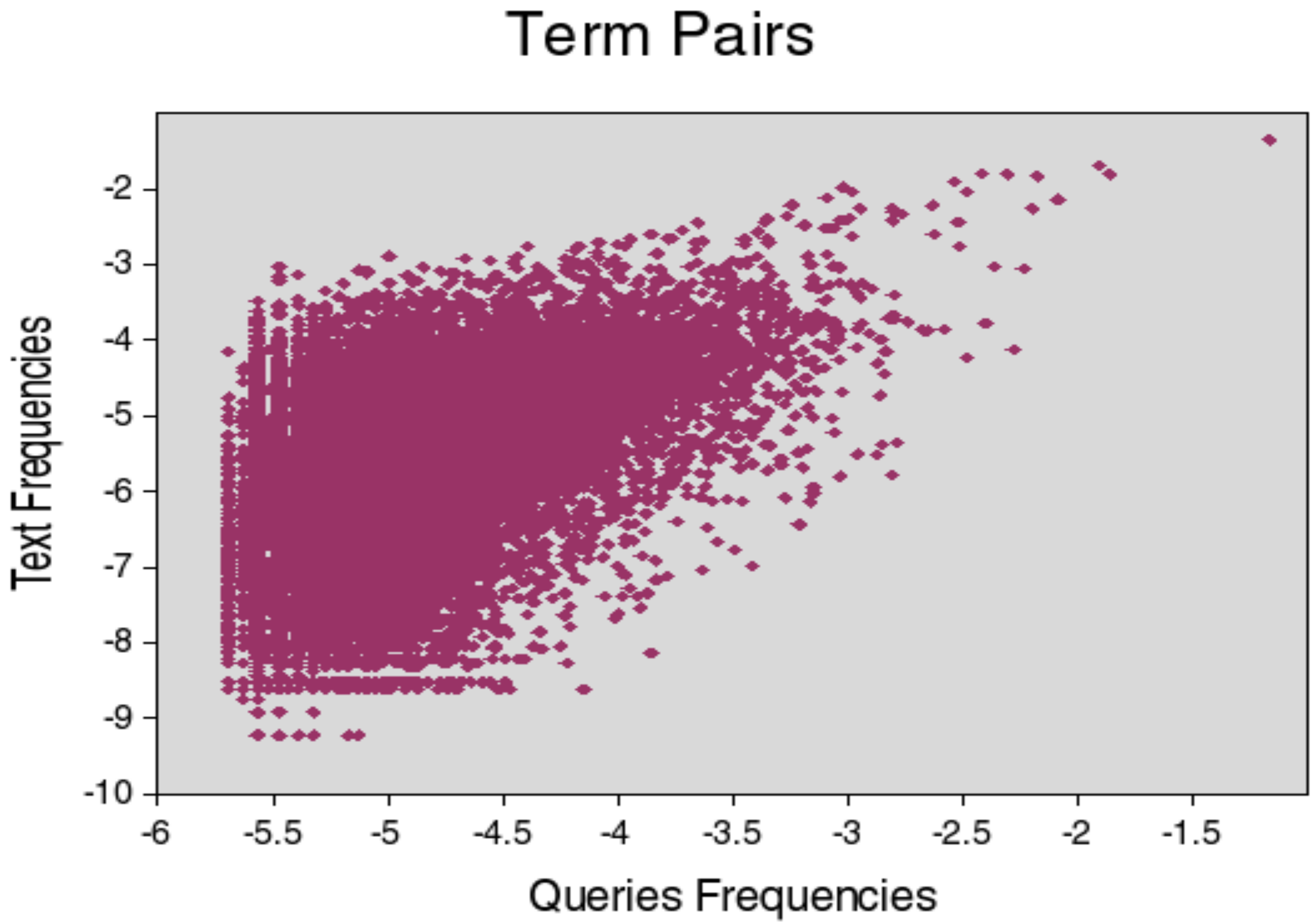
    `Car rental Brasil`

Query Distribution

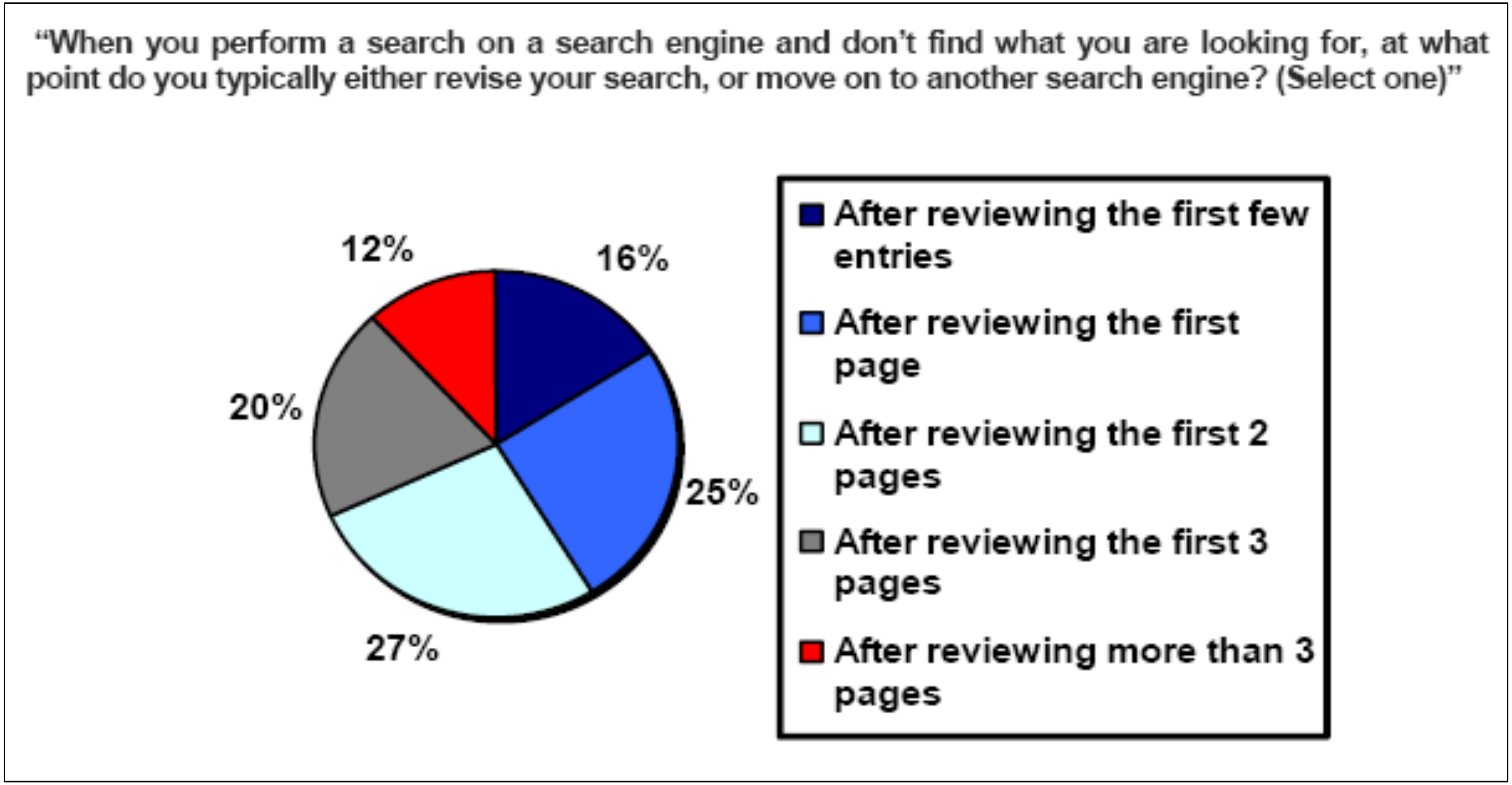


**Power law: few popular broad queries,
many rare specific queries**

Queries and Text



How far do people look for results?



(Source: iprospect.com WhitePaper_2006_SearchEngineUserBehavior.pdf)

Typical Session

- Two queries of
- .. two words, looking at…
- .. two answer pages, doing
- .. two clicks per page

- What is the goal?

MP3

games

cars

britney spears

pictures

ski

Relevance of the Context

- **There is no information without context**
- **Context and hence, content, will be implicit**
- **Balancing act: information vs. form**
- **Brown & Diguid: *The social life of information* (2000)**
  - Current trend: less information, more context
- **News highlights are similar to Web queries**
  - E.g.: *Spell Unchecked (Indian Express, July 24, 2005)*

Context

- ***Who you are*: age, gender, profession, etc.**
- ***Where you are and when*: time, location, speed and direction, etc.**
- ***What you are doing*: interaction history, task in hand, searching device, etc.**
- ***Issues*: privacy, intrusion, will to do it, etc.**
- ***Other sources*: Web, CV, usage logs, computing environment, ...**
- ***Goals*: personalization, localization, better ranking in general, etc.**

Context in Web Queries

- ***Session: (q, (URL, t)*)⁺***

- ***Who you are***: **age, gender, profession (IP), etc.**
- ***Where you are and when***: **time, location (IP), speed and direction, etc.**
- ***What you are doing***: **interaction history, task in hand, etc.**
- ***What you are using***: **searching device (operating system, browser, ...)**

SEARCH GOAL	DESCRIPTION	EXAMPLES
1. Navigational	My goal is to go to specific known website that I already have in mind. The only reason I'm searching is that it's more convenient than typing the URL, or perhaps I don't know the URL.	aloha airlines duke university hospital kelly blue book
2. Informational	My goal is to learn something by reading or viewing web pages	Home page
2.1 Directed	I want to learn something in particular about my topic	
2.1.1 Closed	I want to get an answer to a question that has a single, unambiguous answer.	what is a supercharger 2004 election dates
2.1.2 Open	I want to get an answer to an open-ended question, or one with unconstrained depth.	baseball death and injury why are metals shiny
2.2 Undirected	I want to learn anything/everything about my topic. A query for topic X might be interpreted as "tell me about X."	color blindness jfk jr
2.3 Advice	I want to get advice, ideas, suggestions, or instructions.	help quitting smoking walking with weights
2.4 Locate	My goal is to find out whether/where some real world service or product can be obtained	pella windows phone card
2.5 List	My goal is to get a list of plausible suggested web sites (I.e. the search result list itself), each of which might be candidates for helping me achieve some underlying, unspecified goal	travel amsterdam universities florida newspapers
3. Resource	My goal is to obtain a resource (not information) available on web pages	Hub page
3.1 Download	My goal is to download a resource that must be on my computer or other device to be useful	kazaa lite mame roms
3.2 Entertainment	My goal is to be entertained simply by viewing items available on the result page	Page with resources xxx porno movie free live camera in l.a.
3.3 Interact	My goal is to interact with a resource using another program/service available on the web site I find	weather measure converter
3.4 Obtain	My goal is to obtain a resource that does not require a computer to use. I may print it out, but I can also just look at it on the screen. I'm not obtaining it to learn some information, but because I want to use the resource itself.	free jack o lantern patterns ellis island lesson plans house document no. 587

Rose & Levinson 2004

Kang & Kim, SIGIR 2003

- **Features:**
  - Anchor usage rate
  - Query term distribution in home pages
  - Term dependence
- **Not effective: 60%**

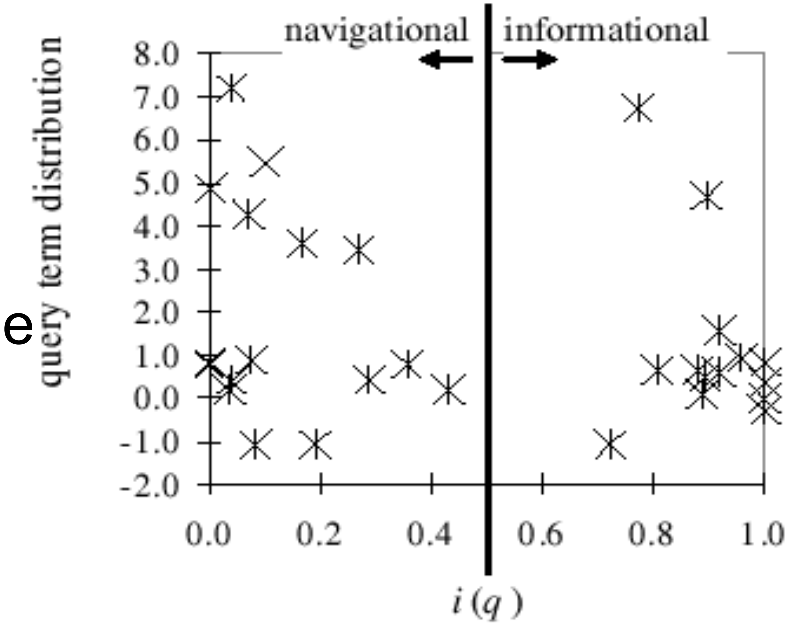


Figure 16: Query term distribution

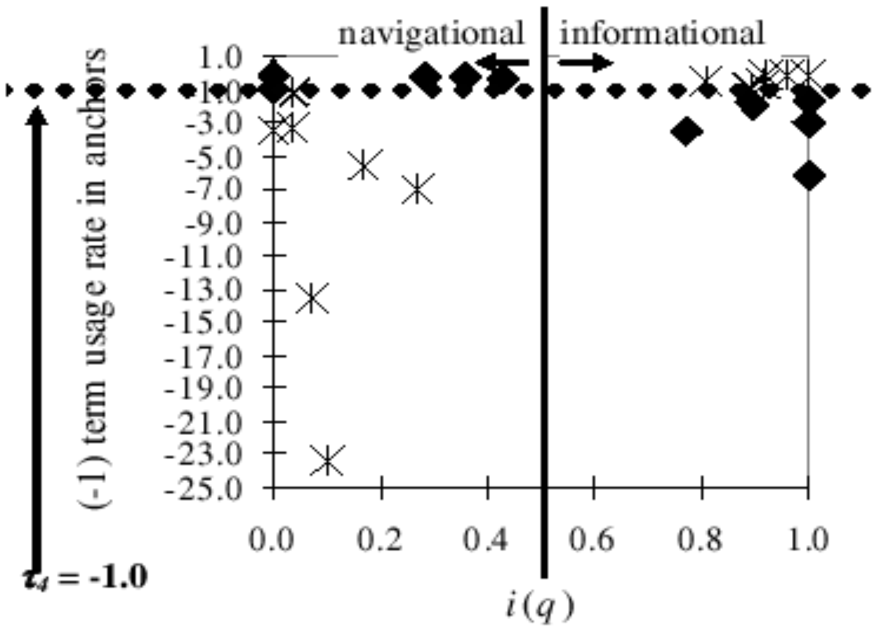


Figure 15: Anchor usage rate

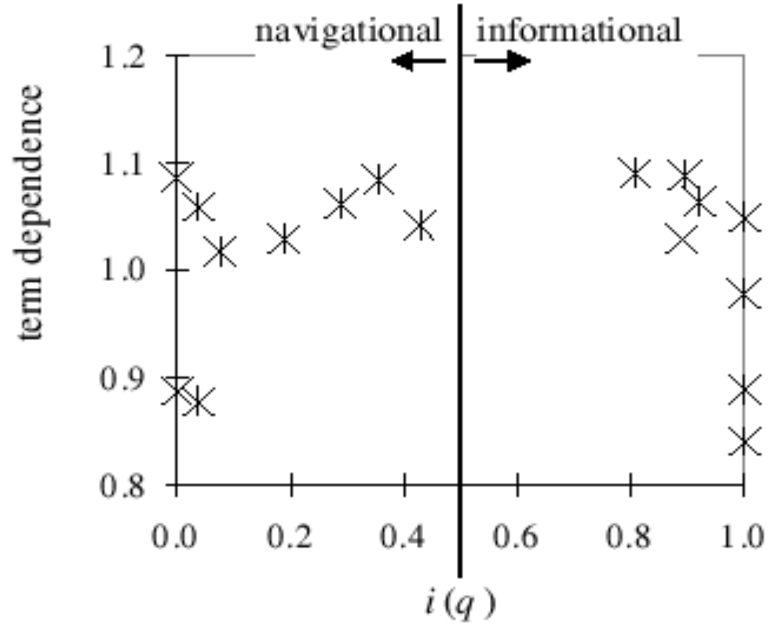


Figure 17: Term dependence

User Goals

- Liu, Lee & Cho, WWW 2005
- Top 50 CS queries
- Manual Query Classification: 28 people
- Informational goal *i(q)*
- *Remove software & person-names*
- *30 queries left*

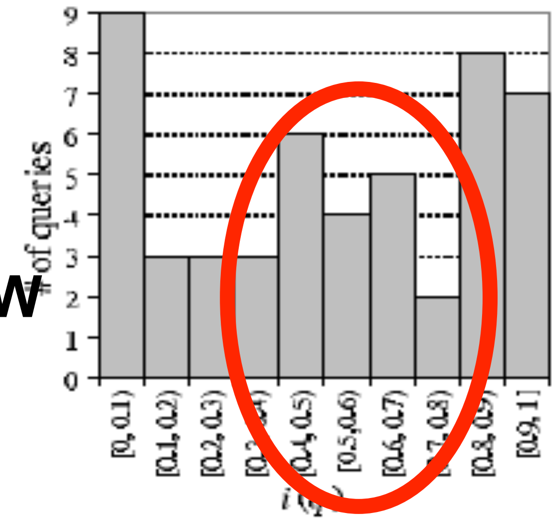


Figure 1: Query distribution along the $i(q)$ axis

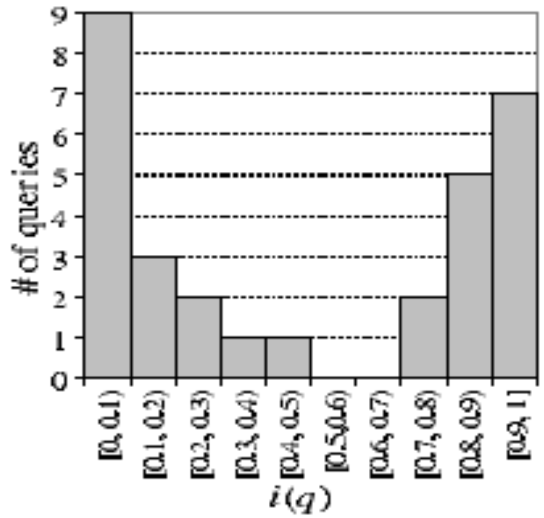


Figure 2: After removing software and person-name queries

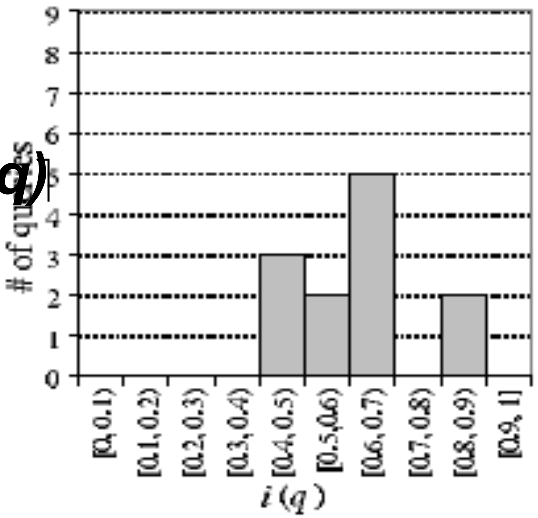


Figure 3: Distribution of the 12 software queries

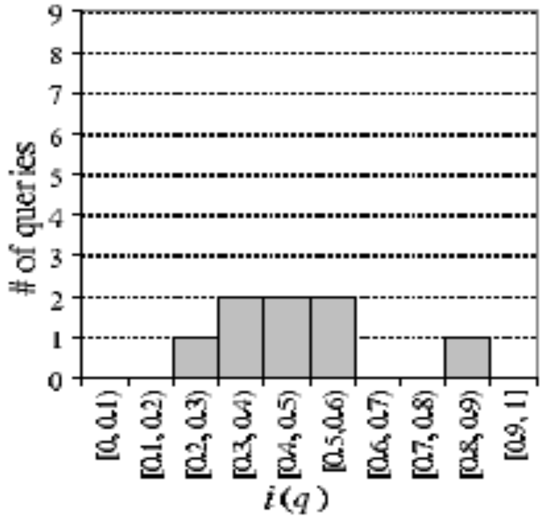
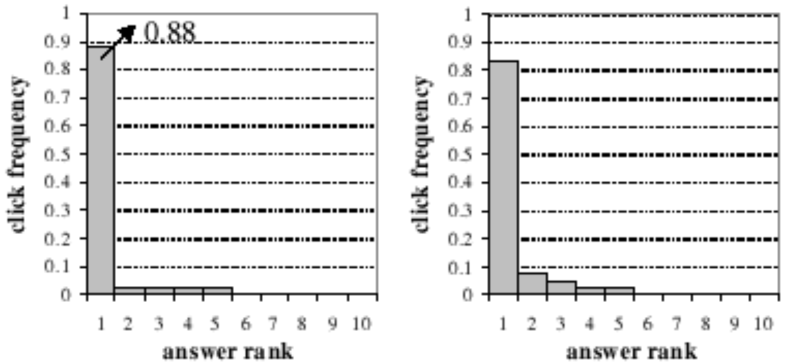


Figure 4: Distribution of the 8 person-name queries

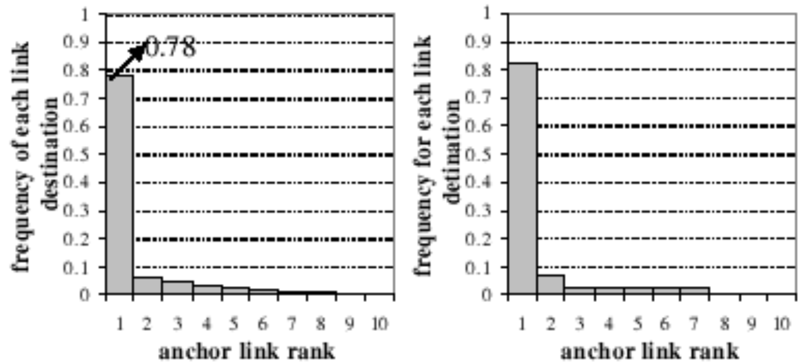
Features

- Click & anchor text distribution



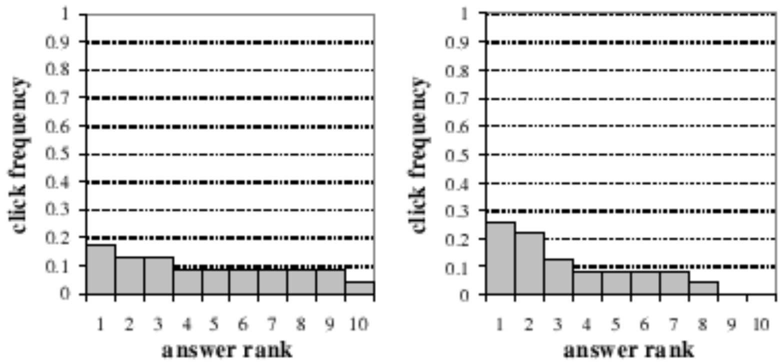
(a) pubmed ($i(q)$=0.1) (b) ucla library ($i(q)$=0)

Figure 5: Click distributions for sample navigational queries



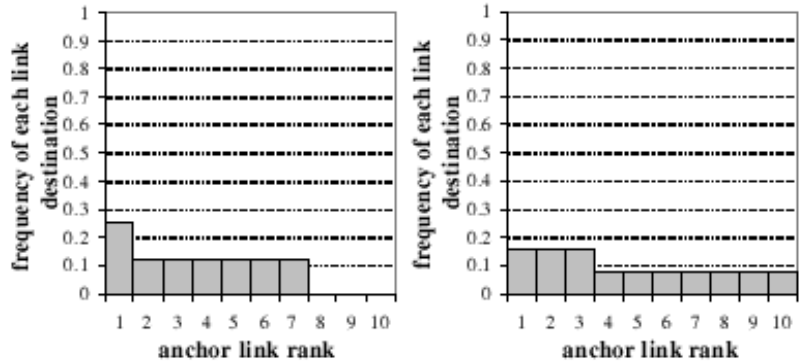
(a) pubmed ($i(q)$=0.1) (b) ucla library ($i(q)$=0)

Figure 7: Anchor-link distributions for sample navigational queries



(a) hidden markov model ($i(q)$=1) (b) simulated annealing ($i(q)$=1)

Figure 6: Click distributions for sample informational queries



(a) hidden markov model ($i(q)$=1) (b) simulated annealing ($i(q)$=1)

Figure 8: Anchor-link distributions for sample informational queries

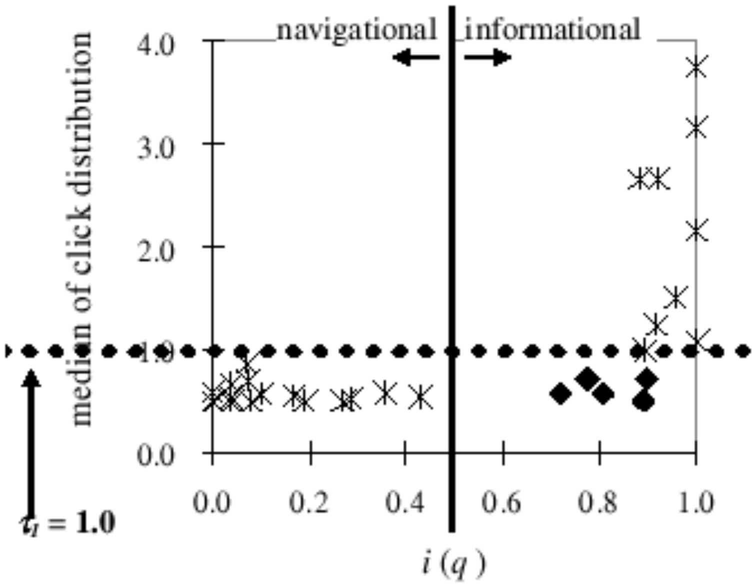


Figure 11: Median of click distribution

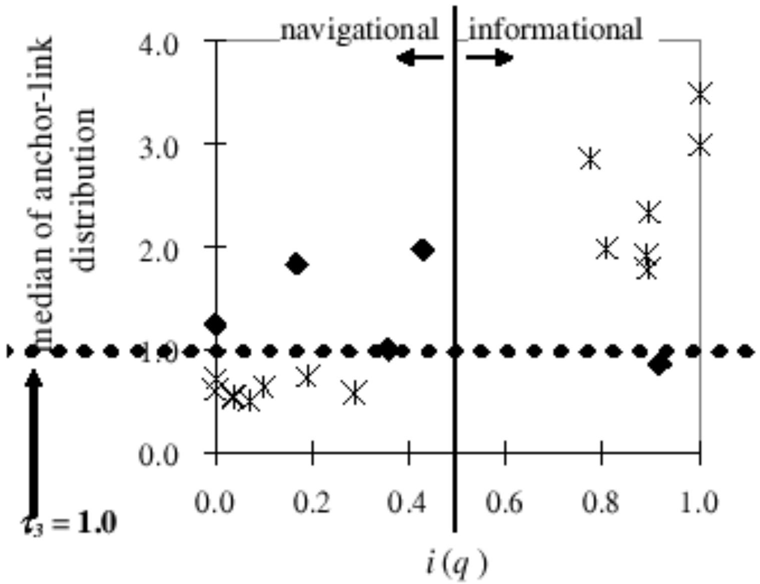


Figure 13: Median of anchor-link distribution

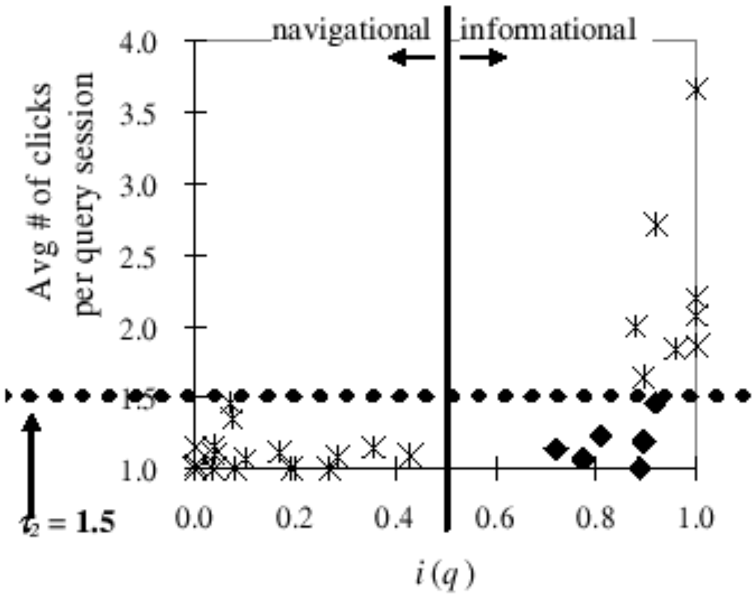


Figure 12: Avg # of clicks per query

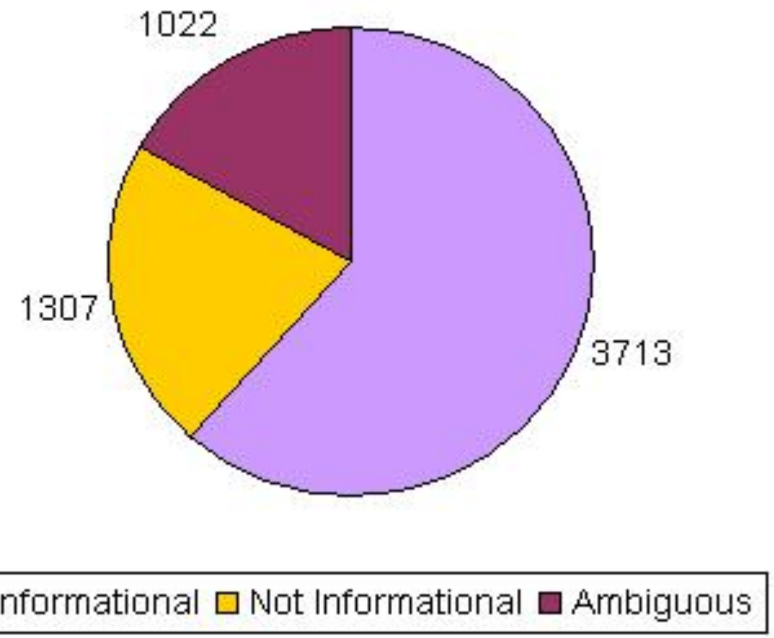
- Prediction power:
- Single features: 80%
- Mixed features: 90%

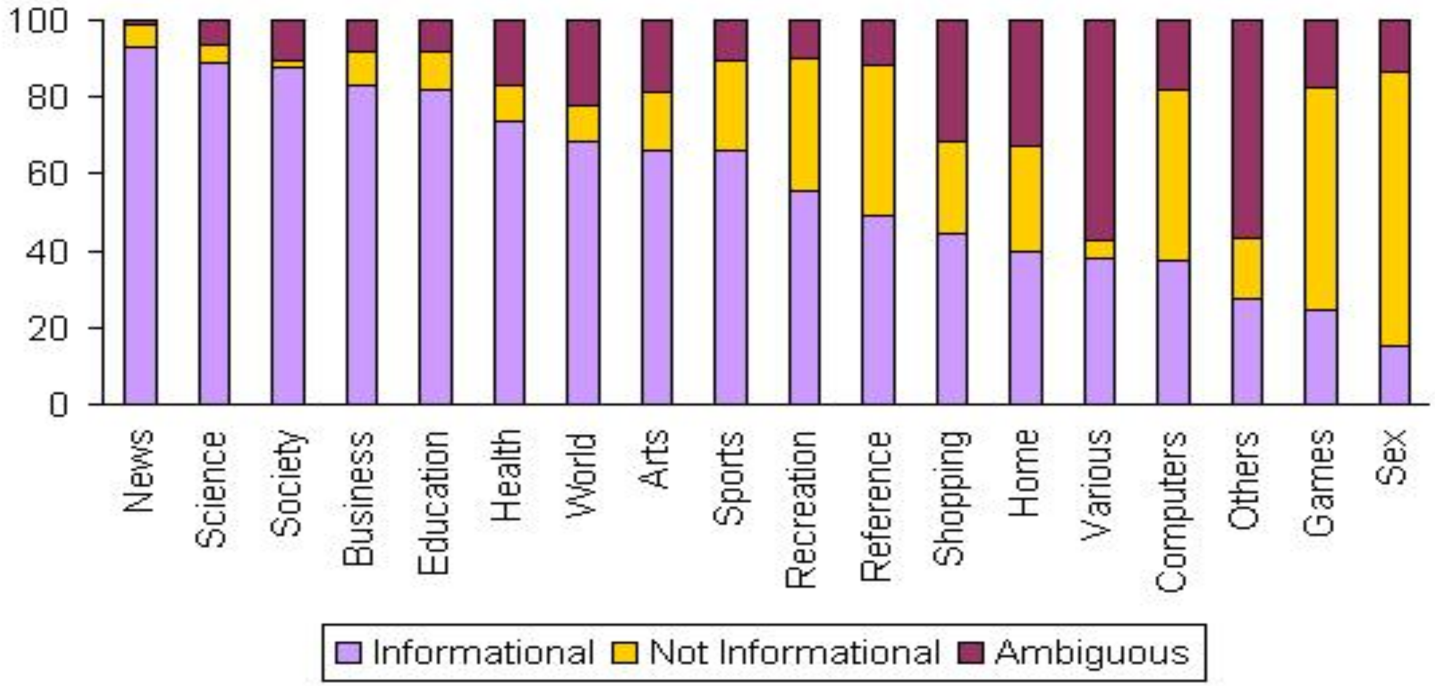
- Drawbacks:
  - Small evaluation
  - a posteriori feature

User Intention

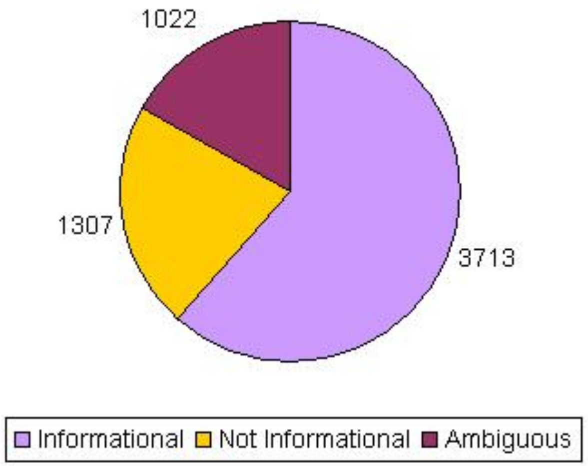
- Manual classification of more than 6,000 popular queries
- Query Intention & topic
- Classification & Clustering
- Machine Learning on all the available attributes
- Baeza-Yates, Calderon & Gonzalez (SPIRE 2006)

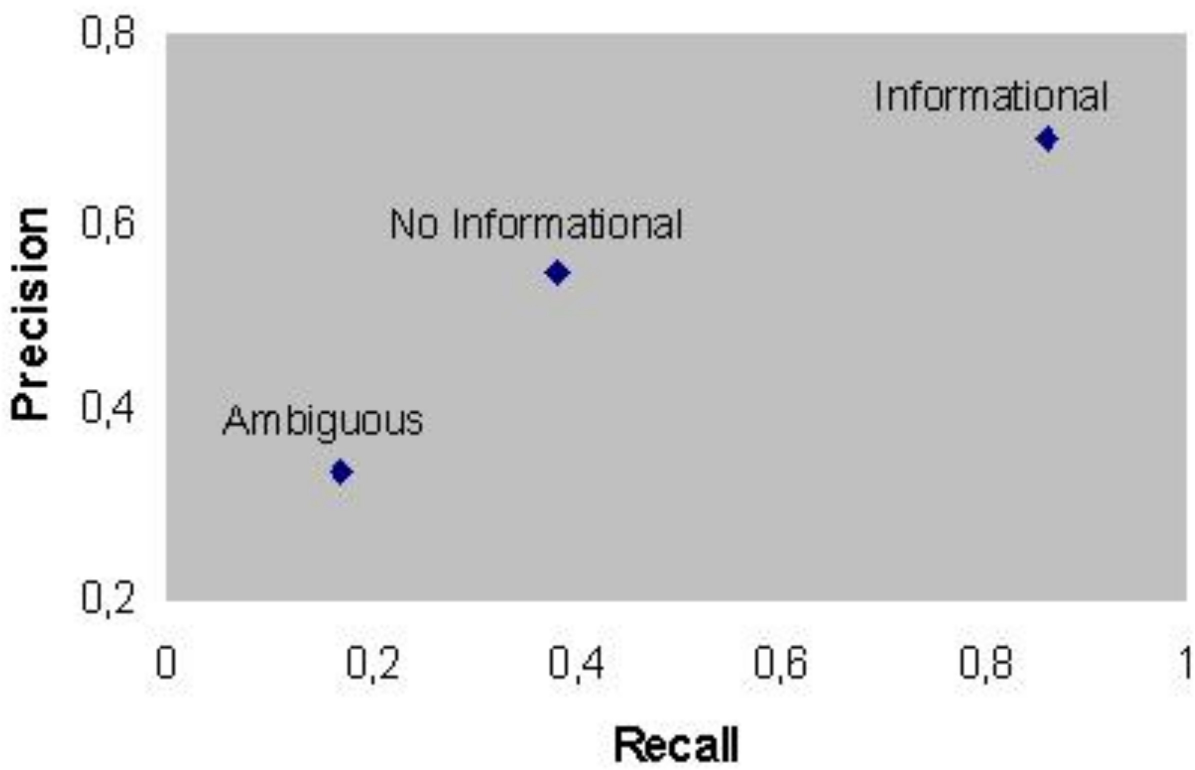
Classified Queries





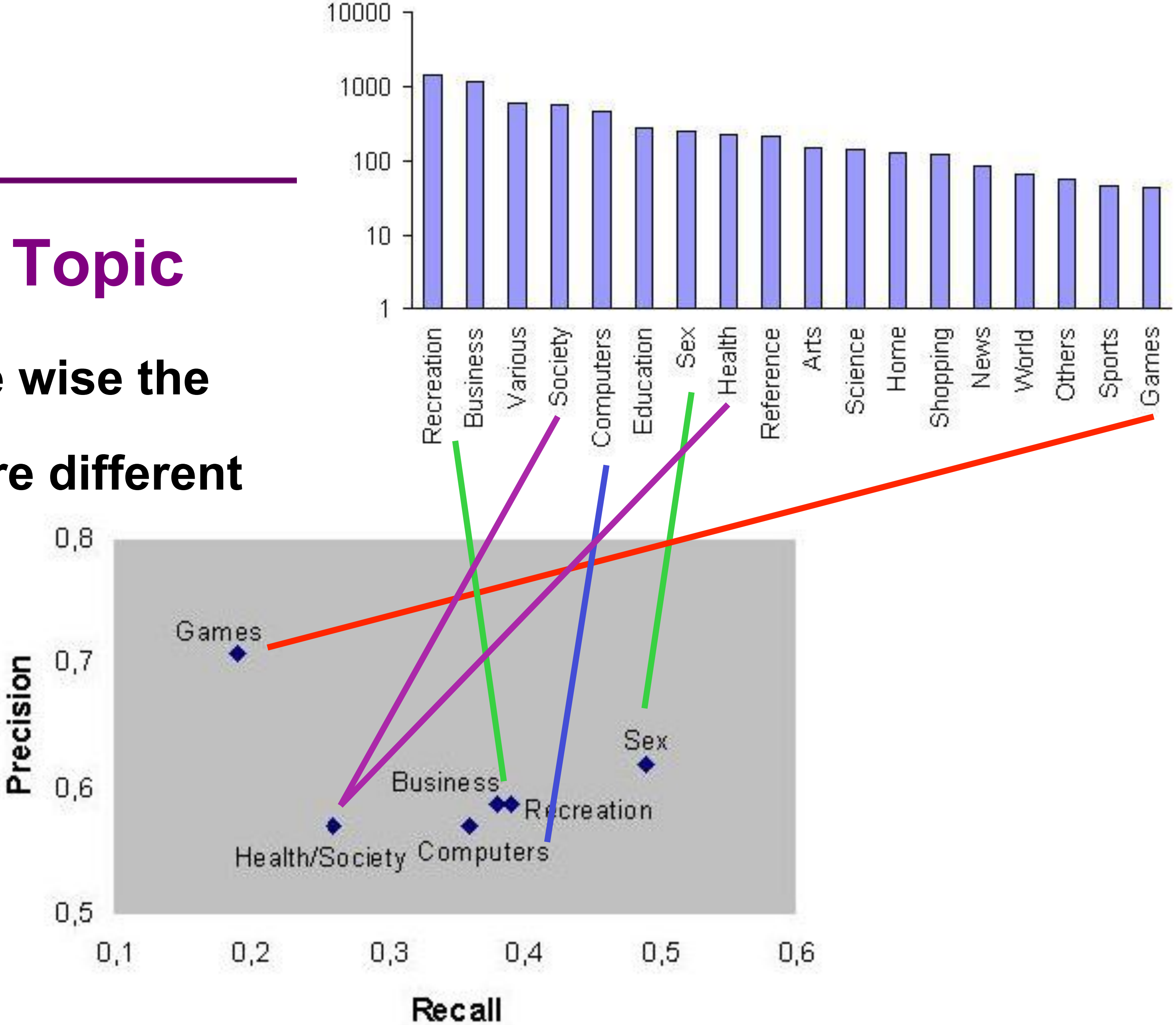
Results: User Intention

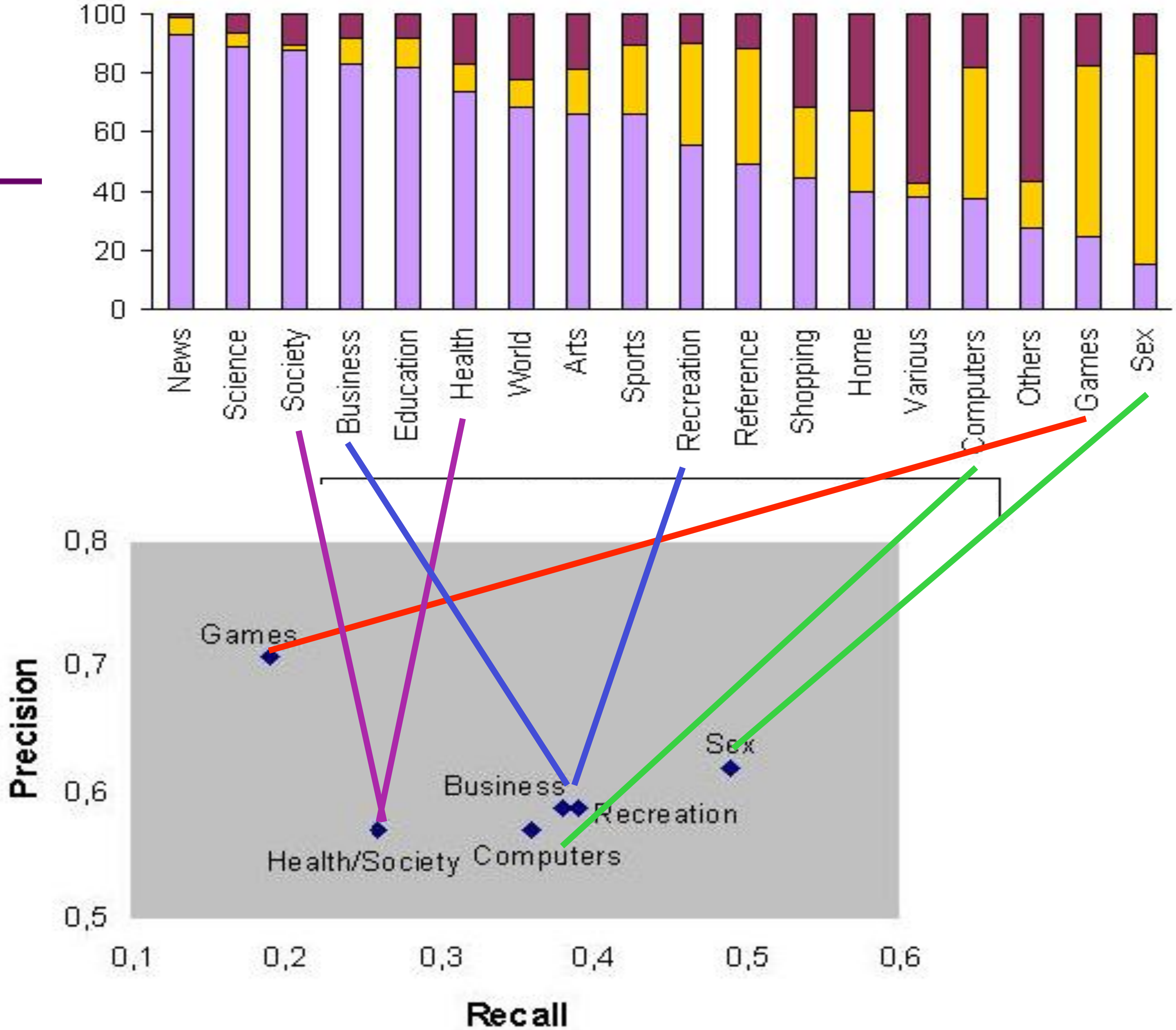




Results: Topic

- **Volume wise the results are different**





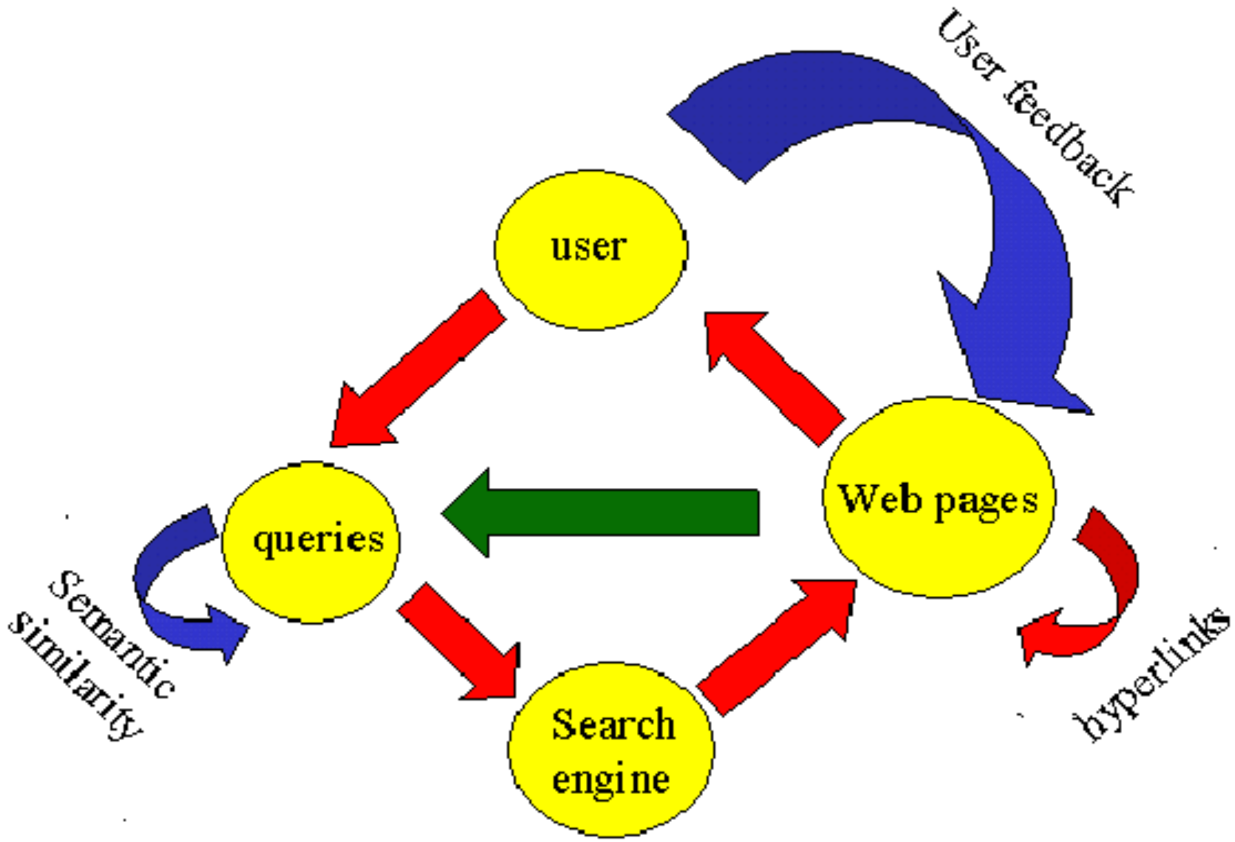
100
80
60
40
20
0
News
Science
Society
Business
Education
Health
World
Arts
Sports
Recreation
Reference
Shopping
Home
Various
Computers
Others
Games
Sex
0,8
0,7
0,6
0,5
Precision
Games
Sex
Business
Recreation
Health/Society
Computers
0,1
0,2
0,3
0,4
0,5
0,6
Recall

Clustering Queries

- **Define relations among queries**
  - Common words: sparse set
  - Common clicked URLs: better
  - Natural clusters
- **Define distance function among queries**
  - Content of clicked URLs
    [Baeza-Yates, Hurtado & Mendoza, 2004]
  - Summary of query answers [Sahami, 2006]

Goals

- Can we cluster queries well?
- Can we assign user goals to clusters?



Our Approach

- **Cluster text of clicked pages**
  - Infer query clusters using a vector model

$$\boldsymbol{q}[i] = \sum_{URL u} \frac{\mathtt{Pop}(q,u) \times \mathtt{Tf}(t_i,u)}{\max_t \mathtt{Tf}(t,u)}$$

- **Pseudo-taxonomies for queries**
  - Real language (slang?) of the Web
  - Can be used for classification purposes

Clusters Examples

Q	Cluster Rank	ISim	ESim	Queries in Cluster	Descriptive keywords
q_1	252	0,447	0,007	car sales, cars Iquique, cars used, diesel, new cars,	cars $(49, 4\%)$, used $(14, 2\%)$, stock $(3, 8\%)$, pickup truck $(3, 7\%)$, jeep $(1, 6\%)$
q_2	497	0,313	0,009	stamp, serigraph inputs, ink reload, cartridge	print $(11, 4\%)$, ink $(7, 3\%)$, stamping $(3, 8\%)$, inkjet $(3, 6\%)$
q_3	84	0,697	0,015	office rental, rentals in Santiago, real state, apartment rental	office $(11, 6\%)$, building $(7, 5\%)$, real state $(5, 9\%)$, real state agents $(4, 2\%)$

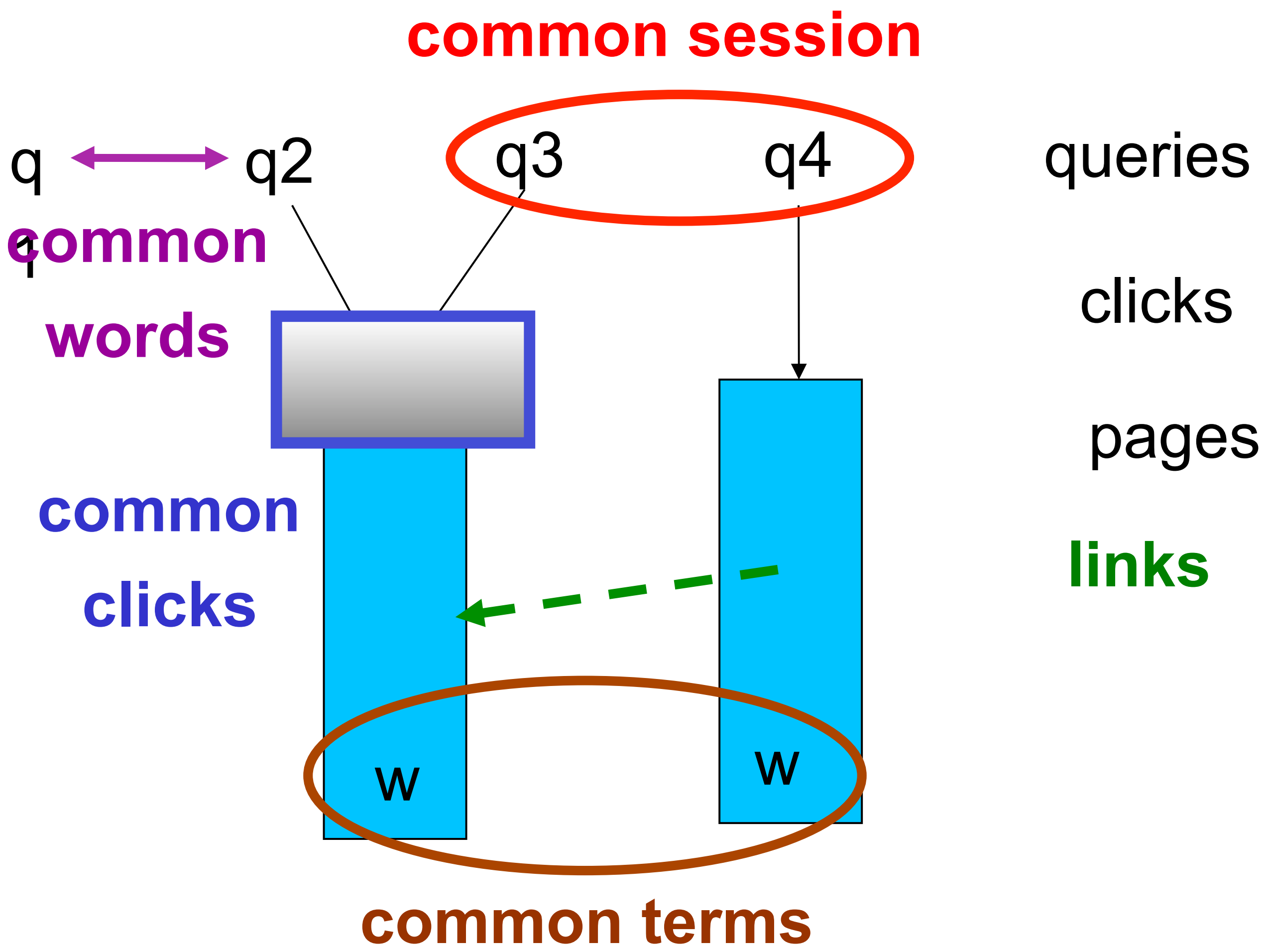
Using the Clusters

- **Improved ranking** **Baeza-Yates, Hurtado & Mendoza Journal of ASIST 2007**
- **Word classification**
  - Synonyms & related terms are in the same cluster
  - Homonyms (polysemy) are in different clusters
- **Query recommendation (ranking queries!)**
  - F $\text{Rank}(q) = \gamma \times \text{Sup}(q, q_{ini}) + (1 - \gamma) \times \text{Clos}(q)$

Query Recommendation

Query	Popularity	Support	Closedness	Rank
rentals apartments viña del mar owners	2	0,133	0,403	0,268
rentals apartments viña del mar	10	0,2	0,259	0,229
viel properties	4	0,1	0,315	0,207
rental house viña del mar	2	0,166	0,121	0,143
house leasing rancagua	8	0,166	0,0385	0,102
quintero	2	0,166	0,024	0,095
rentals apartments cheap vina del mar	3	0,033	0,153	0,093
subsidize renovation urban	5	0,133	0,001	0,067
houses being sold in pucon	10	0	0,114	0,057
apartments selling pucon villarrica	2	0,066	0,015	0,040
portal sell properties	3	0,033	0,023	0,028
sell house	2	0,033	0,017	0,025
sell lots pirque	2	0,033	0,0014	0,017
canete hotels	1	0	0,011	0,005

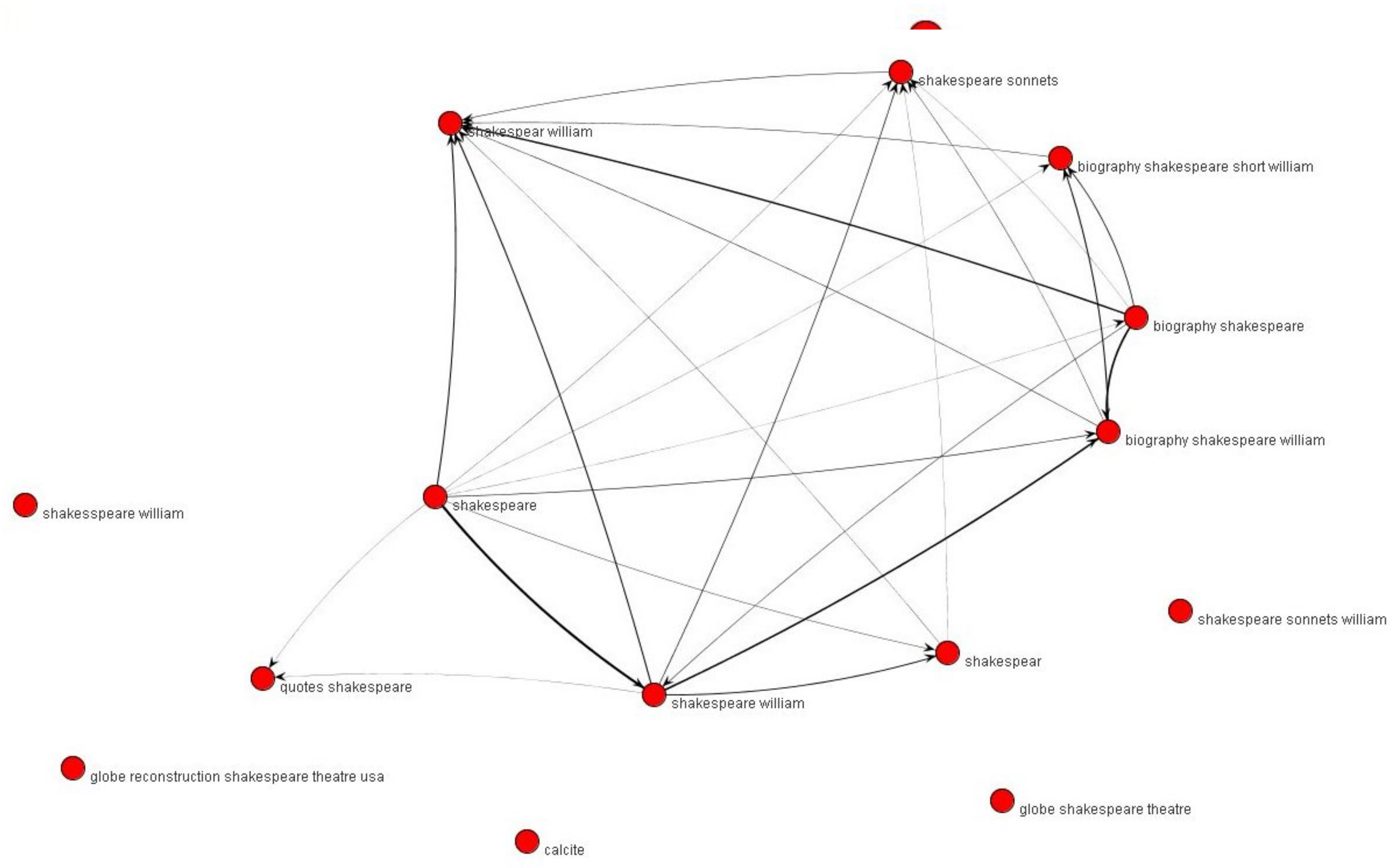
Relating Queries (Baeza-Yates, 2007)



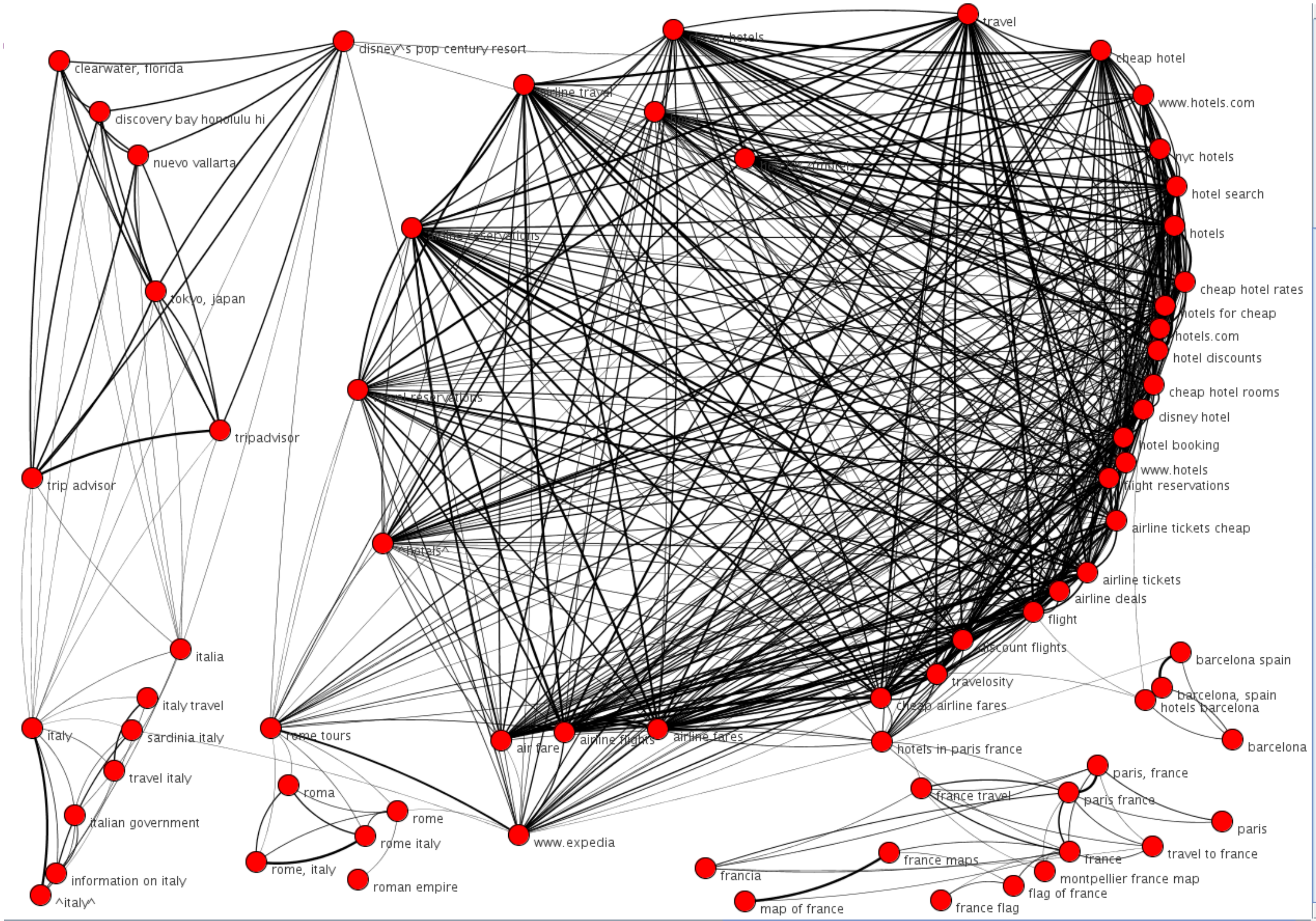
Qualitative Analysis

Graph	Strength	Sparsity	Noise
Word	Medium	High	Polysemy
Session	Medium	High	Physical sessions
Click	High	Medium	**Multitopic pages Click spam**
Link	Weak	Medium	Link spam
Term	Medium	Low	Term spam

Words, Sessions and Clicks



Click Graph



clearwater, florida
disney^s pop century resort
discovery bay honolulu hi
nuevo vallarta
tokyo, japan
tripadvisor
trip advisor
italia
italy travel
italy
sardinia italy
travel italy
italian government
information on italy
^italy^
rome tours
roma
rome
rome italy
rome, italy
roman empire
www.expedia
air fare
airline flights
airline fares
travel
cheap hotel
www.hotels.com
nyc hotels
hotel search
hotels
cheap hotel rates
hotels for cheap
hotels.com
hotel discounts
cheap hotel rooms
disney hotel
hotel booking
www.hotels
flight reservations
airline tickets cheap
airline tickets
airline deals
flight
discount flights
travelosity
cheap airline fares
hotels in paris france
barcelona spain
barcelona, spain
hotels barcelona
barcelona
paris, france
france travel
paris france
paris
travel to france
france
france maps
francia
montpellier france map
flag of france
france flag
map of france

Formal Definition

- **There is an edge between two queries *q* and *q'* if:**
  - There is at least one URL clicked by both
- **Edges can be weighted (for filtering)**
  - We used the cosine similarity in a vector space defined by URL clicks

$$W(e) = \frac{\bar{q} \cdot \bar{q}'}{|\bar{q}|\ |\bar{q}'|} = \frac{\sum_{i \leq D} q(i) \cdot q'(i)}{\sqrt{\sum_{i \leq D} q(i)^2} \cdot \sqrt{\sum_{i \leq D} q'(i)^2}}$$

URL based Vector Space

- **Consider the query *"complex networks"***
- **Suppose for that query the clicks are:**
  - *www.ams.org/featurecolumn/archive/networks1.html* (3 clicks)
  - en.wikipedia.org/wiki/Complex_network (1 click)

0	0	0	0	…	1/4	…	3/4	…	0	0	0	0

"Complex networks"

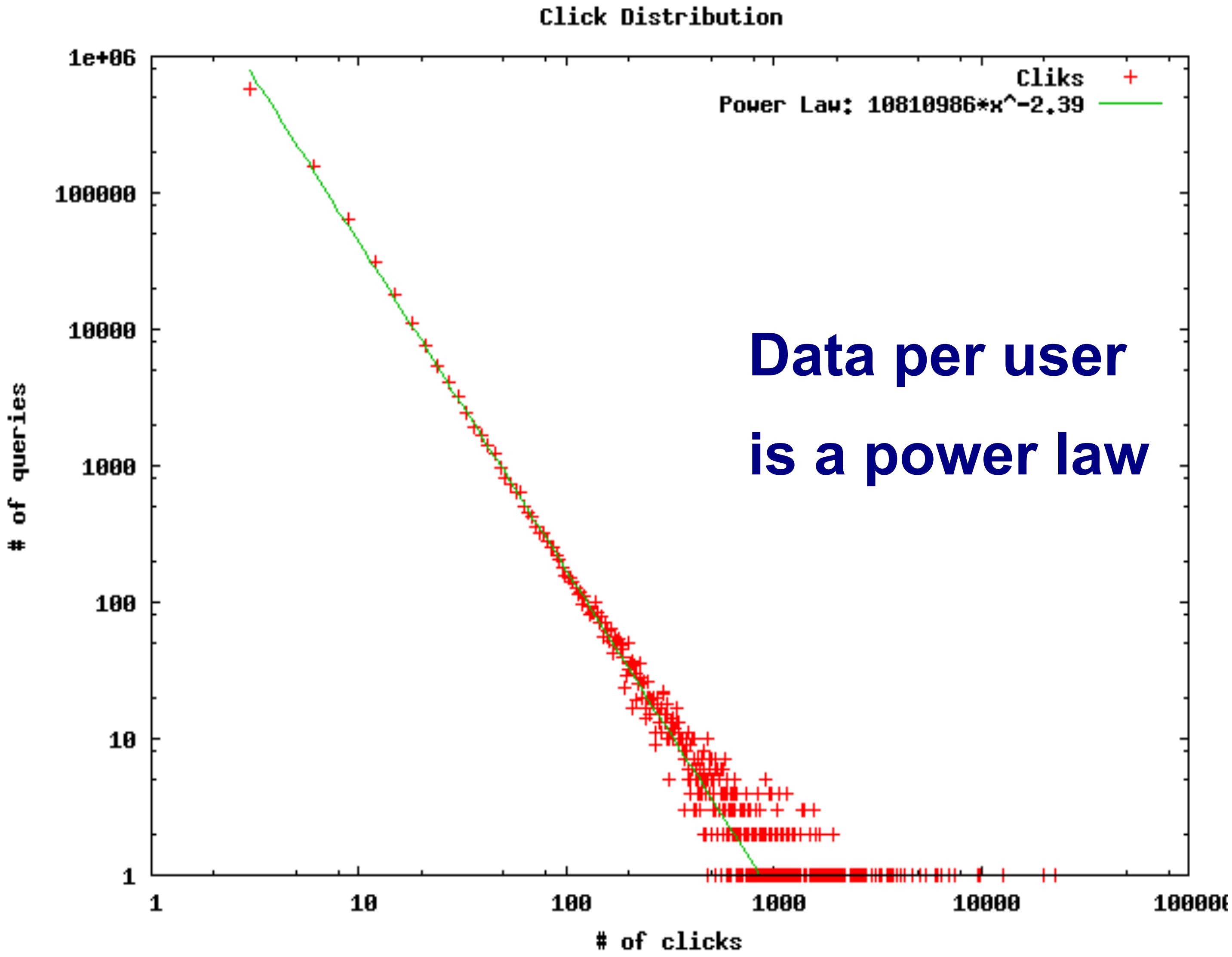
Building the Graph

- **The graph can be built efficiently:**
  - Consider the tuples (query, clicked url)
  - Sort by the second component
  - Each block with the same URL *u* gives the edges induced by *u*
  - Complexity: $O(\max \{M^*|E|, n \log n\})$ where M is the maximum number of URLs between two queries, and n is the number of nodes

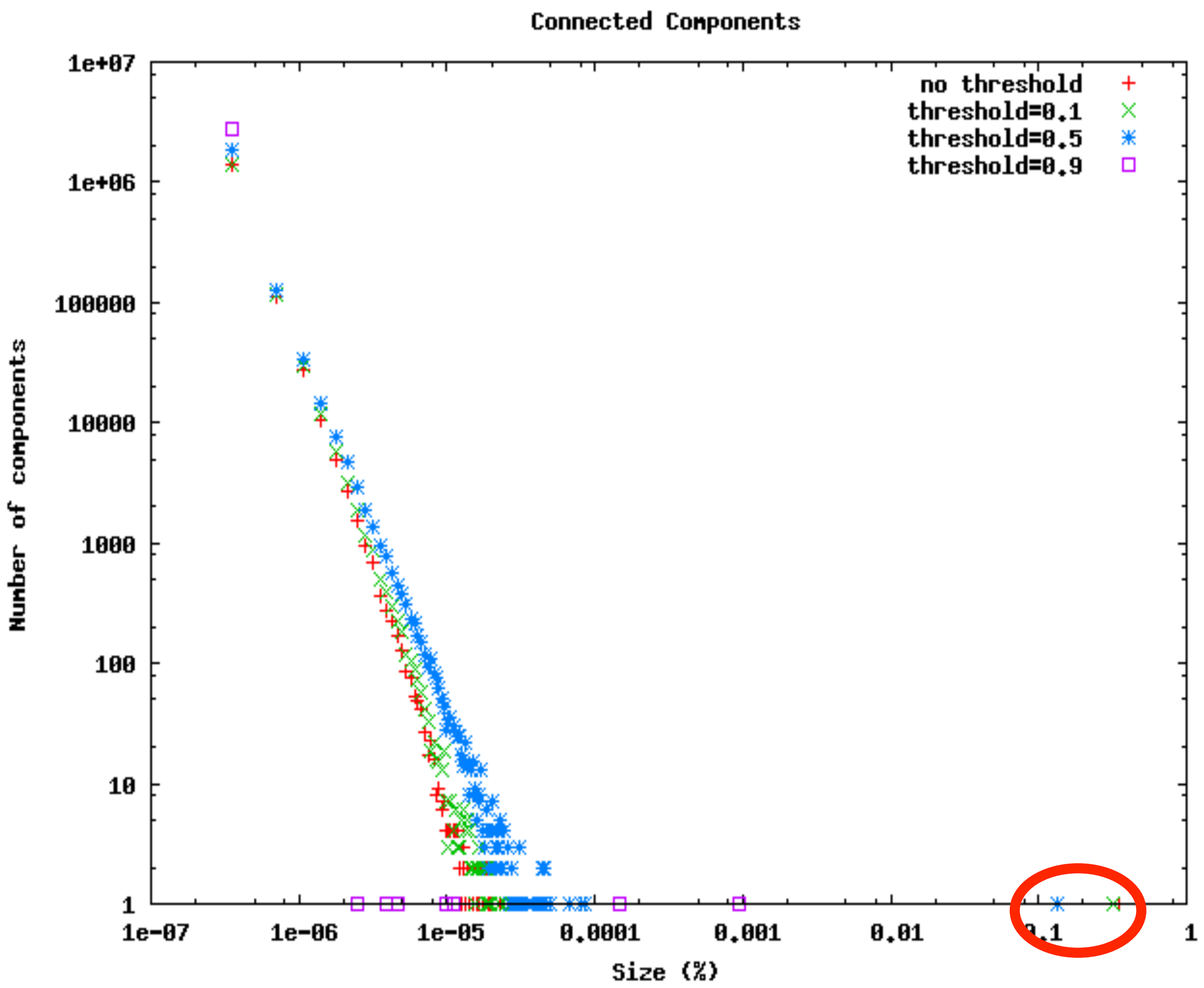
Anatomy of a Click Graph

- **We built graphs using logs with up to 50 millions queries**
  - For all the graphs we studied our findings are qualitatively the same (*scale-free network?*)
- **Here we present the results for the following graph**
  - 20M query occurrences
  - 2.8M distinct queries (nodes)
  - 5M distinct URLs
  - 361M edges

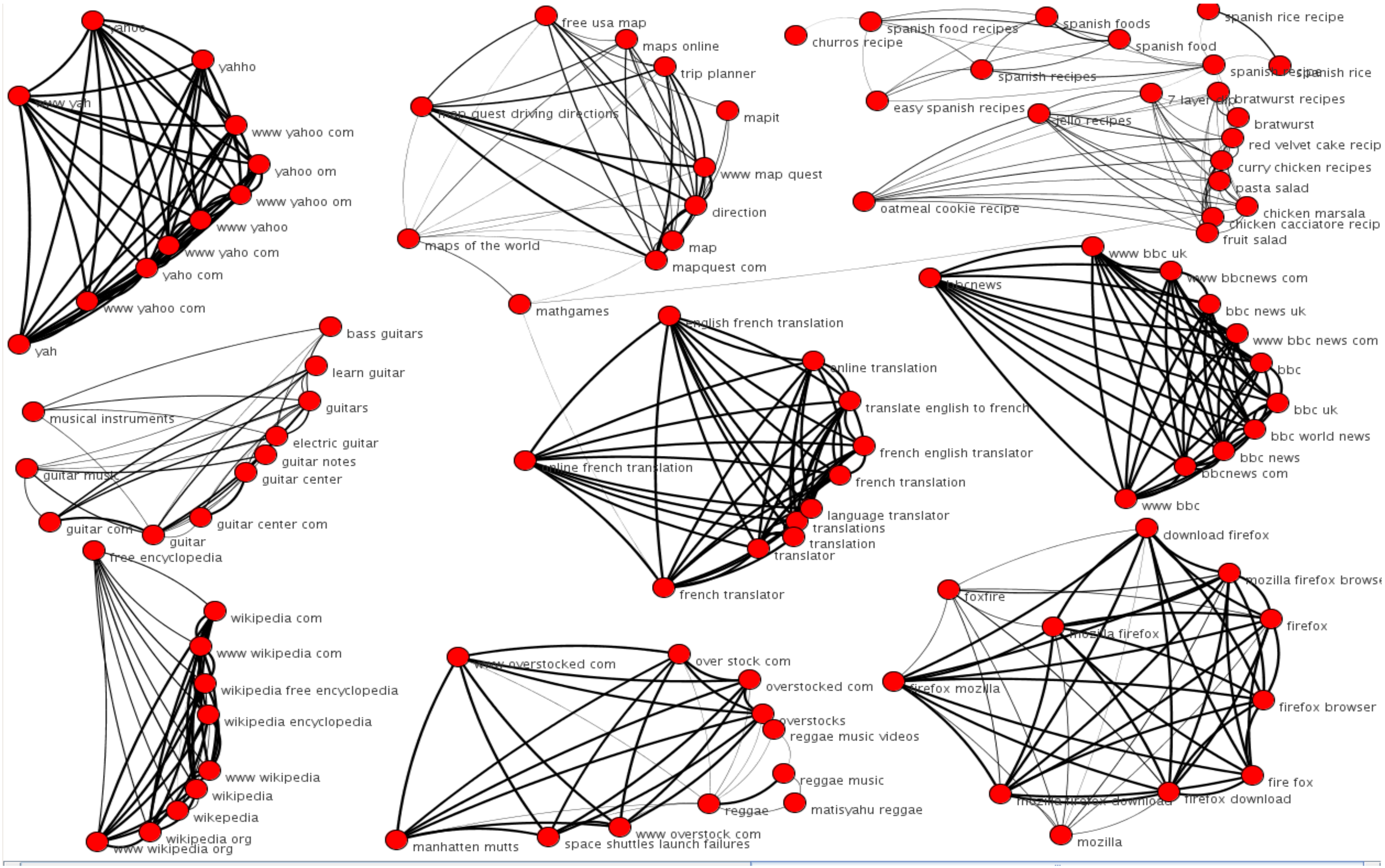
Click Distribution



Connected Components



Implicit Folksonomy?

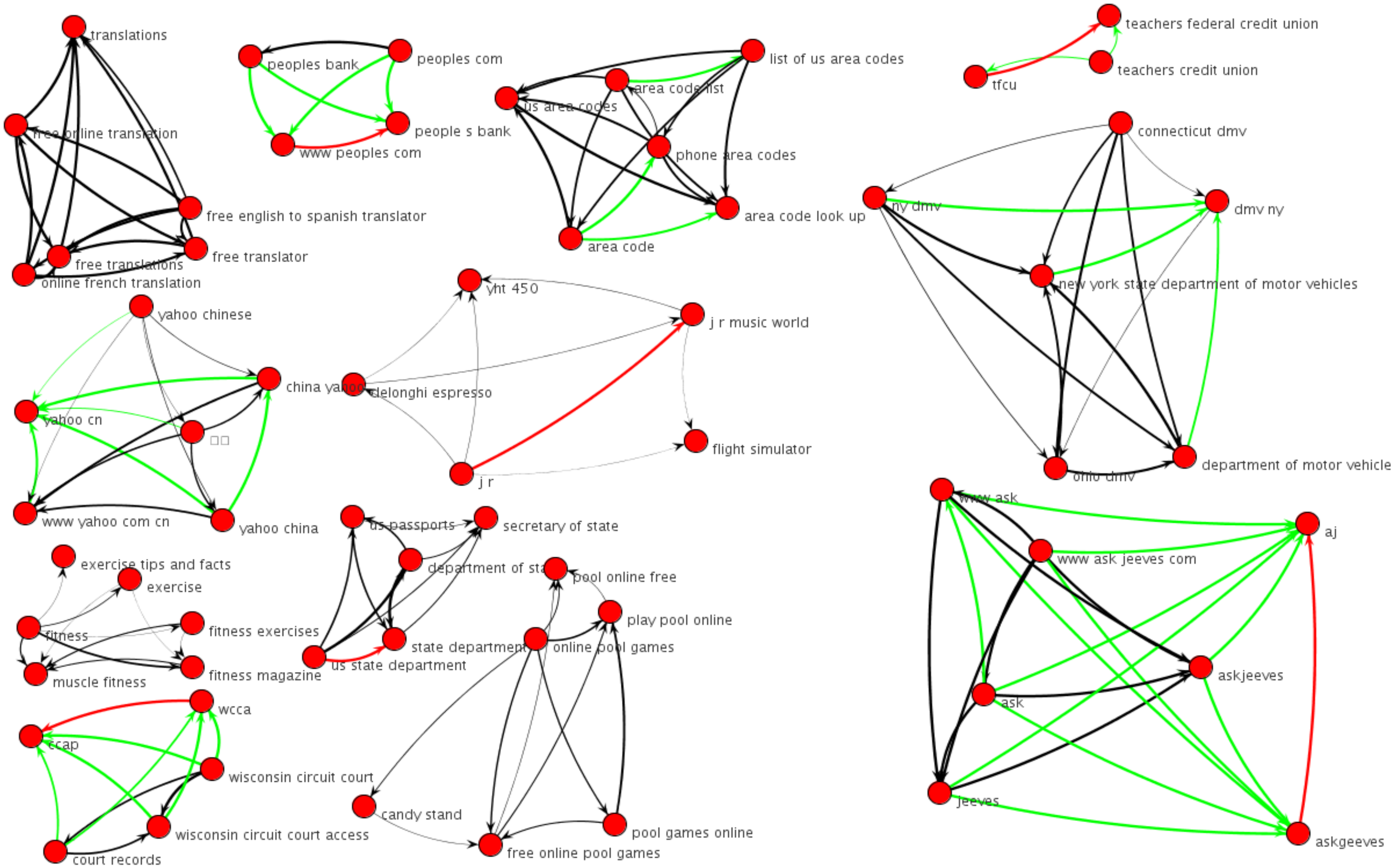


Set Relations and Graph Mining

- **Identical sets: equivalence**
- **Subsets: specificity**
  - directed edges
- **Non empty intersections (with threshold)**
  - degree of relation
- **Dual graph: URLs related by queries**
  - High degree: multi-topical URLs

Baeza-Yates & Tiberi
ACM KDD 2007

Implicit Knowledge? Webslang!



Evaluation: ODP Similarity

- **A simple measure of similarity among queries using ODP categories**
  - Define the similarity between two categories as the length of the longest shared path over the length of the longest path
  - Let $c\_1,..,c\_k$ and $c'\_1,..,c'\_k$ be the top k categories for two queries. Define the similarity (@k) between the two queries as $max\{sim(c\_i,c'\_j) \mid i,j=1,..,K\}$

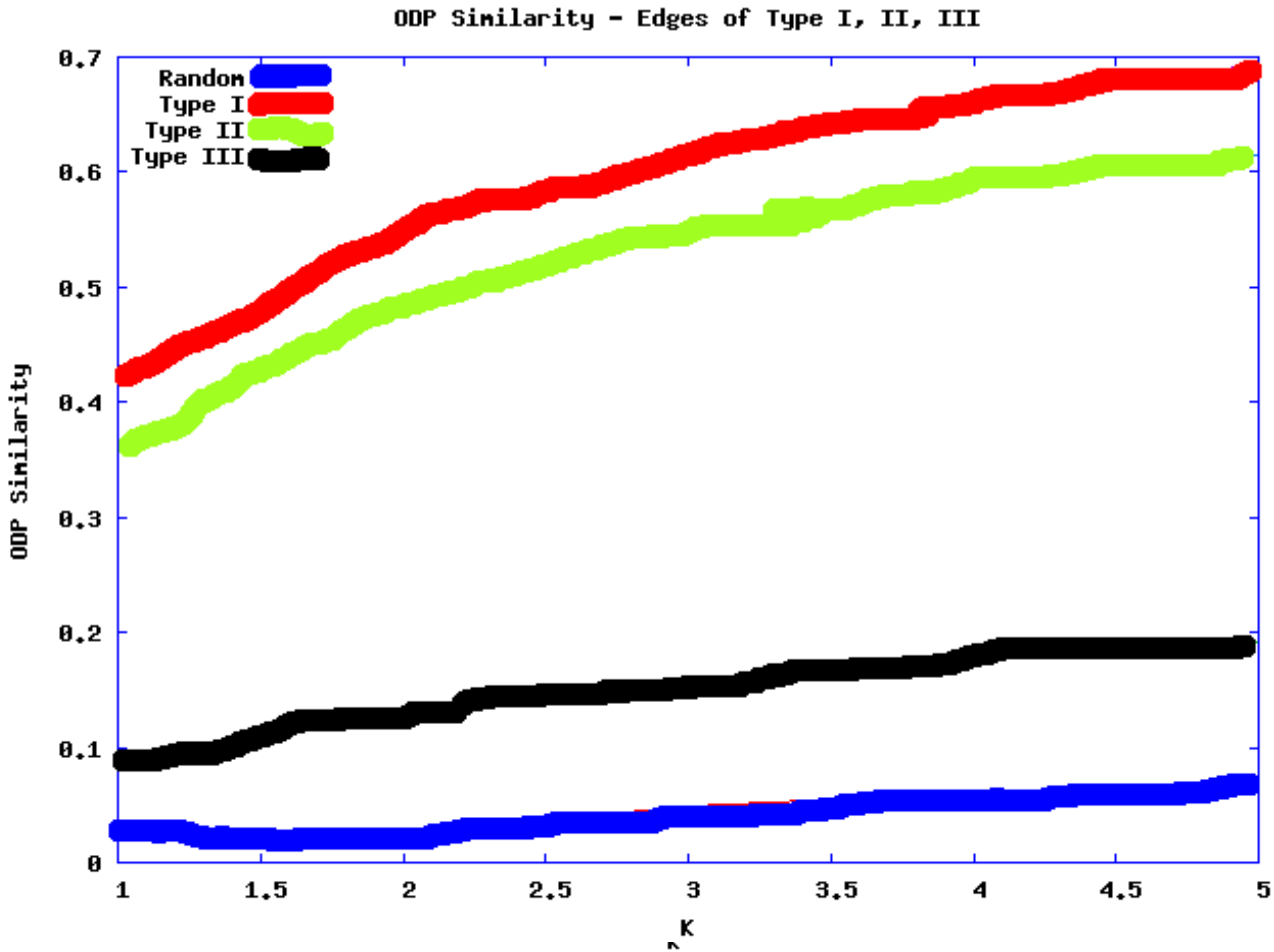
ODP Similarity

- **Suppose you submit the queries "*Spain*" and "*Barcelona*" to ODP.**
- **The first category matches you get are:**
  - Regional/ Europe/ Spain
  - Regional/ Europe/ Spain/ Autonomous Communities/ Catalonia/ Barcelona
- **Similarity @1 is 1/2 because the longest shared path is "Regional/ Europe/ Spain" and the length of the longest is 6**

Experimental Evaluation

- **We evaluated a 1000 thousand edges sample for each kind of relation**
- **We also evaluated a sample of random pairs of not adjacent queries (baseline)**
- **We studied the similarity as a function of *k* (the number of categories used)**

Experimental Evaluation



Open Issues

- **Implicit social network**
  - Any fundamental similarities?
- **How to evaluate with partial knowledge?**
  - Data volume amplifies the problem
- **User aggregation vs. personalization**
  - Optimize common tasks
  - Move away from privacy issues

Final Remarks

Epilogue

- **The Web is scientifically young**
- **The Web is intellectually diverse**
- **The technology mirrors the economic, legal and sociological reality**
- **Web Mining: large potential for many applications**
  - A fast prototyping platform is needed
- **Plenty of open problems**

Overall summary

- **Many open problems and challenges:**
  - Manage and integrate highly heterogeneous information:
  - Content, links, social links, tags, feedback, usage logs, wisdom of crowds, etc.
  - Model and benefit from evolution
  - Battle adversarial attempts and collusions

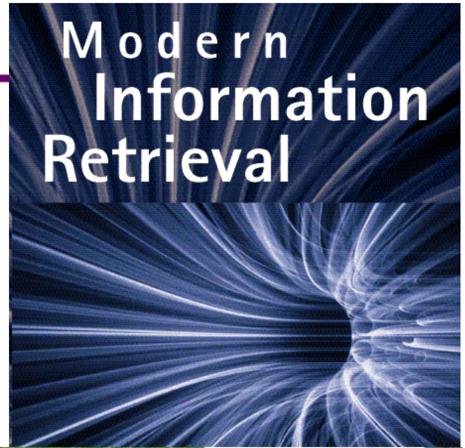
Special thanks

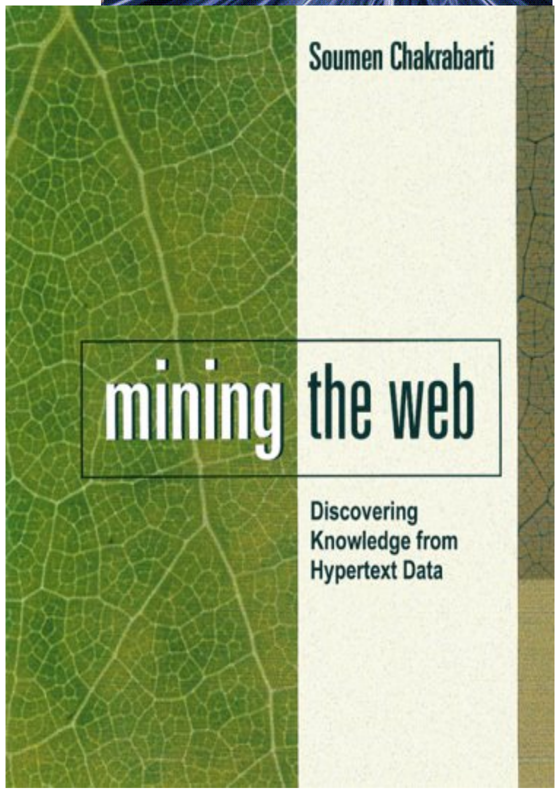
- Andrei Broder
- Carlos Castillo
- Barbara Poblete
- Alvaro Pereira
- Prabhakar Raghavan
- Alessandro Tiberi

Bibliography – General

- **Modern Information Retrieval**

  by R. Baeza-Yates & B. Ribeiro-Neto, Addison-Wesley, 1999. Second edition to appear in 2010.
- **Managing Gigabytes: Compressing and Indexing Documents and Images by I.H. Witten, A. Moffat, and T.C. Bell. Morgan Kaufmann, San Francisco, second edition, 1999.**
- **Mining the Web: Analysis of Hypertext and Semi Structured Data**
  by Soumen Chakrabarti. Morgan Kaufmann; August 15, 2002.
- **The Anatomy of a Large-scale Hypertextual Web Search Engine**
  by S. Brin and L. Page. 7th International WWW Conference, Brisbane, Australia; April 1998.
- **Websites:**
  - http://www.searchenginewatch.com/
  - http://www.searchengineshowdown.com/





an introduction to web mining

Ricardo Baeza-Yates
Aristides Gionis
Yahoo! Research, Barcelona
IJCAI 2011, Barcelona

query-log mining

query-log mining

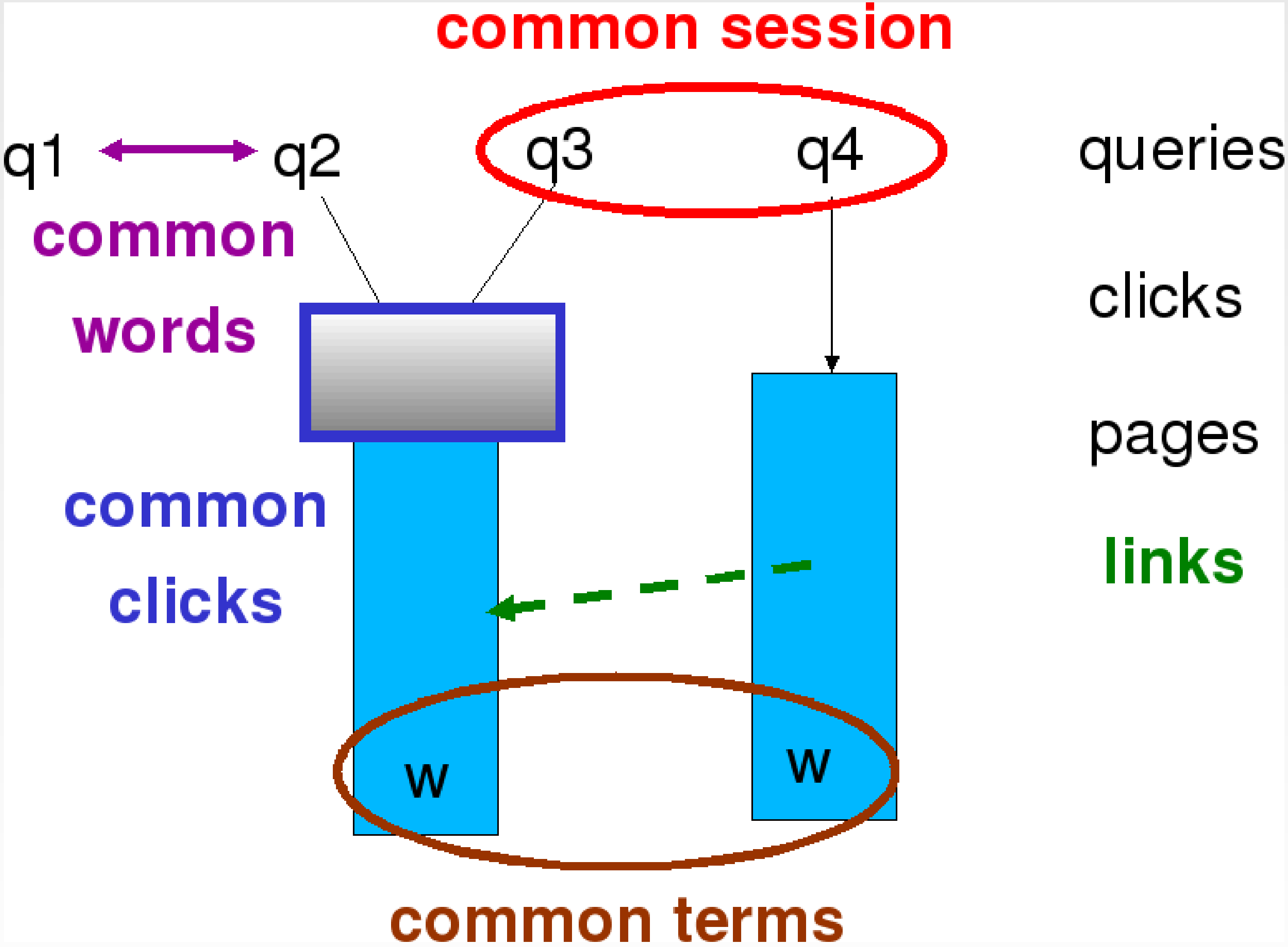
- search engines collect a large amount of query logs
- lots of interesting information
  - analyzing users' behavior
  - creating user profiles
  - personalization
  - creating knowledge bases and folksonomies
  - finding similar concepts
  - building systems for query suggestions and recommendations
  - using statistics for improving systems' performance
  - etc.

query-log mining

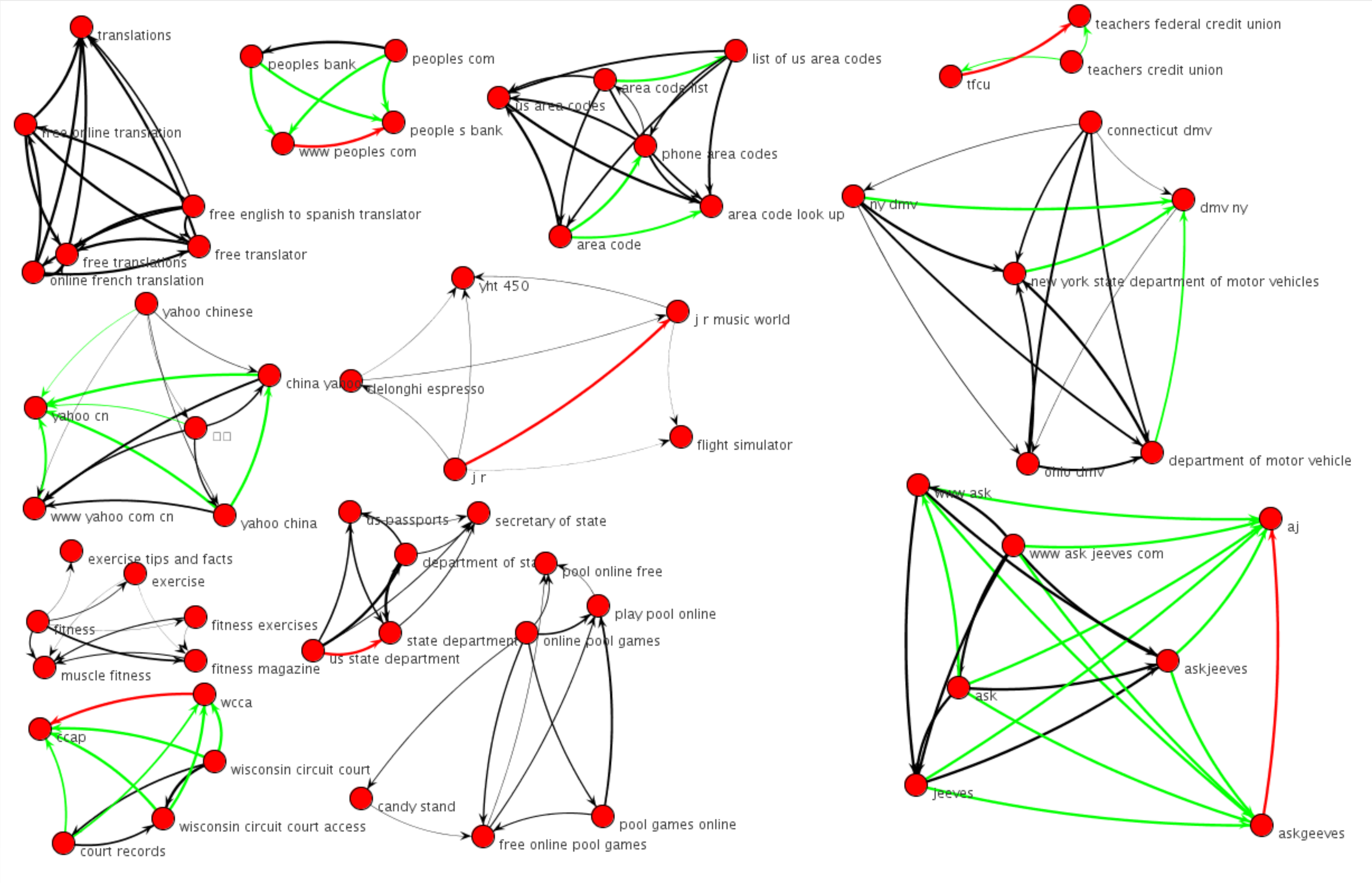
- query-log graphs
- query recommendations

query graphs

different ways to relate queries

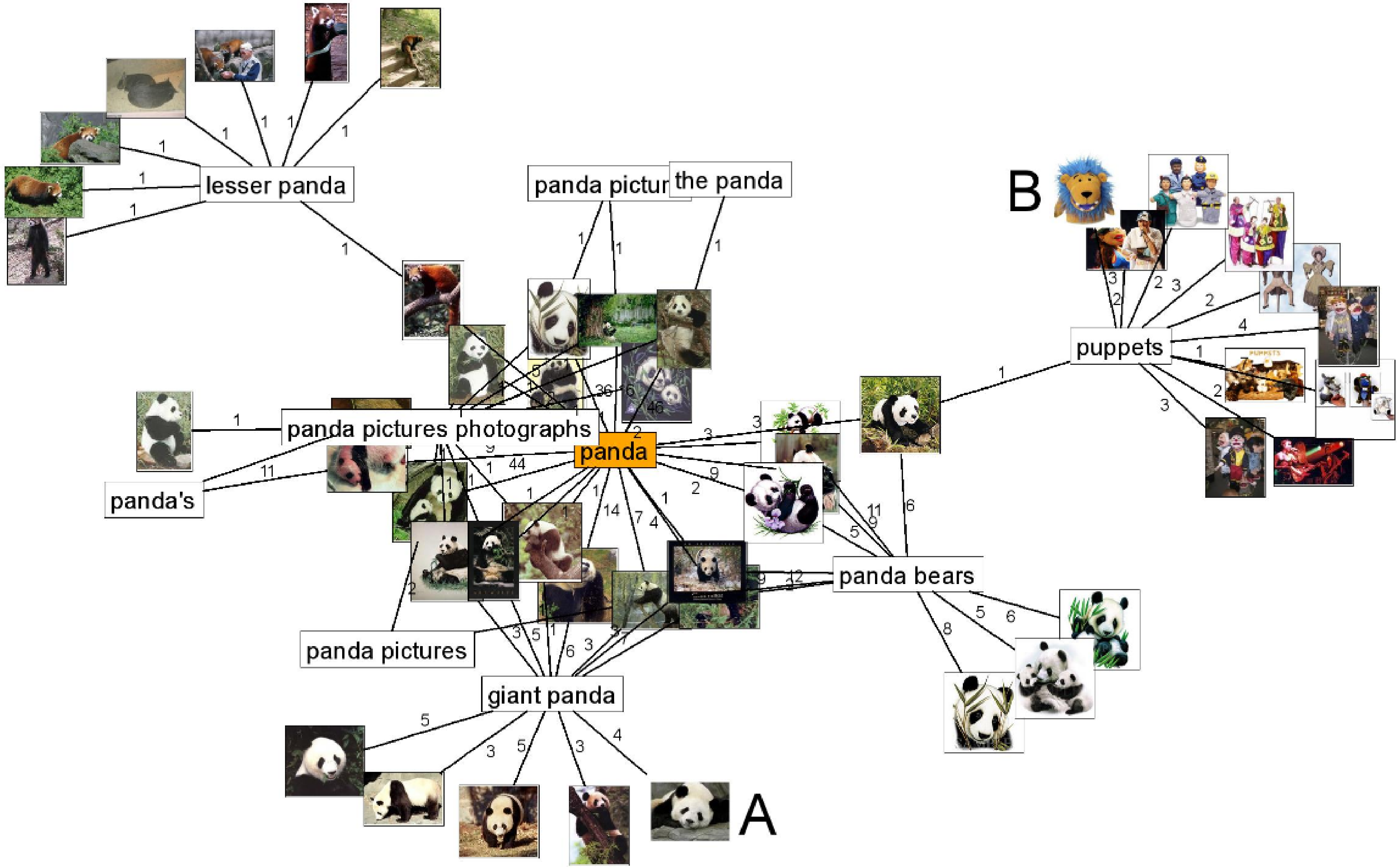


the click graph – implicit knowledge – webslang



[Baeza-Yates and Tiberi, 2007]

the click graph



[Craswell and Szummer, 2007]

applications of the click graph

[Craswell and Szummer, 2007]

- query-to-document search
- query-to-query suggestion
- document-to-query annotation
- document-to-document relevance feedback

the query-flow graph

the query-flow graph

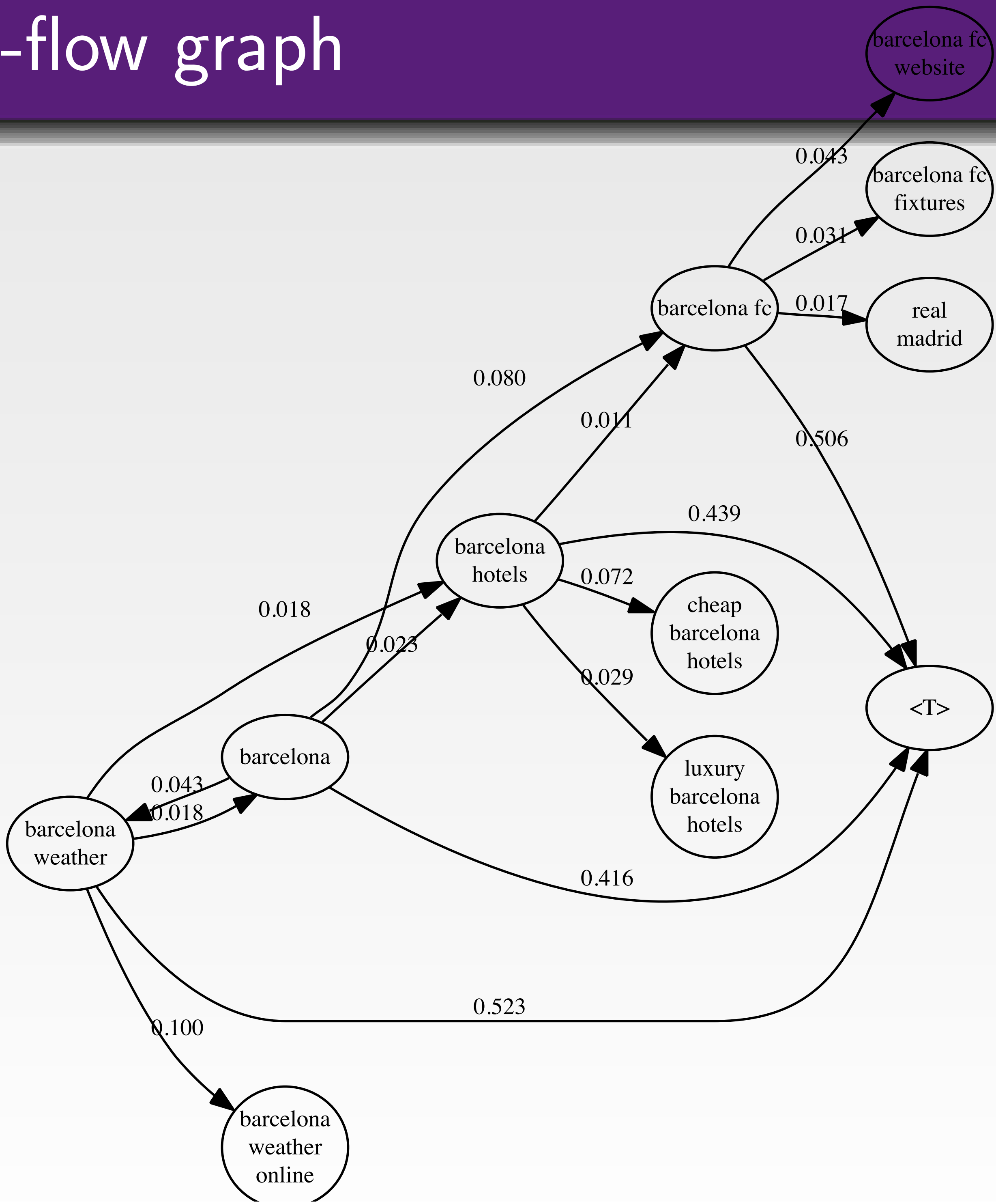
[Boldi et al., 2008]

- take into account temporal information
- captures the "flow" of how users submit queries
- definition:
  - nodes $V = Q \cup \{s, t\}$ the distinct set of queries Q, plus a starting state s and a terminal state t
  - edges $E \subseteq V \times V$
  - weights $w(q, q')$ representing the probability that q and q' are part of the same chain

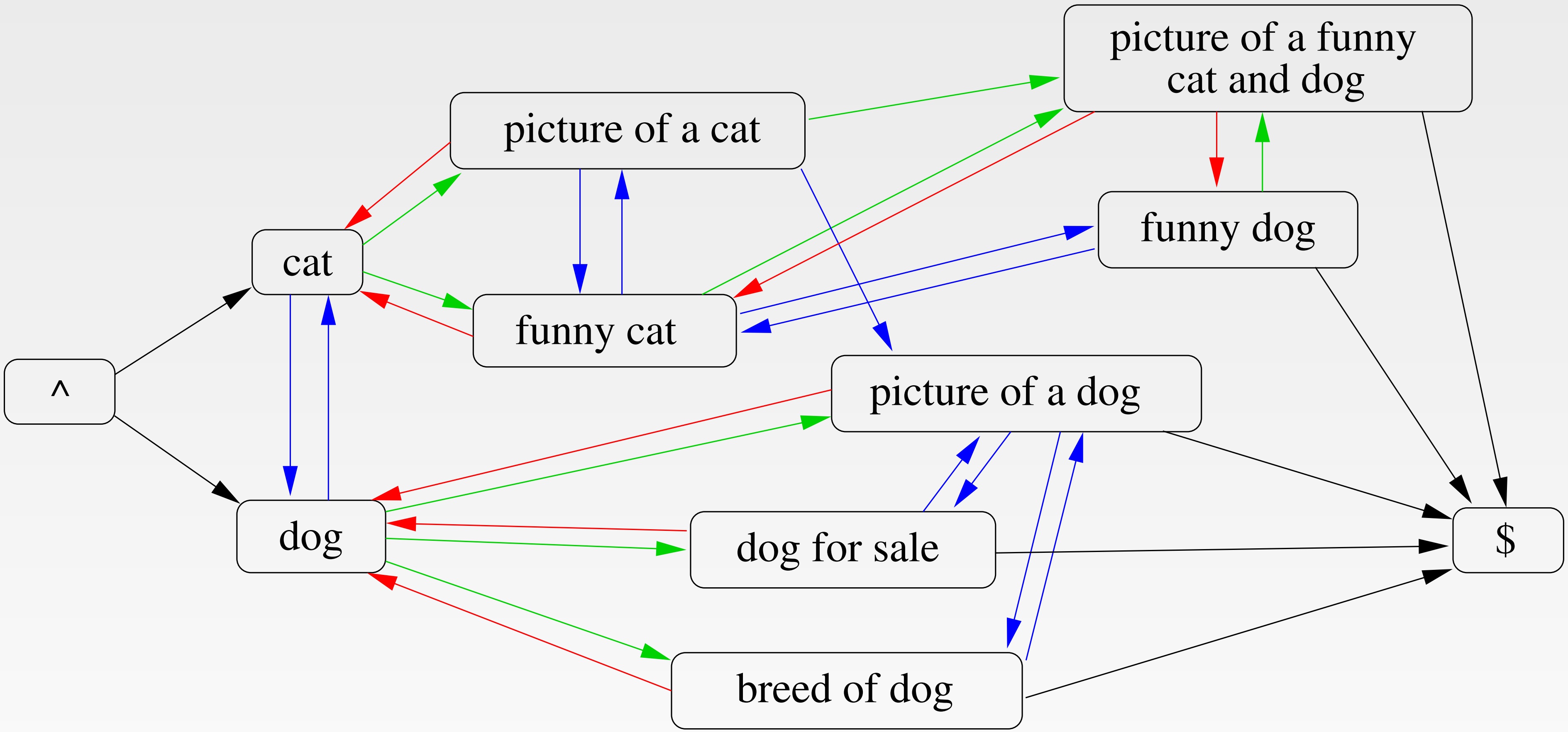
building the query-flow graph

- an edge (q, q') if q and q' are consequetive in at least one session
- weights $w(q, q')$ learned by machine learning
- features used
  - textual features: cosine similarity, Jaccard coefficient, size of intersection, etc.
  - session features: the number of sessions, the average session length, the average number of clicks in the sessions, the average position of the queries in the sessions, etc. and
  - time-related features: average time difference, etc.

query-flow graph



query-flow graph



application: session segmentation

- user submits queries by switching contexts:
  work, go to a movie, buy a product, work
- problem: given a long session of queries
  find a segmentation into logical sessions
- re-order the query sequence in order to maximize likelihood
- solved as a traveling salesman problem

query recommendations

the general theme

- given an input query q
- identify similar queries q
- rank them and present them to the user
- all graphs we studied can be used for both tasks: similarity and ranking

recommendations using the query-flow graph

recommendations using the query-flow graph

[Boldi et al., 2008]

- perform a random walk on the query-flow graph
- teleportation to the submitted query
- teleportation to previous queries to take into account the user history
- normalize PageRank score to unbiasing for very popular queries

example : `apple`

Max. weight	s_q	$\hat{s}_q$	$\bar{s}_q$
t	t	apple	apple
apple ipod	apple	apple fruit	apple ipod
apple store	apple ipod	apple ipod	apple trailers
apple trailers	apple store	apple belgium	apple store
amazon	apple trailers	eating apple	apple mac
apple mac	google	apple.nl	apple fruit
itunes	amazon	apple monitor	apple usa
pc world	argos	apple usa	apple ipod nano
argos	itunes	apple jobs	apple.com/ipod...

example : jeep

Max. weight	s_q	$\hat{s}_q$	$\bar{s}_q$
t	t	jeep	jeep
jeep cherokee	jeep	jeep trails	jeep cherokee
jeep grand ...	jeep cherokee	jeep kinderk...	jeep trails
jeep wrangler	jeep grand ...	jeep compass	jeep compass
land rover	bmw	jeep cherokee	jeep kinderkled...
landrover	jeep wrangler	swain and jon...	jeep grand ...
ebay	land rover	jeep bag	jeep wrangler
chrysler	landrover	country living ...	chryslar
bmw	chrysler	buy range rov...	jeepcj7
nissan	google	craviotto snare	buses to Knowl...

example : `banana` → `apple`

`banana` → `apple`	`banana`
banana	banana
apple	eating bugs
usb no	banana holiday
banana cs	opening a banana
giant chocolate bar	banana shoe
where is the seed in anut	fruit banana
banana shoe	recipe 22 feb 08
fruit banana	banana jules oliver
banana cloths	banana cs
eating bugs	banana cloths

example : beatles → apple

beatles → apple	beatles
beatles	beatles
apple	scarring
apple ipod	paul mcartney
scarring	yarns from ireland
srg peppers artwork	statutory instrument A55
ill get you	silver beatles tribute band
bashles	beatles mp3
dundee folk songs	GHOST'S
the beatles love album	ill get you
place lyrics beatles	fugees triger finger remix

recommendations as shortcuts to QFG

QFG-based recommendations

[Anagnostopoulos et al., 2010]

- model user behavior as a random walk on QFG
- a user starts at query q_0 and follows a path p of reformulations on QFG before terminating
- consider weight function $w(q)$
  - e.g., query quality, user satisfaction, monetization, etc.
- utility function $U(p)$

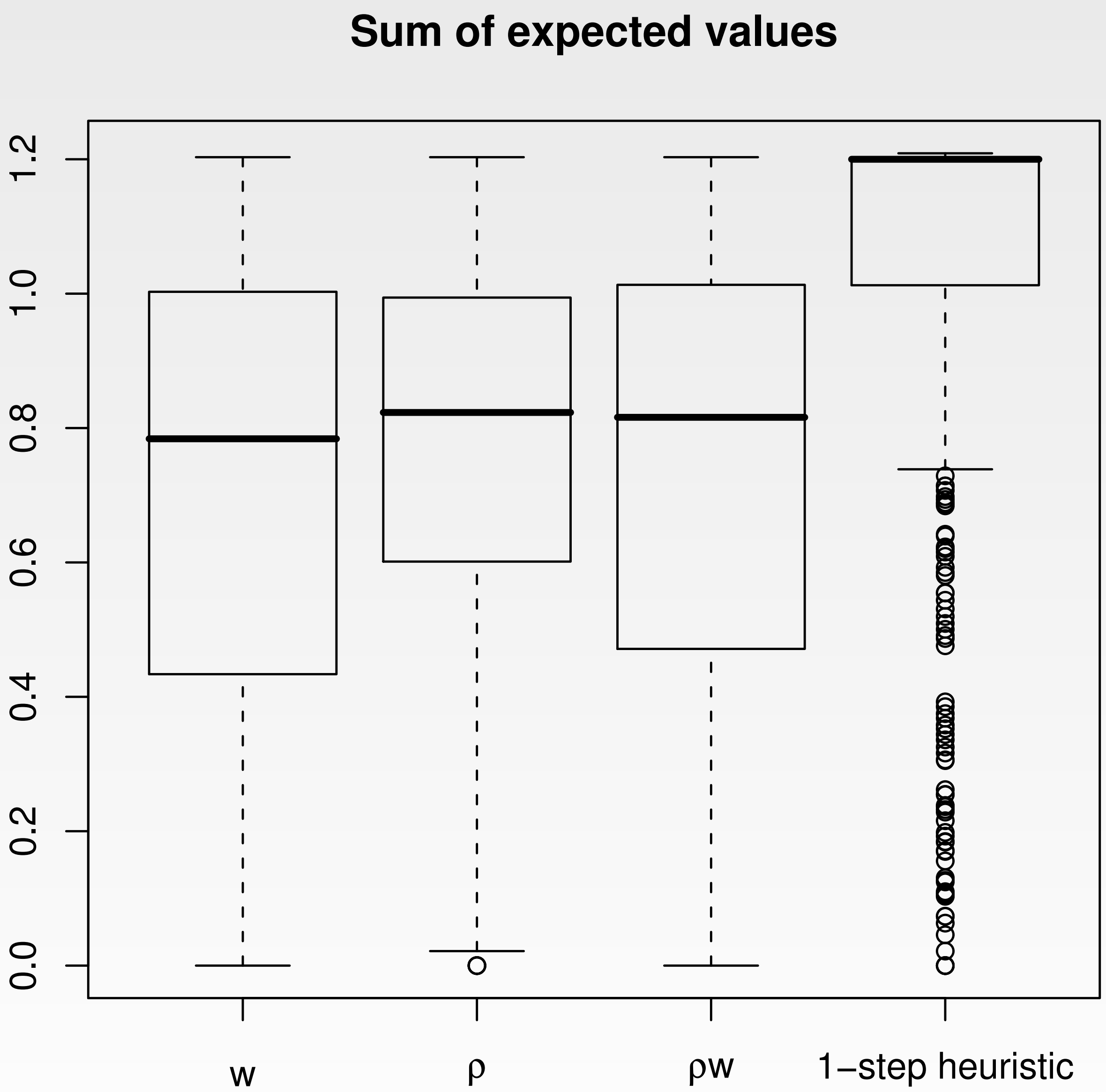
$$U(p) = \sum_{q \in p} w(q), \qquad \text{or} \qquad U(p) = w(q_{k-1}),$$

where $p = \langle q_0 \dots q_{k-1} T \rangle$

QFG-based recommendations

- random walk on QFG is modeled by stochastic matrix P
- recommendations R modify P to $P' = \alpha P + (1 - \alpha)R$
- problem definition: for each query q find k recommendations $R(q)$ in order to maximize expected utility achieved on the modified graph P'
- a general problem formulation for suggesting shortcuts (web graph, social networks, etc.)

utility


Sum of expected values
0.0
0.2
0.4
0.6
0.8
1.0
1.2
w
ρ
ρw
1−step heuristic

QFG projections for diverse recommendations

diverse recommendations

[Bordino et al., 2010]

- we want not only relevant and high-quality recommendations, but also a diverse set
- we want recommendations that take to different "directions" in the QFG
- need notions of distance of queries in the QFG
- use spectral embeddings
    - project a graph in a low dimensional space, so that embedding minimizes total edge distortion
- finding diverse recommendations reduces to a geometric problem

example: time

Spectral projection on 2-hop neighborhood

time	time magazine	new york times	time zone	world time	what time is it	time warner	time warner cable
time magazine		0.9953	0.0162	0.1422	0.1049	-0.6071	-0.6056
new york times	0.9953		-0.0051	0.1248	0.0893	-0.6478	-0.6462
time zone	0.0162	-0.0051		0.9903	0.9891	-0.5234	-0.5254
world time	0.1422	0.1248	0.9903		0.9970	-0.6263	-0.6282
what time is it	0.1049	0.0893	0.9891	0.9970		-0.6244	-0.6263
time warner	-0.6071	-0.6478	-0.5234	-0.6263	-0.6244		0.9999
time warner cable	-0.6056	-0.6462	-0.5254	-0.6282	-0.6263	0.9999	

properties of web graphs

properties of graphs at different levels

different families of web graphs arise from different phenomena

are there any typical patterns?

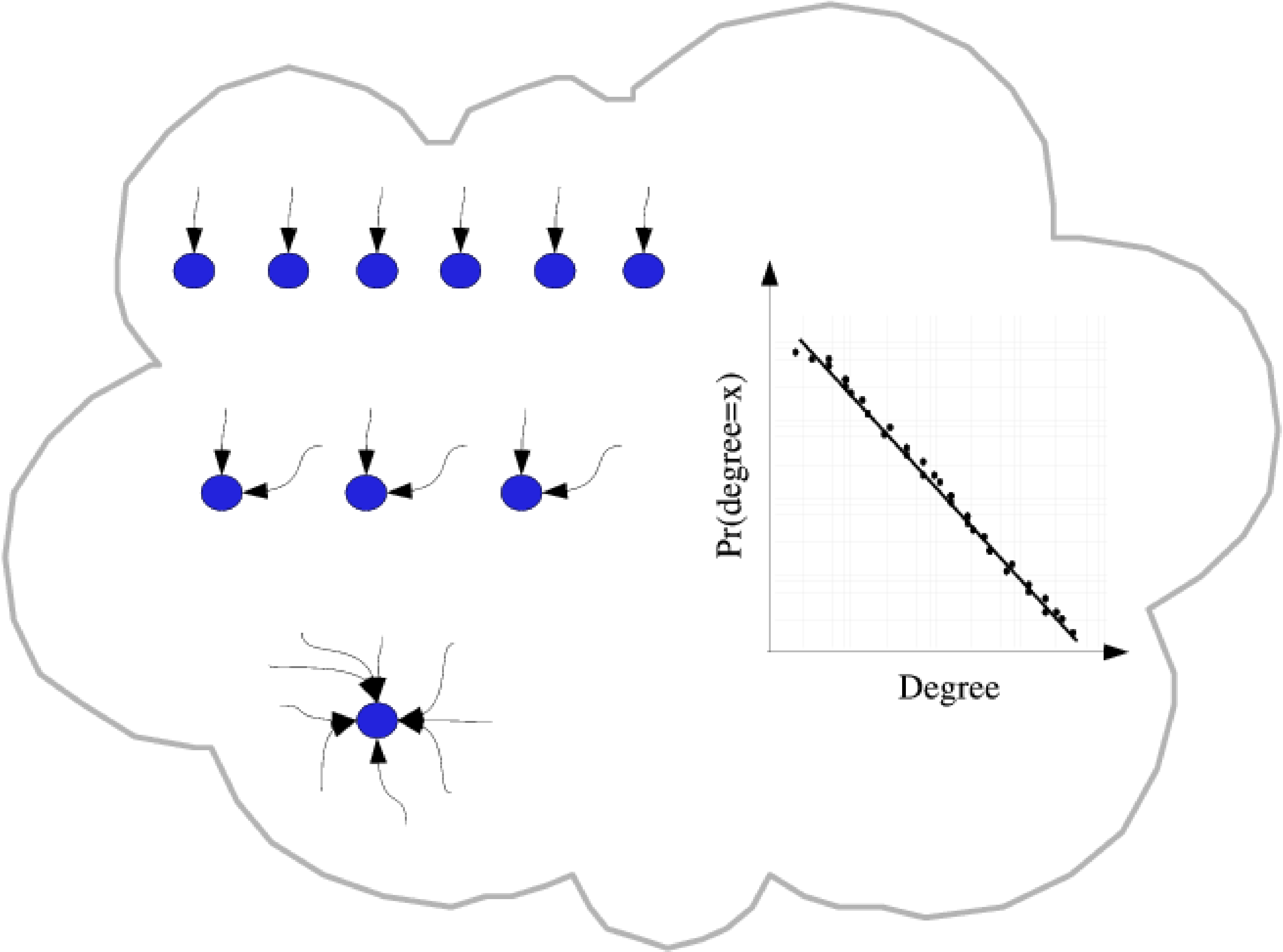
at which level should we look for commonalities?

- degree distribution — microscopic
- communities — mesoscopic
- small diameters — macroscopic

degree distribution

- consider C_k the number of vertices u with degree $d(u) = k$.
  then
  $$C_k = ck^{-\gamma},$$
  with $\gamma > 1$, or
  $$\ln C_k = \ln c - \gamma \ln k$$
- so, plotting $\ln C_k$ versus $\ln k$ gives a straight line with slope $-\gamma$
- *heavy-tail distribution*: there is a non-negligible fraction of nodes that has very high degree (hubs)

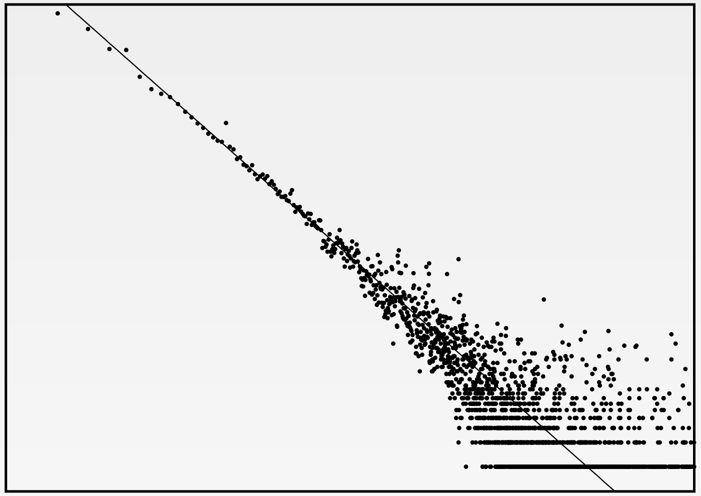
degree distribution



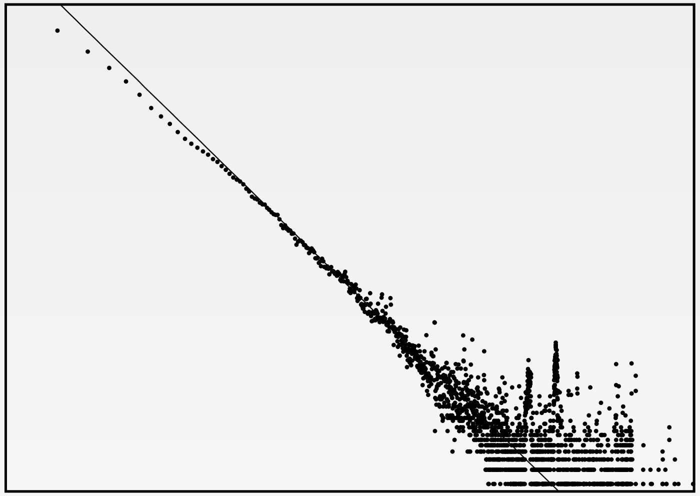
degree distribution

indegree distributions of web graphs within national domains

Greece



Spain

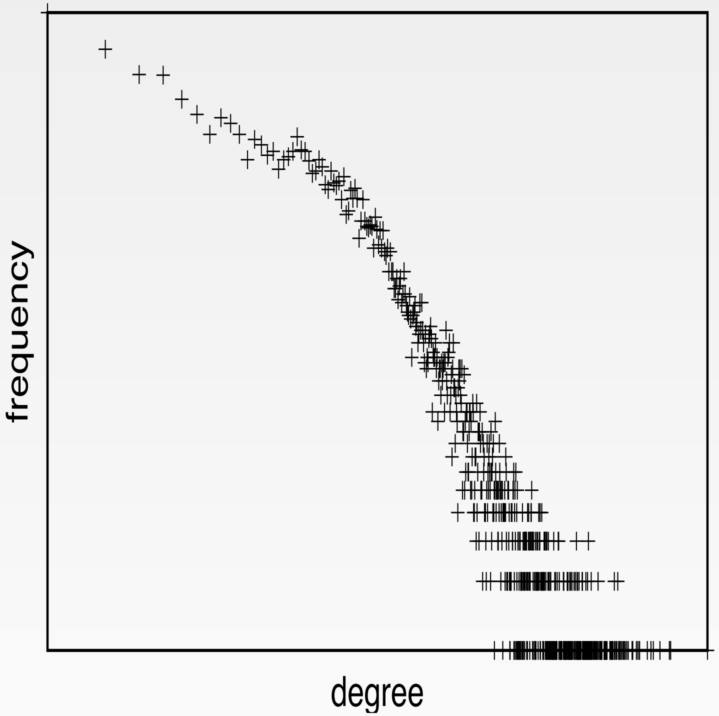


[Baeza-Yates and Castillo, 2005]

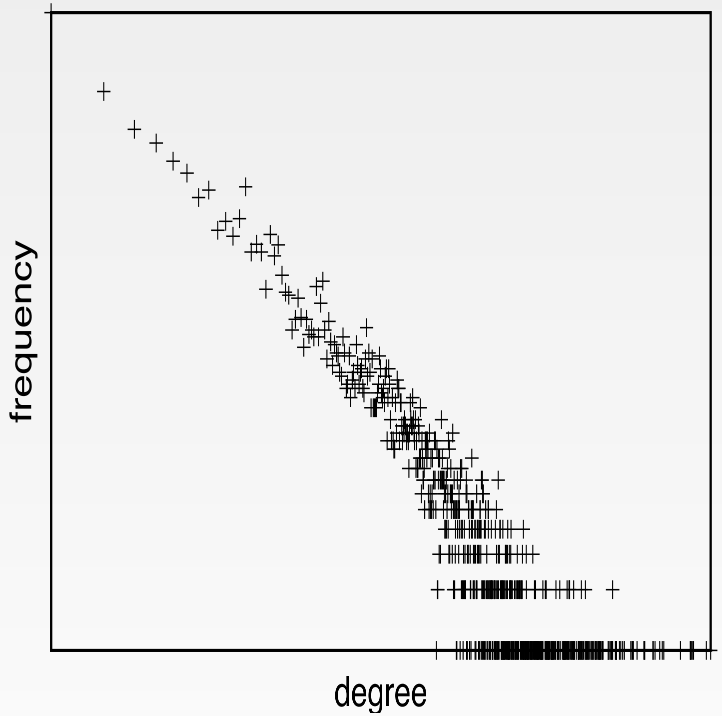
degree distribution

...and more "straight" lines

in-degrees of UK hostgraph



out-degrees of UK hostgraph



community structure

- intuitively a subset of vertices that are more connected to each other than to other vertices in the graph
- a proposed measure is *clustering coefficient*

$$C_1 = \frac{3 \times \text{ number of triangles in the network}}{\text{number of connected triples of vertices}}$$

- captures "transitivity of clustering"
- if u is connected to v and
  v is connected to w, it is also likely that
  u is connected to w

community structure

- alternative definition
- *local clustering coefficient*

$$C_i = \frac{\text{number of triangles connected to vertex } i}{\text{number of triples centered at vertex } i}$$

- *global clustering coefficient*

$$C_2 = \frac{1}{n} \sum_i C_i$$

- community structure is captured by large values of clustering coefficient

small diameter

diameter of many real graphs is small (e.g., $D = 6$ is famous)

proposed measures

- *hop-plots:* plot of $|N_h(u)|$, the number of neighbors of u at distance at most h, as a function of h
  [M. Faloutsos, 1999] conjectured that it grows exponentially and considered *hop exponent*
- *effective diameter:* upper bound of the shortest path of 90% of the pairs of vertices
- *average diameter:* average of the shortest paths over all pairs of vertices
- *characteristic path length:* median of the shortest paths over all pairs of vertices

measurements on real graphs

graph	n	m	α	C_1	C_2	ℓ
film actors	449 913	25 516 482	2.3	0.20	0.78	3.48
internet	10 697	31 992	2.5	0.03	0.39	3.31
protein interactions	2 115	2 240	2.4	0.07	0.07	6.80

[Newman, 2003]

random graphs

- Erdös-Rényi random graphs have been used as point of reference
- the basic random graph model:
- n : the number of vertices
- $0 \leq p \leq 1$
- for each pair (u, v), independently generate the edge (u, v) with probability p
- $G_{n,p}$ a family of graphs, in which a graph with m edges appears with probability $p^m(1-p)^{\binom{n}{2}-m}$
- $z = np$

random graphs

- do they satisfy properties similar with those of real graphs?
- typical distance $d = \frac{\ln n}{\ln z}$ ✔
  - number of vertices at distance l is $\simeq z^l$, set $z^d \simeq n$
- Poisson degree distribution ✘

$$p_k = \binom{n}{k} p^k (1-p)^{n-k} \simeq \frac{z^k e^{-z}}{k}$$

  - highly concentrated around the mean ($z = np$)
  - probability of very high degree nodes is exponentially small
- clustering coefficient $C = p$ ✘
  - probability that two neighbors of a vertex are connected is independent of the local structure

other properties

- degree correlations
- distribution of size of connected components
- resilience
- eigenvalues
- distribution of motifs

properties of evolving graphs

- [Leskovec et al., 2005] discovered two interesting and counter-intuitive phenomena
- densification power law

$$|E_t| \propto |V_t|^{\alpha} \qquad 1 \leq \alpha \leq 2$$

- diameter is shrinking

algorithmic tools

efficiency considerations

- data in the web and social-media are typically of extremely large scale (easily reach to billions)
- how to locate similar objects fast?
- how to cluster objects?
- how to compute simple statistics?

hashing and sketching

- hashing: hash objects in such a way that similar objects have larger probability of mapped to the same value than non-similar objects
- sketching: create sketches that summarize the data and allow to estimate simple statistics with small space
- probabilistic/approximate methods

locality sensitive hashing

a family $\mathcal{H}$ is called (R, cR, p_1, p_2)-sensitive if for any two objects p and q

- if $d(p,q) \leq R$, then $\Pr_{\mathcal{H}}[h(p) = h(q)] \geq p_1$
- if $d(p,q) \geq cR$, then $\Pr_{\mathcal{H}}[h(p) = h(q)] \leq p_2$

interesting case when $p_1 > p_2$

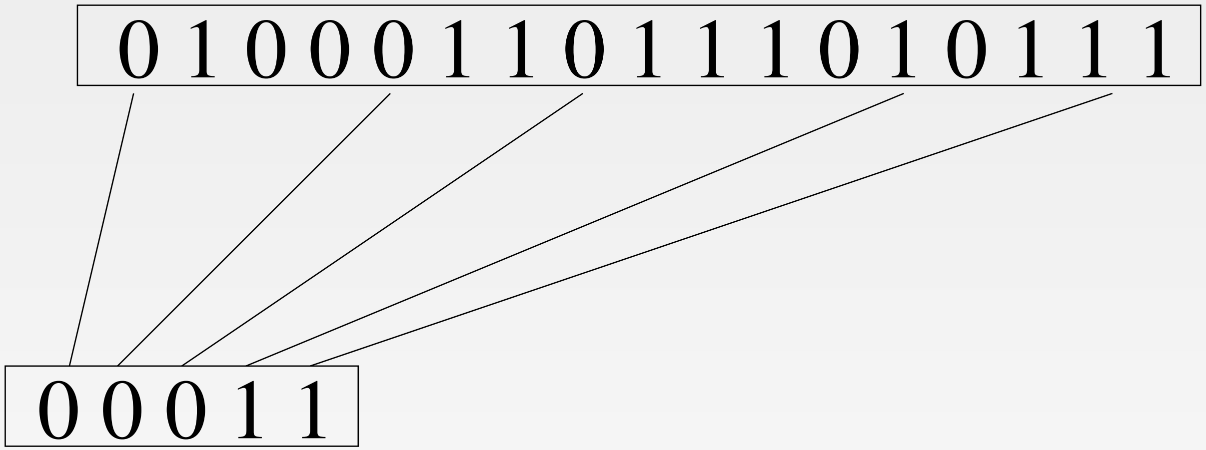
locality sensitive hashing: example

- objects in a Hamming space $\{0,1\}^d$ – binary vectors
- $\mathcal{H} : \{0,1\}^d \to \{0,1\}$ sample the i bit:
- $\mathcal{H} = \{h(x) = x_i \mid i = 1, \ldots, d\}$
- for two vectors x and y with distance r, it is $\Pr_{\mathcal{H}}[h(x) = h(y)] = 1 - \frac{r}{d}$
- thus $p_1 = 1 - \frac{R}{d}$ and $p_2 = 1 - \frac{cR}{d}$
- gap between p_1 and p_2 too small
- probability amplification

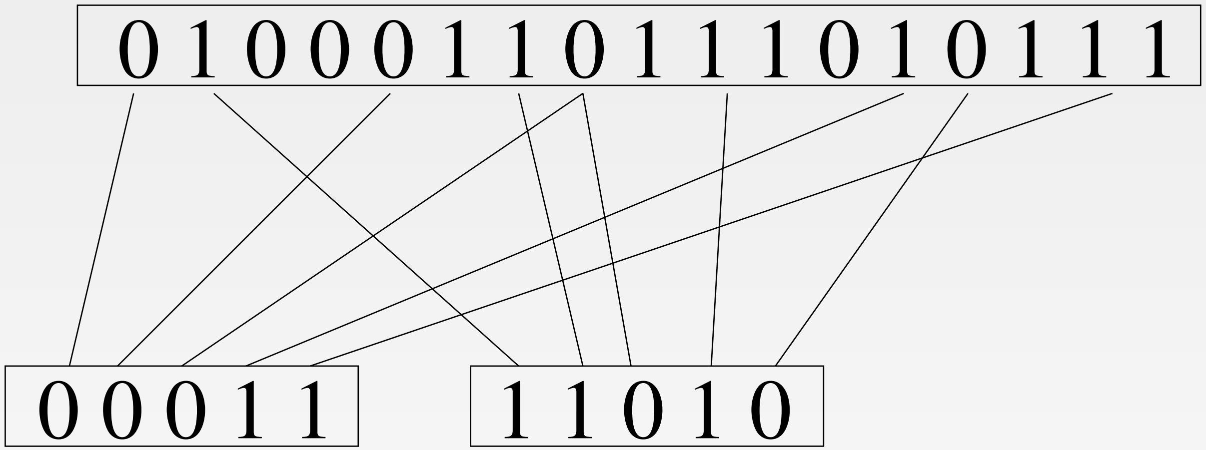
locality sensitive hashing: Hamming distance

$\boxed{0\ 1\ 0\ 0\ 0\ 1\ 1\ 0\ 1\ 1\ 1\ 0\ 1\ 0\ 1\ 1\ 1}$

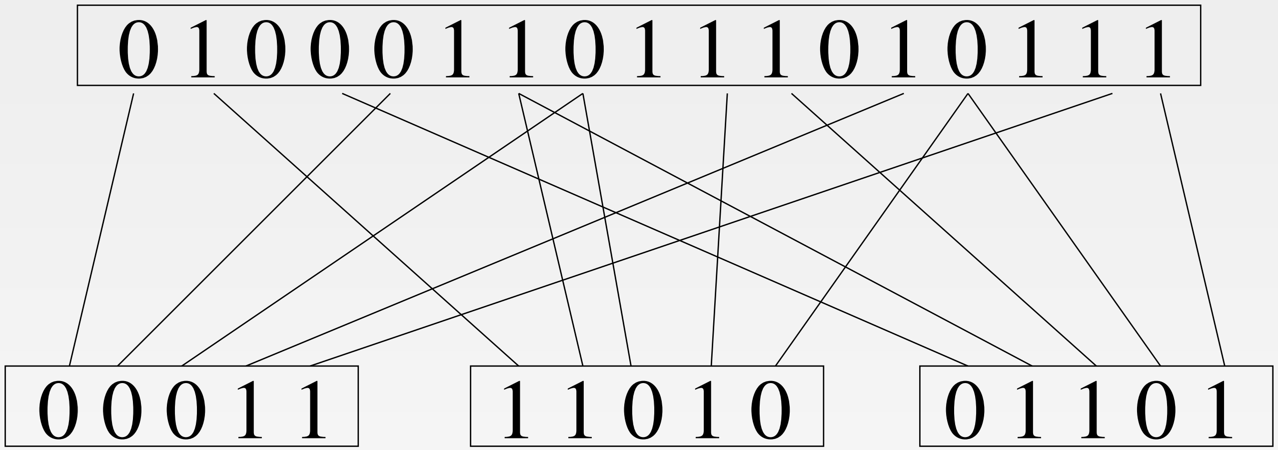
locality sensitive hashing: Hamming distance



locality sensitive hashing: Hamming distance



locality sensitive hashing: Hamming distance

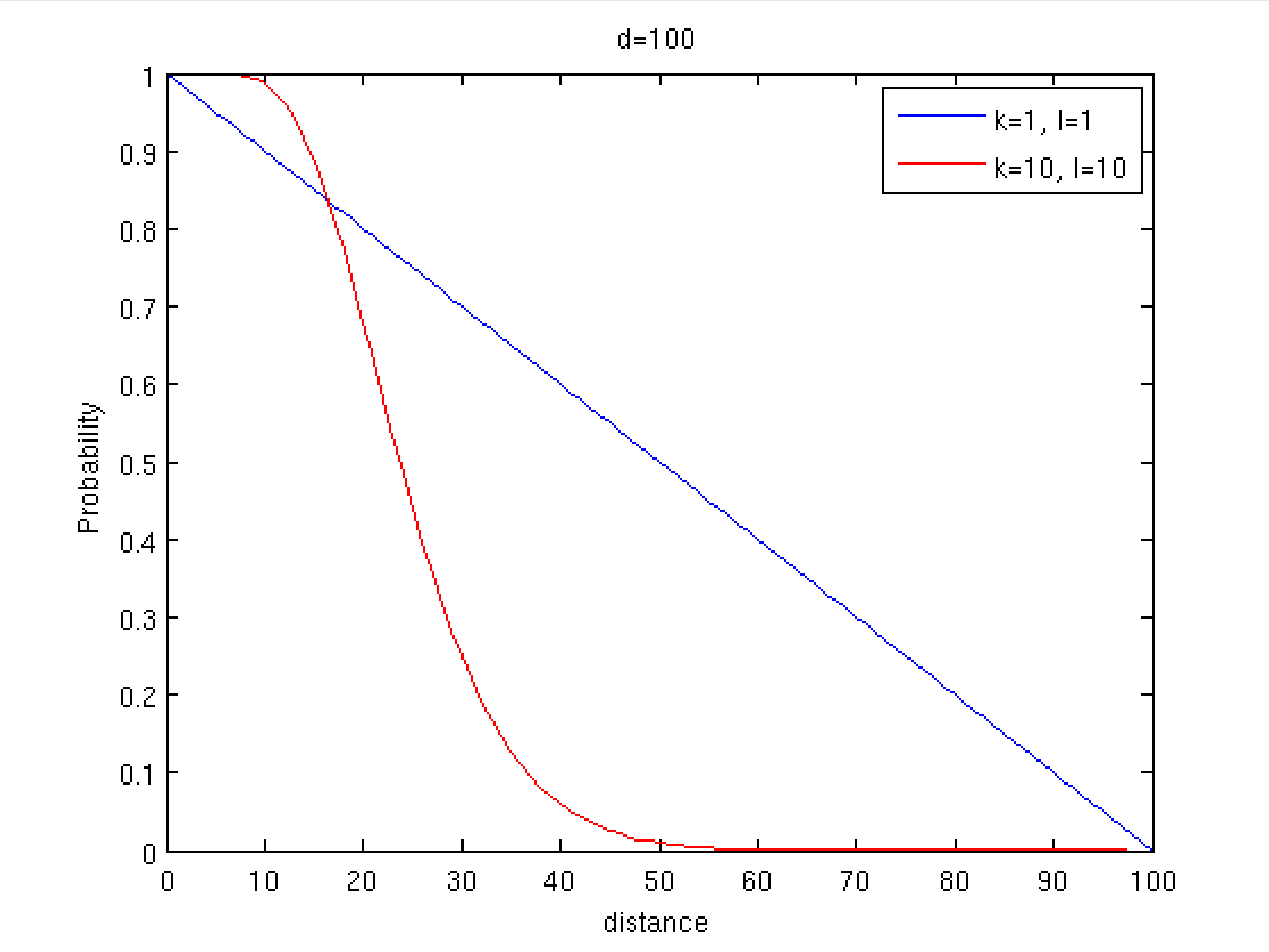


locality sensitive hashing: Hamming distance

Probability of collision

$$\Pr[h(x) = h(y)] = 1 - (1 - (1 - \frac{r}{d})^k)^l$$

locality sensitive hashing: Hamming distance



homework

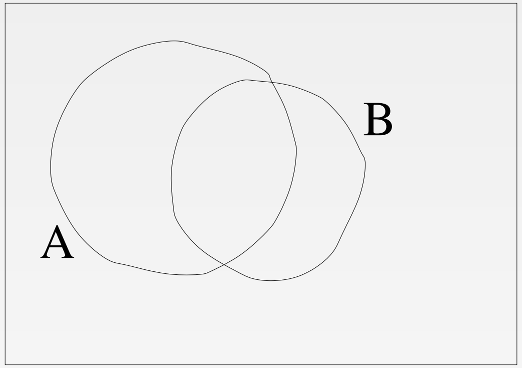
how to apply the locality sensitive hashing for vectors of integers, not just binary vectors?

vectors $\mathbf{x} = \{x_1, \ldots, x_d\}$

L_1 distance $||\mathbf{x} - \mathbf{y}||_1 = \sum_{i=1}^{d} |x_i - y_i|$

Jaccard coefficient

- for two sets $A, B \subseteq U$ define $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$
- measure of similarity of the sets



- can we design a locality sensitive hashing family for Jaccard?

min-wise independent permutations

- $\pi : U \to U$ a random permutation of U
- $h(A) = \min\{\pi(x) \mid x \in A\}$
- then
$$\Pr[h(A) = h(B)] = J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$
- amplify the probability as before:
  - repeat many times,
  - concatenate into blocks
  - consider objects similar if they collide in at least one block

homework

show that for $h(A) = \min\{\pi(x) \mid x \in A\}$ with π a random permutation

$$\Pr[h(A) = h(B)] = J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

homework

design a locality-sensitive hashing scheme for vectors according to the cosine similarity measure

vectors $\mathbf{x} = \{x_1, \ldots, x_d\}$

distance $1 - \cos(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}||_2 \, ||\mathbf{y}||_2}$

computing statistics on data streams

- $X = (x_1, x_2, \ldots, x_m)$ a sequence of elements
- each x_i is a member of the set $N = \{1, \ldots, n\}$
- $m_i = |\{j : x_j = i\}|$ the number of occurrences of i
- define

$$F_k = \sum_{i=1}^{n} m_i^k$$

- F_0 is the number of distinct elements
- F_1 is the length of the sequence
- F_2 index of homogeneity, size of self-join, and other applications

computing statistics on data streams

- How to compute the frequency moments using less than $O(n \log m)$ space?
- sketching: create a sketch that takes much less space and gives an estimation of F_k

estimating the number of distinct values (F_0)

Theorem For every $c > 2$, the algorithm computes a number Y using $O(logn)$ memory bits, such that the probability that the ratio between Y and F_0 is not between $1/c$ and c is at most $2/c$.

estimating the number of distinct values (F_0)

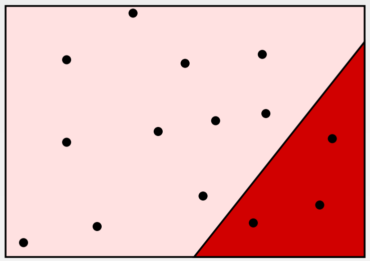
- [Flajolet and Martin, 1985]
- consider a bit vector of length $O(\log n)$
- upon seen x_i, set:
  - the 1st bit with probability $1/2$
  - the 2nd bit with probability $1/4$
  - ...
  - the i-th bit with probability $1/2^i$
- important: bits are set deterministically for each x_i
- let R be the index of the largest bit set
- return $Y = 2^R$

estimating number of distinct values (F_0)

Theorem. For every $c > 2$, the algorithm computes a number Y using $O(logn)$ memory bits, such that the probability that the ratio between Y and F_0 is not between $1/c$ and c is at most $2/c$.

estimator theorem

- consider a set of items U
- a fraction ρ of them have a specific property
- estimate ρ by sampling



- how many samples N are needed?

$$N \geq \frac{4}{\epsilon^2 \rho} \log \frac{2}{\delta}.$$

for an ϵ-approximation with probability at least $1 - \delta$

- notice: it does not depend on $|U|$ (!)

applications of the algorithmic tools to real scenarios diameter

diameter

- how to compute the diameter of a graph?
- matrix multiplication in $O(n^{2.376})$ time,
  but $O(n^2)$ space
- BFS from a vertex takes $O(n+m)$ time,
  but need to do it from every vertex, so $O(mn)$
- resort to approximations again

approximating the diameter

- [Palmer et al., 2002], see also [Cohen, 1997]
- define:

  Individual neighborhood function

  $$N(u, h) = |\{v \mid d(u, v) \leq h\}|$$

  Neighborhood function

  $$N(h) = |\{(u, v) \mid d(u, v) \leq h\}| = \sum_u N(u, h)$$

- $N(h)$ can be used to obtain diameter, effective diameter, etc.

approximating the diameter

- define: $M(u,h) = \{v \mid d(u,v) \leq h\}$, e.g., $M(u,0) = \{u\}$
- algorithm based on the idea that
  $x \in M(u,h)$ if $(u,v) \in E$ and $x \in M(v,h-1)$

  ANF [Palmer et al., 2002]

  $M(u,0) = \{u\}$ for all $u \in V$
  for each distance h **do**
  $\quad M(u,h) = M(u,h-1)$ for all $u \in V$
  $\quad$ **for each** edge (u,v) **do**
  $\quad\quad M(u,h) = M(u,h) \cup M(v,h-1)$

- keep $M(u,h)$ in memory, make a passes over the edges
- how to maintain $M(u,h)$?

approximating the diameter

- how to maintain $M(u, h)$ that it counts *distinct* vertices?
- the problem of counting distinct elements in data streams
- ANF uses the sketching algorithm of [Flajolet and Martin, 1985] with $O(\log n)$ space (but other counting algorithms can be used [Bar-Yossef et al., 2002])
- what if the $M(u, h)$ sketches do not fit in memory?
- split $M(u, h)$ sketches into in-memory blocks, load one block at the time, and process edges from that block

clustering coefficient and triangles

clustering coefficient

$$C = \frac{3 \times \text{ number of triangles in the network}}{\text{number of connected triples of vertices}}$$

- how to compute it?
- how to compute the number of triangles in a graph?
- assume that the graph is very large, stored in disk
- [Buriol et al., 2006]
- count triangles, when graph is seen as a data stream
- two models:
  - edges are stored in any order
  - edges in order — all edges incident to one vertex are stored sequentially

counting triangles

- brute-force algorithm is checking every triple of vertices
- obtain an approximation by sampling triples
- let T be the set of all triples and
  T_i the set of triples that have i edges, $i = 0, 1, 2, 3$
- by the estimator theorem, to get an ϵ-approximation, with probability $1 - \delta$, the number of samples should be

$$N \geq O(\frac{|T|}{|T_3|} \frac{1}{\epsilon^2} \log \frac{1}{\delta})$$

- but $|T|$ can be very large compared to $|T_3|$

sampling algorithm for counting triangles

- incidence model
- 2-pass algorithm
- consider sample space $\mathcal{S} = \{b\text{-}a\text{-}c \mid (a,b),(a,c) \in E\}$
- $|\mathcal{S}| = \sum_i d_i(d_i - 1)/2$

1: sample $X \subseteq \mathcal{S}$ (paths b-a-c)
2: estimate fraction of X for which edge (b,c) is present
3: scale by $|\mathcal{S}|$

- gives (ϵ, δ) approximation

counting triangles — incidence stream model

SampleTriangle [Buriol et al., 2006]

1st pass

count the number of paths of length 2 in the stream

2nd pass

uniformly choose one path (a, b, c)

3rd pass

if $((b, c) \in E)$ $\beta = 1$ else $\beta = 0$

return β

we have $\mathrm{E}[\beta] = \frac{3|T_3|}{|T_2|+3|T_3|}$, with $|T_2| + 3|T_3| = \sum_u \frac{d_u(d_u-1)}{2}$, so

$$|T_3| = \mathrm{E}[\beta] \sum_u \frac{d_u(d_u-1)}{6}$$

and space needed is $O((1 + \frac{|T_2|}{|T_3|}) \frac{1}{\epsilon^2} \log \frac{1}{\delta})$

properties of the sampling space

it should be possible to

- estimate the size of the sampling space
- sample an element uniformly at random

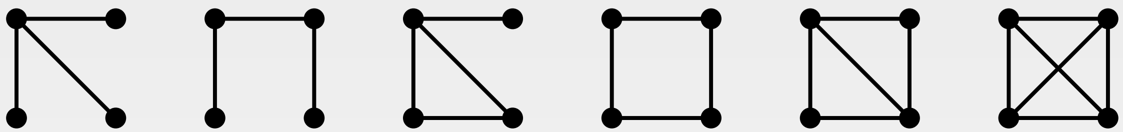
counting graph minors

counting other minors

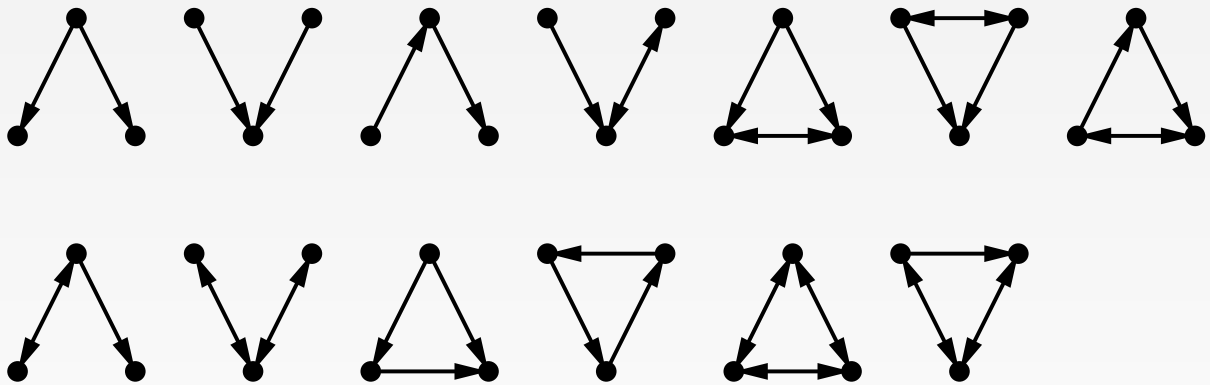
- count all minors in a very large graphs
  - connected subgraphs
  - size 3 and 4
  - directed or undirected graphs
- why?
  - modeling networks, “signature” structures, e.g., copying model
  - anomaly detection, e.g., spam link farms
  - indexing graph databases

counting minors in large graphs

- characterize a graph by the distribution of its minors



All undirected minors of size 4



All directed minors of size 3

sampling algorithm for counting triangles

- incidence model
- 2-pass algorithm
- consider sample space $\mathcal{S} = \{b\text{-}a\text{-}c \mid (a, b), (a, c) \in E\}$
- $|\mathcal{S}| = \sum_i d_i(d_i - 1)/2$

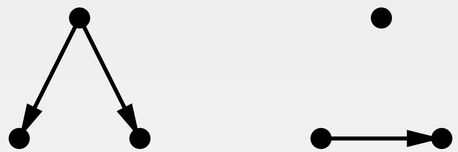
1: sample $X \subseteq \mathcal{S}$ (paths b-a-c)
2: estimate fraction of X for which edge (b, c) is present
3: scale by $|\mathcal{S}|$

- gives (ϵ, δ) approximation
- adapt the algorithm to count all minors of size 3 and 4 and directed and undirected graphs

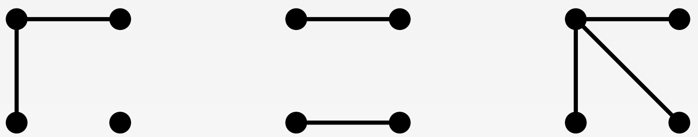
adapting the algorithm

sampling spaces:

- 3-node directed



- 4-node undirected



are the sampling space properties satisfied?

datasets

graph class	type	# instances
synthetic	un/directed	39
wikipedia	un/directed	7
webgraphs	un/directed	5
cellular	directed	43
citation	directed	3
food webs	directed	6
word adjacency	directed	4
author collaboration	undirected	5
autonomous systems	undirected	12
protein interaction	undirected	3
US road	undirected	12

clustering of undirected graphs

assigned to	0	1	2	3	4	5	6
AS graph	12	0	0	0	0	0	0
collaboration	0	0	3	2	0	0	0
protein	1	0	0	1	0	0	1
road-graph	0	12	0	0	0	0	0
wikipedia	0	0	0	0	2	5	0
synthetic	11	0	0	0	0	0	28
webgraph	2	0	0	1	0	0	0

clustering of directed graphs

feature class	error compared to ground truth
standard topological properties (81)	74.00%
minors of size 3	77.78%
minors of size 4	84.26%
minors of size 3 and 4	90.74%

local statistics

compute local statistics in large graphs

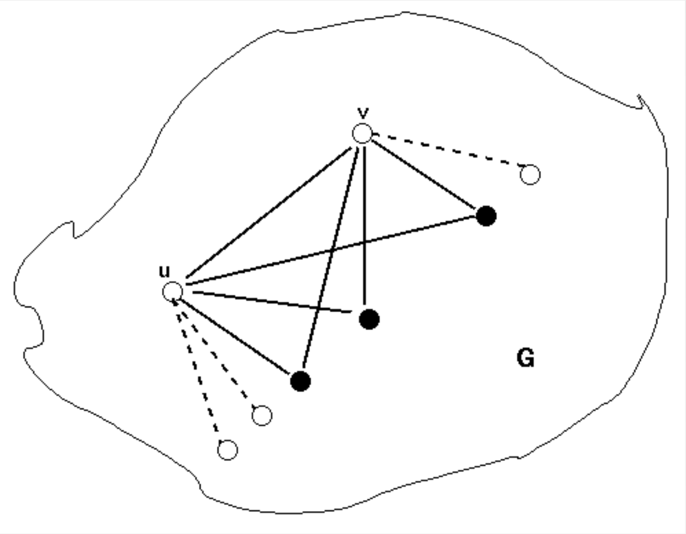
- our goal: compute triangle counts for all vertices
- local clustering coefficient and related statistics
- motivation
  - motifs can be used to characterize network families [Alon, 2007]
  - analysis of social or biological networks
  - thematic relationships in the web
  - web spam
- applications: spam detection and content quality analysis in social media

two algorithms

1. external memory
   - keep a counter for each vertex (main memory)
   - keep a counter for each edge (secondary memory)

2. main memory
   - keep a counter for each vertex

number of triangles for edges and nodes



- neighbors: $N(u) = \{v : (u, v) \in E\}$
- degree: $d(u) = |N(u)|$
- edge triangles: $T_{uv} = |N(u) \cap N(v)|$
- vertex triangles: $T(u) = \frac{1}{2} \sum_{v \in N(u)} T_{uv}$

computing triangles: algorithm idea

- consider the Jaccard coefficient between two sets A and B:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

- if we knew $J(N(u), N(v)) = J$, then:

$$T_{uv} = |N(u) \cap N(v)| = \frac{J}{J+1}(|N(u)| + |N(v)|)$$

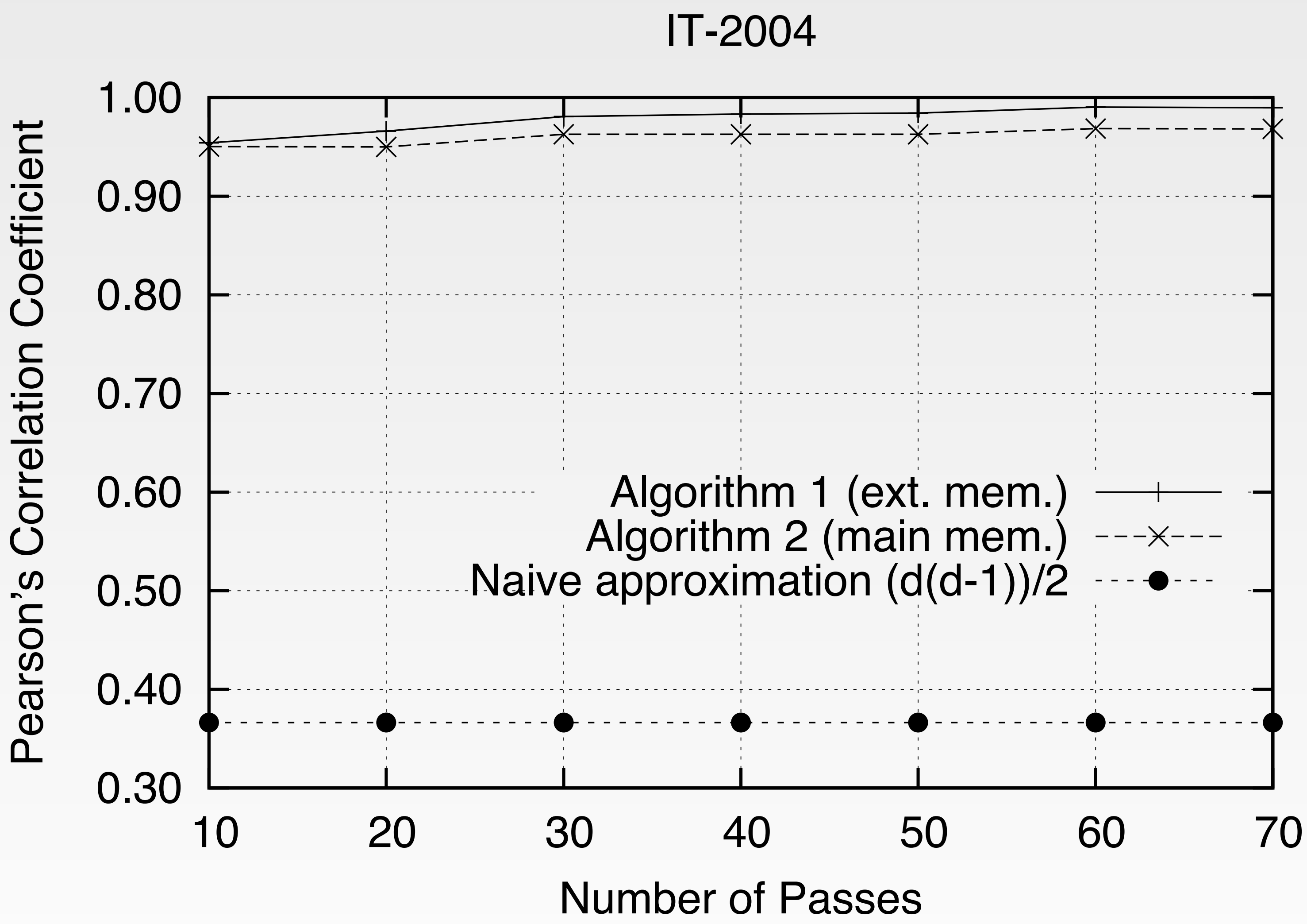
- and then:

$$T(u) = \frac{1}{2} \sum_{v \in N(u)} T_{uv}$$

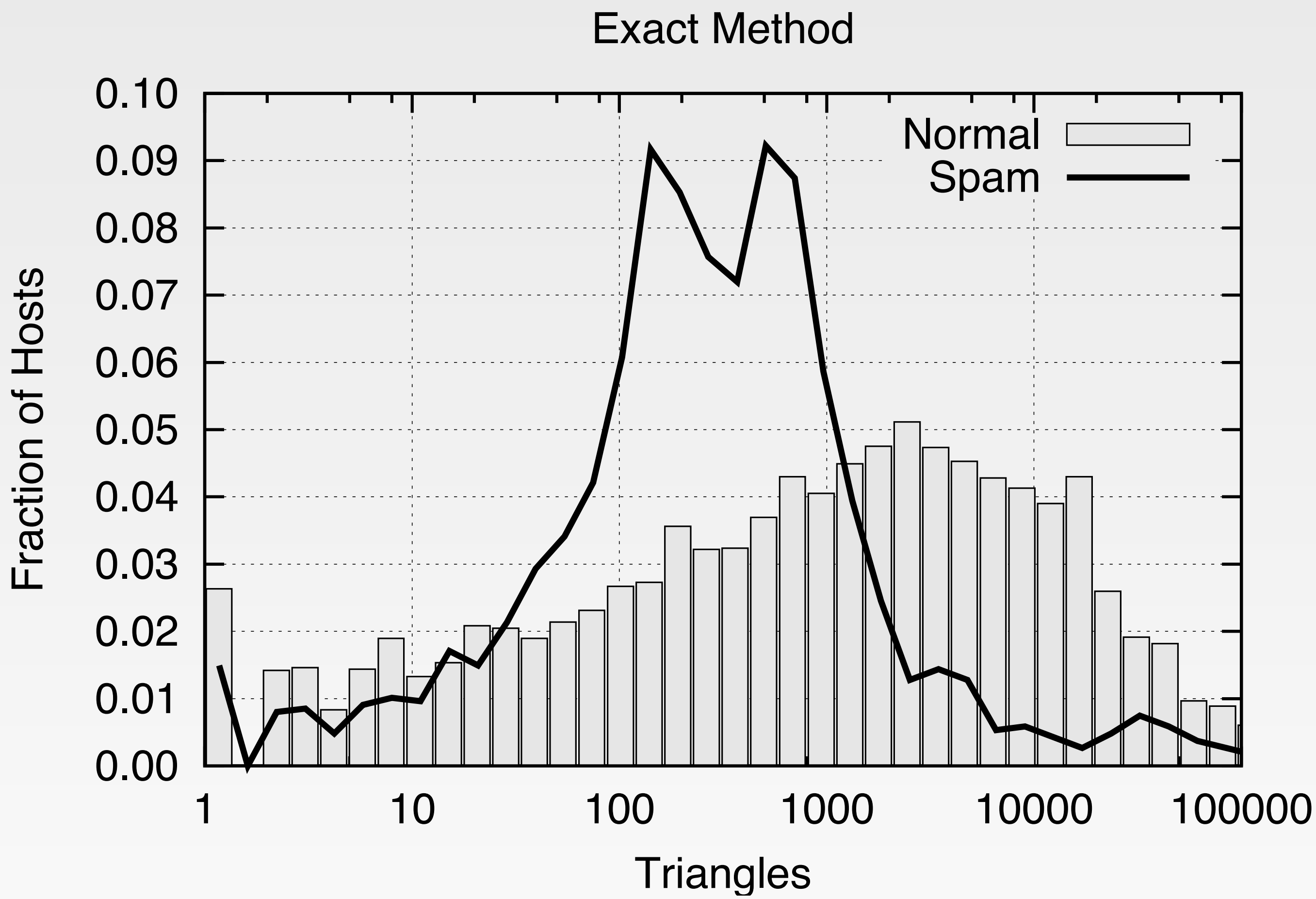
external-memory algorithm

- semi-stream model
- keep min-hash values for the graph nodes (in memory)
- keep counters for edges (on disk)
- use counters for edges to estimate number of triangles and local clustering coefficient

quality of approximation

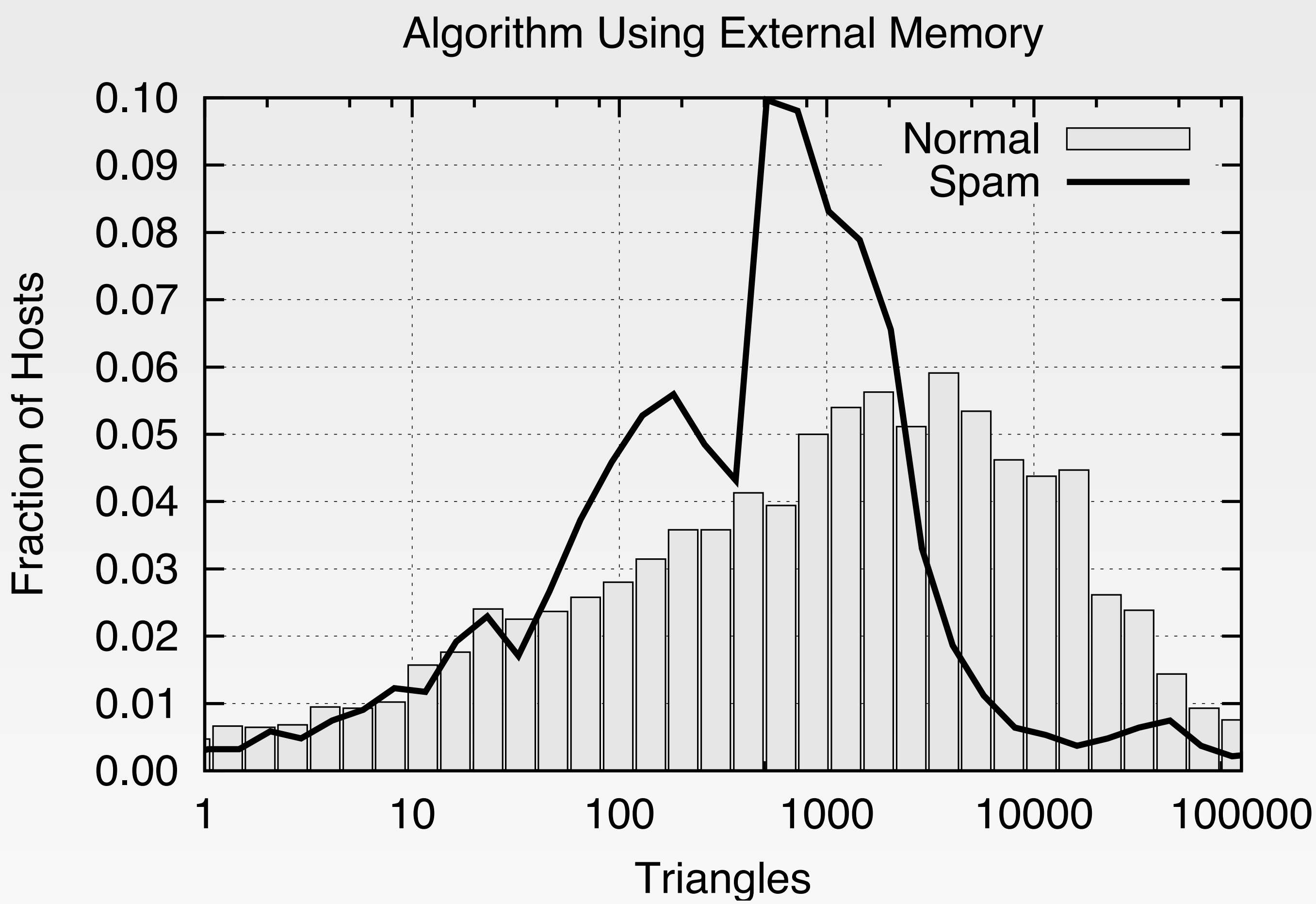


applications: spam detection



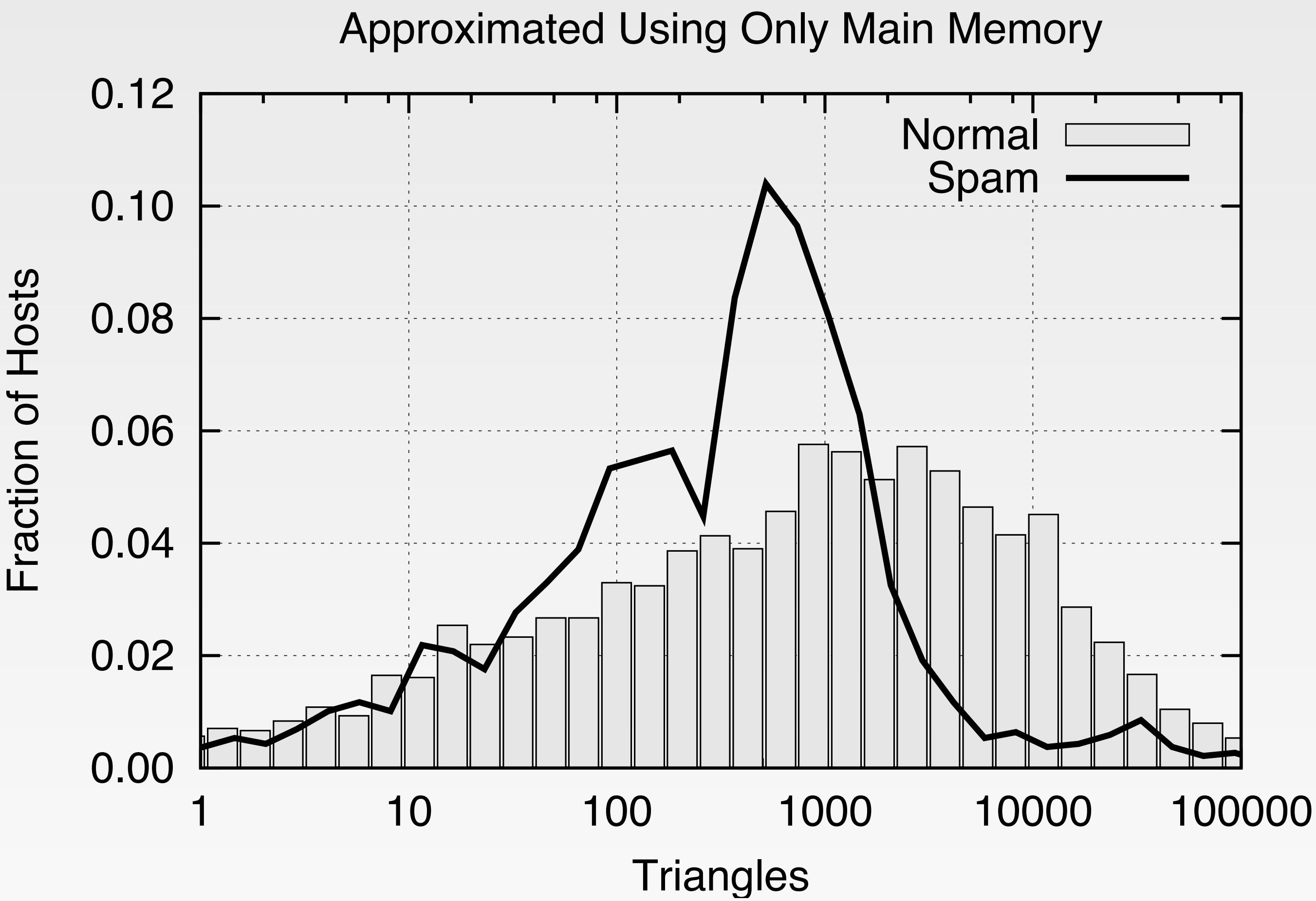
Separation of non-spam and spam hosts in the histogram of triangles

applications: spam detection



Separation of non-spam and spam hosts in the histogram of triangles

applications: spam detection



Separation of non-spam and spam hosts in the histogram of triangles

applications: spam detection

number of triangles feature is ranked 60-th out of 221 for spam detection estimating the size of the web

what is the size of the web?

- issues
  - the web is really infinite
  - dynamic content, e.g., calendar
  - *soft 404*: `www.yahoo.com/anything` is a valid page
  - static web contains syntactic duplication, mostly due to mirroring ($\approx$20-30%)
- who cares?
  - media, and consequently the user
  - engine design
  - engine crawl policy
  - impact on recall

what can we attempt to measure?

- the relative size of search engines
- the notion of a page being indexed is still reasonably well defined
  - document extension: e.g., Yahoo indexes pages not yet crawled by indexing anchor-text
  - document restriction: some engines restrict what is indexed (first n words, only relevant words, etc.)

relative size of search engines

- [Bharat and Broder, 1998]
- main idea:

$$\Pr[A \cap B \mid A] = \frac{s(A \cap B)}{s(A)} \text{ and } \Pr[A \cap B \mid B] = \frac{s(A \cap B)}{s(B)}$$

- therefore

$$\frac{s(A)}{s(B)} = \frac{\Pr[A \cap B \mid B]}{\Pr[A \cap B \mid A]}$$

- need:
  - sampling a random page from the index of a search engine
  - checking if a page exists at the index of a search engine

sampling and checking pages

- [Bharat and Broder, 1998]
- both tasks by using the public interface search engines
- sampling:
  - construct a large lexicon
  - use the lexicon to fire random queries
  - sample a page from the results
  - (introduces query and ranking biases)
- checking:
  - construct a strong query from the most k most distinctive terms of the page
  - (in order to deal with aliases, mirror pages, etc.)

random-walk sampling

- [Bar-Yossef and Gurevich, 2006]
- define a graph on documents and queries:
  - edge (d, q) indicates that document d is a result of a query q
- random walk gives biased samples
- bias depends on the degree of docs and queries
- use Monte Carlo methods to unbias the samples and obtain uniform samples
- paper shows how to obtain estimates of the degrees and weights needed for the unbiasing

results of random-walk sampling

- [Bar-Yossef and Gurevich, 2006]

	G	M	Y
G		46%	45%
M	55%		51%
Y	44%	22%	

near-duplicate detection

mirror sites

- mirroring is systematic replication of web pages across hosts
- single largest cause of duplication on the web

why detect mirrors?

- smart crawling
  - fetch from the fastest or freshest server
  - avoid duplication
- better connectivity analysis
  - combine inlinks
  - avoid double counting outlinks
- redundancy in result listings
  - "if that fails you can try: `<mirror>/samepath`"
- proxy caching

study the genealogy of the web

- new pages copy content from existing pages
- web genealogy study:
  - how textual content of source pages (parents) are reused to compose part of new Web pages (children)
  - not near-duplicates, as similarities of short passages are also identified
- how can search engines benefit?
  - by associating more relevance to a parent page?
  - by trying to decrease the bias?

study the genealogy of the web

- define concepts such as
  - parents
  - children
  - sterile parents
  - orphans
  - etc.
- correlate well-studied measures (such as PageRank) for different types of documents and draw interesting conclusions

more about syntactic similarity

- bag of words representation
  - each document D is represented as the set $b(D)$ of words that it contains
- define similarity between two documents using Jaccard

$$\mathrm{sim}(A, B) = \frac{|b(A) \cap b(B)|}{|b(A) \cup b(B)|}$$

- two documents considered near-duplicates if $\mathrm{sim}(A, B) \geq \alpha$

more about syntactic similarity

- bag of words representation does not capture well the concept of syntactic similarity
- shingles

  "a rose is a rose is a rose" becomes

  a rose is a

  rose is a rose

  is a rose is

  a rose is a

  rose is a rose

- bag representation of shingles
- same complexity

algorithm for mirror detection

- locality sensitive hashing using min-wise independent permutations
- documents are hashed according to the LSH scheme
- mirror documents hashed to the same value (w.h.p.)
- sort the documents and examine the ones with the same values
- fine parameter tuning is required to make it scalable and successful

indexing distances in large graphs

indexing distances in large graphs

[Potamias et al., 2009]

- motivation: context-sensitive search and social search
- input: consider a graph $G = (V, E)$
- and nodes s and t in V
- goal: compute (fast) shortest-path distance $d(s, t)$ from s to t

- BFS takes $O(m)$
- too expensive for large graphs

landmark-based approach

- precompute: distance from each node to a fixed ladmark l
- then

$$|d(s,l) - d(t,l)| \leq d(s,t) \leq d(s,l) + d(l,t)$$

- precompute: distances to d landmarks, $l_1, \ldots, l_d$

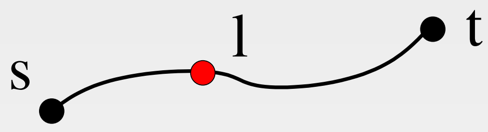
$$\max_i |d(s,l_i) - d(t,l_i)| \leq d(s,t) \leq \min_i (d(s,l_i) + d(l_i,t))$$

- obtain a range estimate in time $O(d)$ (i.e., constant)

landmark-based approach

- motivated by indexing general metric spaces
- used for estimating latency in the internet [Ng and Zhang, 2008]
- already used for social search [Vieira et al., 2007] distance from each node to a fixed ladmark l
- typically randomly chosen landmarks
- in our work: we investigate techniques for selecting good landmarks

good landmarks

if 

then $d(s,t)=d(s,l)+d(l,t)$

if 

then $|d(s,l)-d(t,l)|=d(s,t)$

good (upper-bound) landmarks

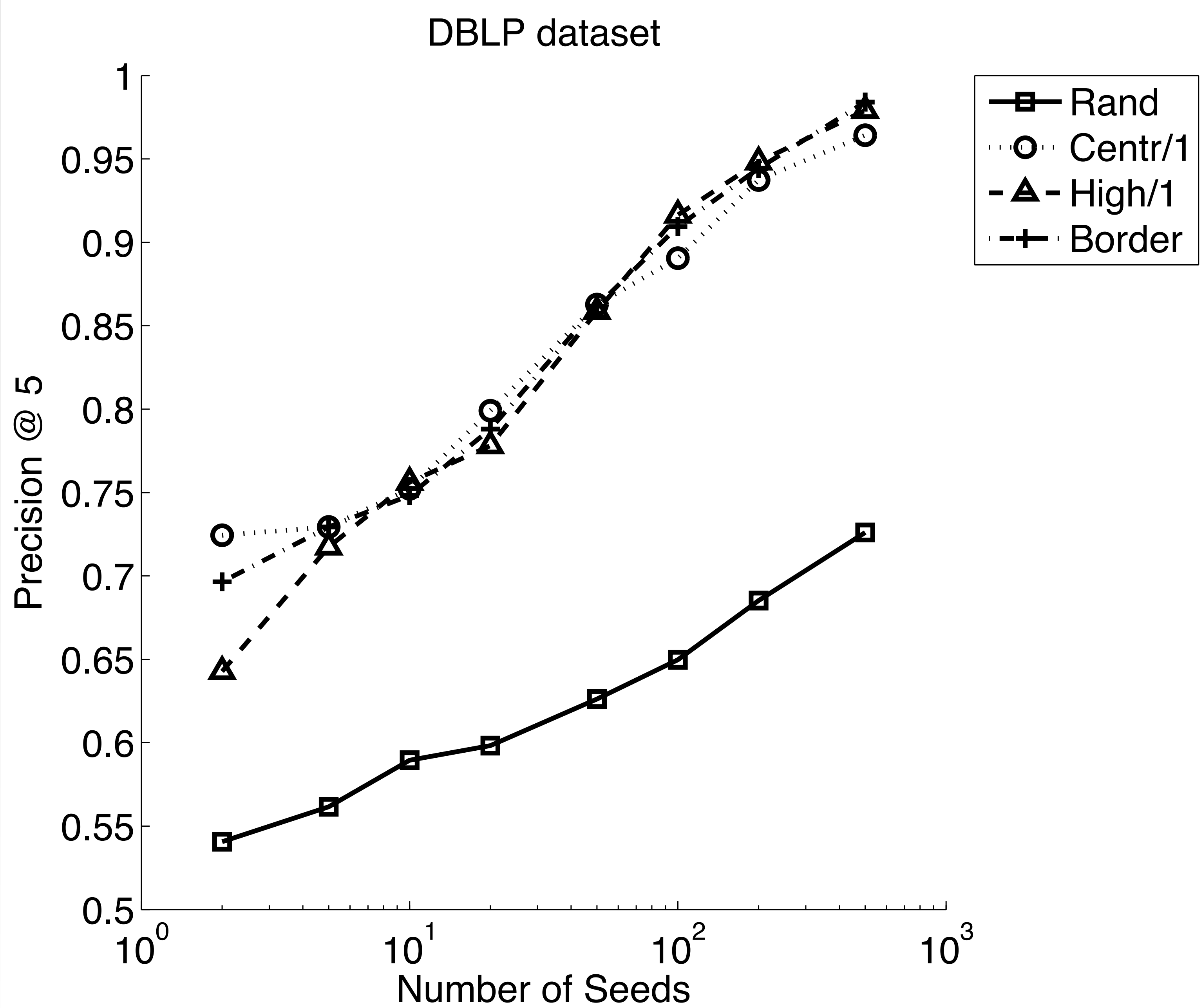
- a landmark l covers a pair (s, t) if l is on a shortest path from s to t
- problem definition: find a set $L \subseteq V$ of k landmarks that cover as many pairs $(s, t) \in V \times V$ as possible

- NP-hard
- for $k = 1$: the node with the highest centrality betweenness
- for $k > 1$: apply a "natural" set-cover approach (but $O(n^3)$)

landmark selection heuristics

- high-degree nodes
- high-centrality nodes
- “constrained” versions
  - once a node is selected none of its neighbors is selected
- “clustered” versions
  - cluster the graph and select one landmark per cluster
  - select landmarks on the “borders” between clusters

DBLP — precision @ 5



comparing with exact algorithm

[Goldberg and Harrelson, 2005]

Ours (10%)	Fl.-E	Fl.-I	Wiki	DBLP	Y!IM
Method	CENT	CENT	CENT/P	BORD/P	BORD/P
Landmarks used	20	100	500	50	50
Nodes visited	1	1	1	1	1
Operations	20	100	500	50	50
CPU ticks	2	10	50	5	5

ALT (exact)	Fl.-E	Fl.-I	Wiki	DBLP	Y!IM
Method	Ikeda	Ikeda	Ikeda	Ikeda	Ikeda
Landmarks used	8	4	4	8	4
Nodes visited	7245	10337	19616	2458	2162
Operations	56502	41349	78647	19666	8648
CPU ticks	7062	10519	25868	1536	1856

mining graph evolution rules

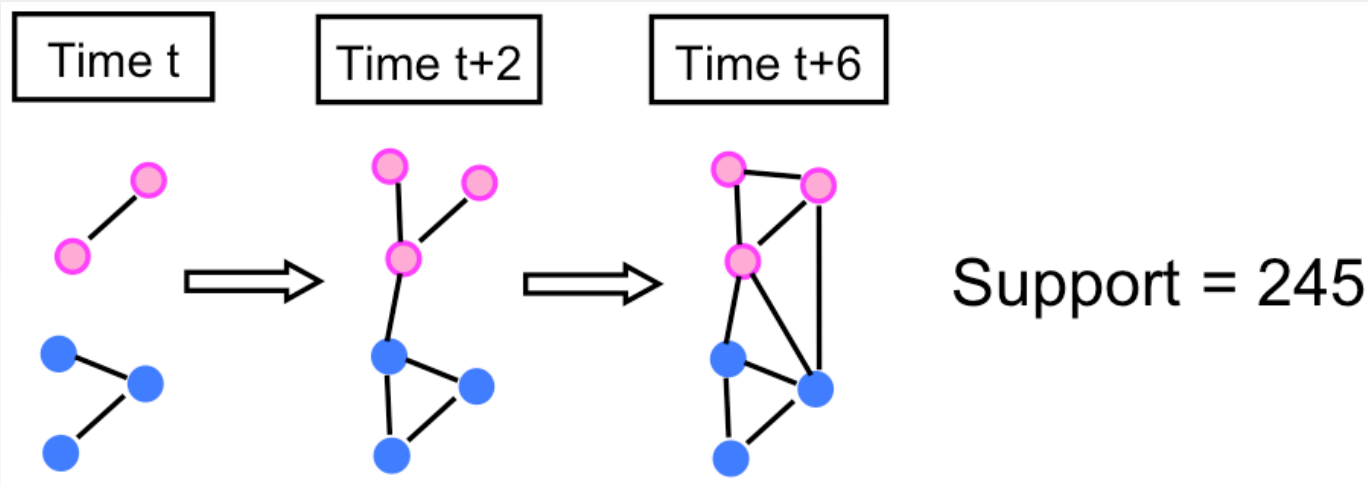
motivation

- objective: study the evolution of a graph over time
- traditionally, study static properties of graphs, e.g.,
  - degree distribution
  - small-world structure
  - communities
- more recently, study evolution of graphs at macroscopic level
  - models of evolution
  - densification law and shrinking diameter
- focus on microscopic level
- adopt a frequent-pattern mining approach

the problem

[Berlingerio et al., 2009]

- given a sequence of snapshots of a dynamic network $G_1, \ldots, G_k$, and a minimum support threshold σ
- find frequent patterns, such as:



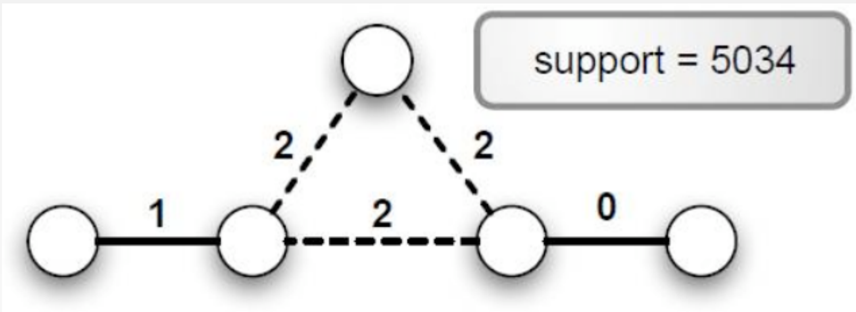
- nodes and edges can have labels
- from those patterns find graph-evolution rules, which satisfy also a confidence threshold γ

the approach

- trick: represent a sequence of snapshots, as a single graph with time-stamped edges
- adapt existing technology for mining single graphs

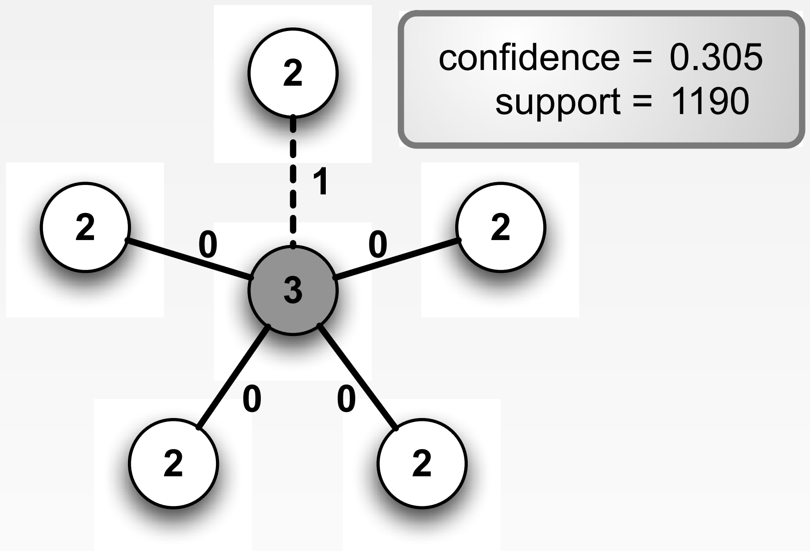
patterns

- two types of patters
  - absolute-time – less interesting
  - relative-time – more interesting

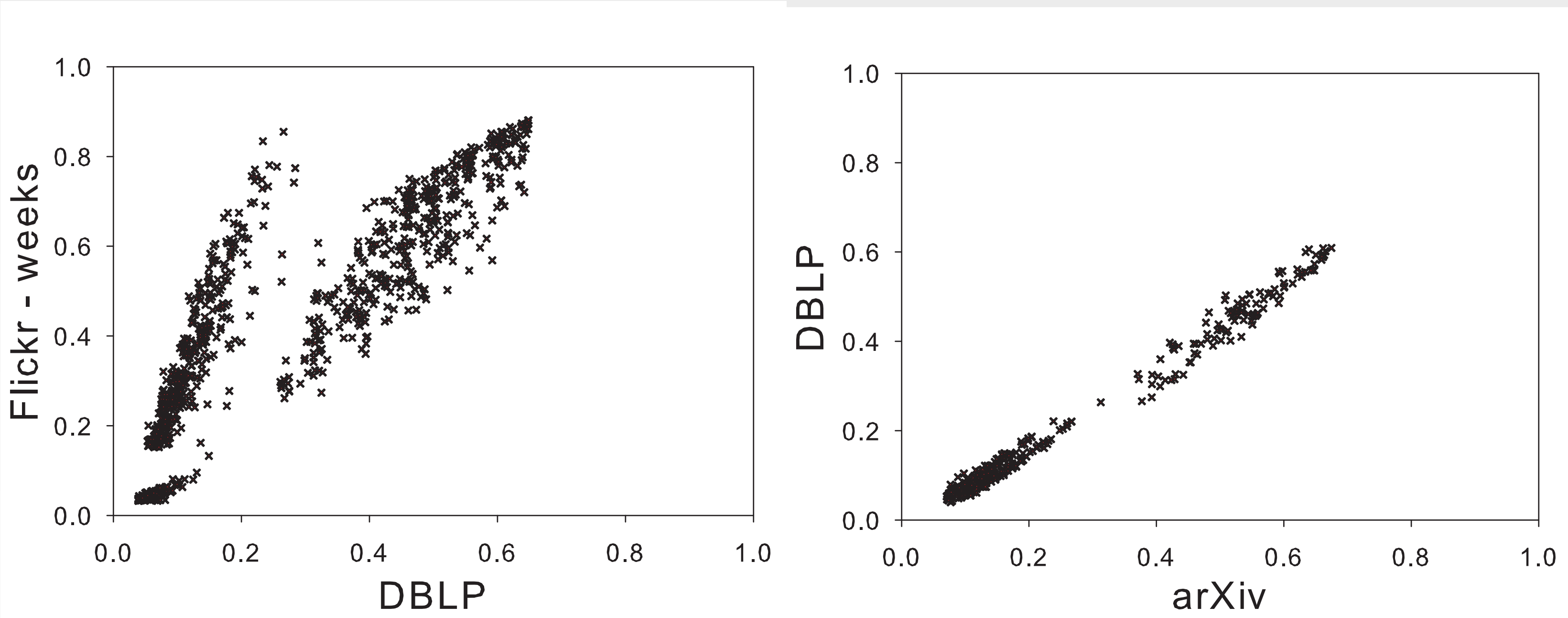


graph-evolution rules

- a rule has the form body $\rightarrow$ head
  - body: all the edges except the most recent ones
  - head: the complete pattern
- confidence: relative frequency

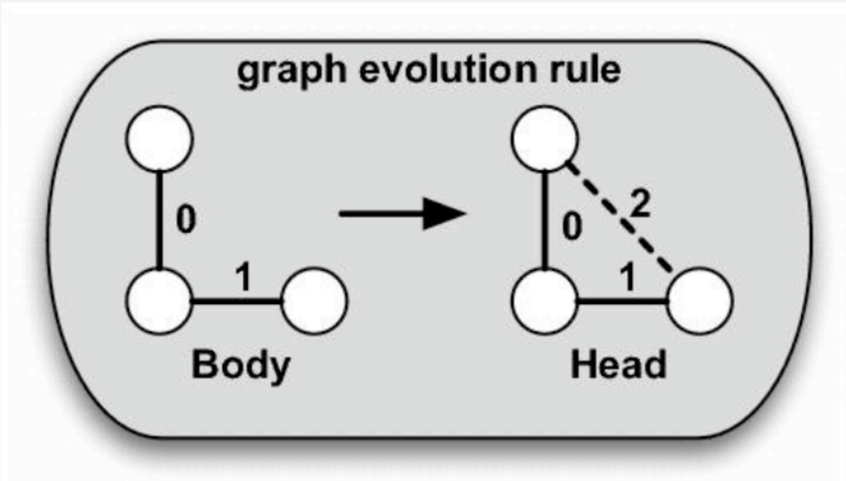


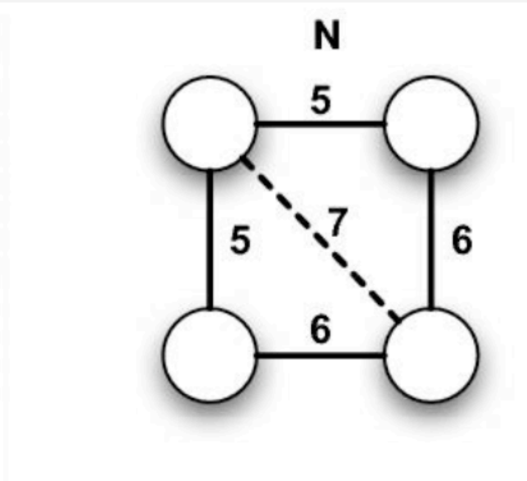
graph-evolution rules



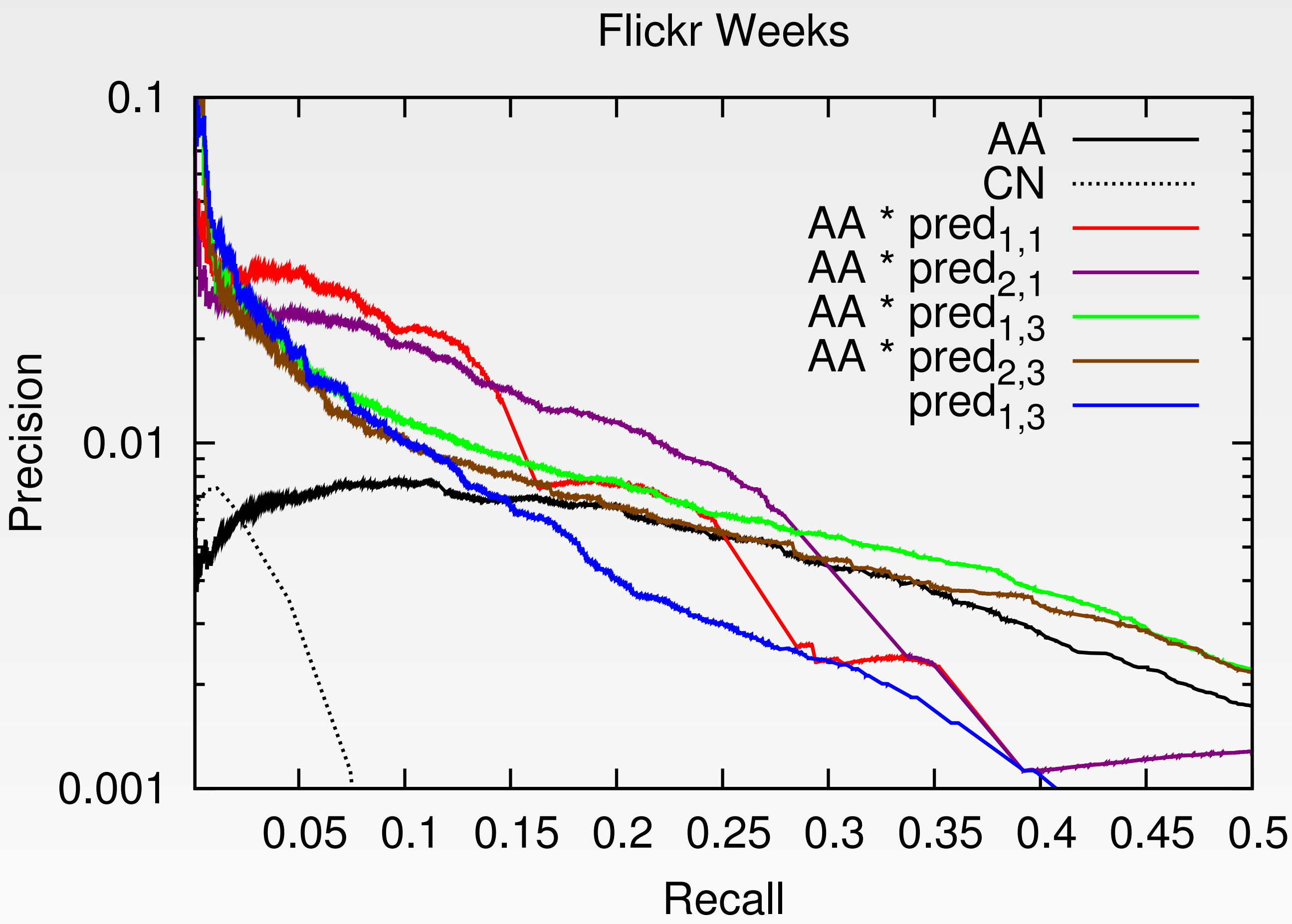
application to link prediction

- link-prediction problem: predict future edges in the network [Nowell and Kleinberg, 2003]
- approach:
  1. find rules from previous snapshots
  2. identify embeddings of the body of the rule
  3. predict new edges





application to link prediction



advantages of the approach

- predict arrival of new nodes
- predict arrival time
- predict future using present and past

thank you!

references I

Alon, U. (2007).
Network motifs: theory and experimental approaches.
Nature Reviews Genetics.

Anagnostopoulos, A., Becchetti, L., Castillo, C., and Gionis, A. (2010).
An optimization framework for query recommendation.
In *WSDM*.

Baeza-Yates, R. and Castillo, C. (2005).
Link analysis in national Web domains.
In Beigbeder, M. and Yee, W. G., editors, *Workshop on Open Source Web Information Retrieval (OSWIR)*, pages 15–18, Compiegne, France.

references II

Baeza-Yates, R. and Tiberi, A. (2007).
Extracting semantic relations from query logs.
In *KDD*.

Bar-Yossef, Z. and Gurevich, M. (2006).
Random sampling from a search engine's index.
In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 367–376, New York, NY, USA. ACM Press.

Bar-Yossef, Z., Jayram, T. S., Kumar, R., Sivakumar, D., and Trevisan, L. (2002).
Counting distinct elements in a data stream.
In *Proceedings of the 6th International Workshop on Randomization and Approximation Techniques (RANDOM)*, pages 1–10, Cambridge, Ma, USA. Springer-Verlag.

references III

Barabási, A. L. and Albert, R. (1999).
Emergence of scaling in random networks.
Science, 286(5439):509–512.

Berlingerio, M., Bonchi, F., Bringmann, B., and Gionis, A. (2009).
Mining graph evolution rules.
In *ECML/PKDD (1)*, pages 115–130.

Bharat, K. and Broder, A. (1998).
Estimating the size of the web.
In *Proceedings of the wWW conference*.

Boldi, P., Bonchi, F., Castillo, C., Donato, D., Gionis, A., and Vigna, S. (2008).
The query-flow graph: model and applications.
In *Proceeding of the 17th ACM conference on Information and knowledge management (CIKM)*.

references IV

Bordino, I., Castillo, C., Donato, D., and Gionis, A. (2010).
Query similarity by projecting the query-flow graph.
In *SIGIR*.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000).
Graph structure in the web: Experiments and models.
In *Proceedings of the Ninth Conference on World Wide Web*, pages 309–320, Amsterdam, Netherlands. ACM Press.

Buriol, L. S., Frahling, G., Leonardi, S., Marchetti-Spaccamela, A., and Sohler, C. (2006).
Counting triangles in data streams.
In *PODS '06: Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 253–262, New York, NY, USA. ACM Press.

references V

Cohen, E. (1997).

Size-estimation framework with applications to transitive closure and reachability.

Journal of Computer and System Sciences, 55(3):441–453.

Craswell, N. and Szummer, M. (2007).

Random walks on the click graph.

In *Proceedings of the 30th annual international ACM conference on Research and development in information retrieval (SIGIR)*.

Flajolet, P. and Martin, N. G. (1985).

Probabilistic counting algorithms for data base applications.

Journal of Computer and System Sciences, 31(2):182–209.

Goldberg, A. and Harrelson, C. (2005).

Computing the shortest path: A* search meets graph.

In *SODA*.

references VI

Gyöngyi, Z., Garcia-Molina, H., and Pedersen, J. (2004).
Combating Web spam with TrustRank.
In *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB)*, pages 576–587, Toronto, Canada. Morgan Kaufmann.

Haveliwala, T. H. (2002).
Topic-sensitive pagerank.
In *Proceedings of the Eleventh World Wide Web Conference*, pages 517–526, Honolulu, Hawaii, USA. ACM Press.

Huberman, B. A. and Adamic, L. A. (1999).
Growth dynamics of the world-wide web.
Nature, 399.

references VII

Kleinberg, J. M. (1999).

Authoritative sources in a hyperlinked environment.

Journal of the ACM, 46(5):604–632.

Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. S. (1999).

The Web as a graph: measurements, models and methods.

In *Proceedings of the 5th Annual International Computing and Combinatorics Conference (COCOON)*, volume 1627 of *Lecture Notes in Computer Science*, pages 1–18, Tokyo, Japan. Springer.

Lempel, R. and Moran, S. (2000).

The stochastic approach for link-structure analysis (salsa) and the tkc effect.

In *ACM Transactions on Information Systems*, pages 387–401.

references VIII

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005).

Graphs over time: densification laws, shrinking diameters and possible explanations.

In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187, New York, NY, USA. ACM Press.

M. Faloutsos, P. Faloutsos, C. F. (1999).

On power-law relationships of the internet topology.

In *SIGCOMM*.

Mitzenmacher, M. (2004).

A brief history of generative models for power law and lognormal distributions.

Internet Mathematics, 1(2):226–251.

references IX

Newman, M. E. J. (2003).
The structure and function of complex networks.

Ng, A. Y., Zheng, A. X., and Jordan, M. I. (2001).
Stable algorithms for link analysis.
In *NIPS*, pages 258–266. ACM Press.

Ng, E. and Zhang, H. (2008).
Predicting internet network distance with coordinate-based approaches.
In *INFOCOMM*.

Nowell, D. L. and Kleinberg, J. (2003).
The link prediction problem for social networks.
In *CIKM*.

references X

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998).
The PageRank citation ranking: bringing order to the Web.
Technical report, Stanford Digital Library Technologies Project.

Palmer, C. R., Gibbons, P. B., and Faloutsos, C. (2002).
ANF: a fast and scalable tool for data mining in massive graphs.
In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 81–90, New York, NY, USA. ACM Press.

Potamias, M., Bonchi, F., Castillo, C., and Gionis, A. (2009).
Fast shortest path distance estimation in large networks.
In *CIKM*.

Simon, H. A. (1955).
On a class of skew distribution functions.
Biometrika, 42(3/4):425.

references XI

Solis, B. (2007).

The definition of social media.

Vieira, M. V., Fonseca, B. M., Damazio, R., Golgher, P. B., de Castro Reis, D., and Ribeiro-Neto, B. A. (2007).

Efficient search ranking in social networks.

In *CIKM*.

Yule, G. U. (1925).

A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis.

Philosophical transactions of the Royal Society of London, 213:21–87.