# An introduction to Web Mining

## part I

**Ricardo Baeza-Yates, Aristides Gionis**

**Yahoo! Research**

**Barcelona, Spain & Santiago, Chile**

**ECML/PKDD 2008 Antwerp**

# Contents of the tutorial

1. Motivation of web mining
2. The mining process
   - data anonymization and data modeling
3. The basic methods
   - usage mining, link mining, algorithmic tools, finding communities
4. Detailed examples
   - Size of the web, near-duplicate detection, spam detection based on content and links

# Disclaimer

- Topics reflect the presenters' subjective choices
- Cannot be complete and cover all topics
- Your feedback will be highly appreciated

rby@yahoo-inc.com<br>
gionis@yahoo-inc.com

# Intended audience

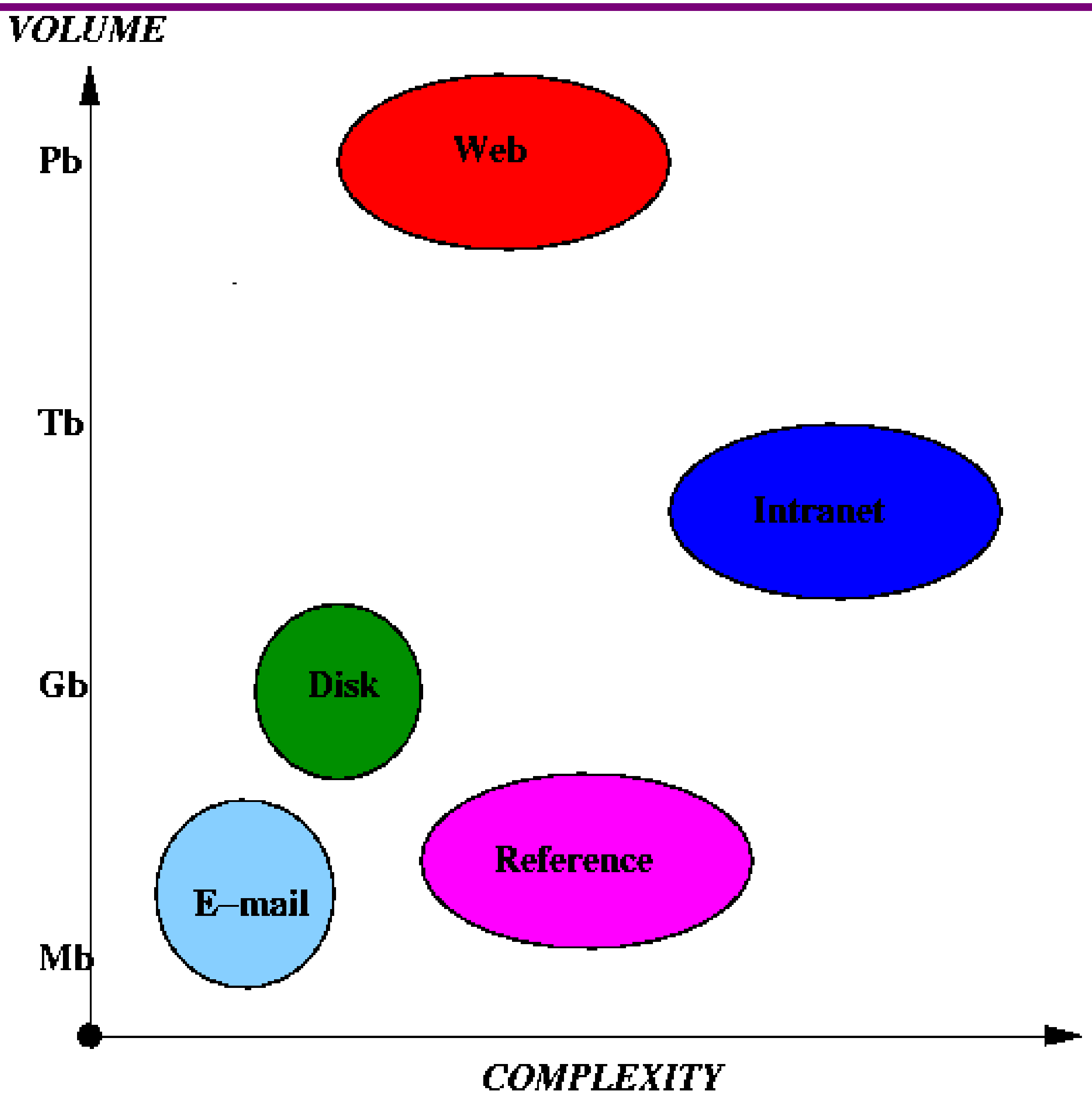
- Beginning research students who want to work in the area of Web mining
- Researchers who want would like to work in Web mining and want to obtain a view of the problems, issues, and solutions

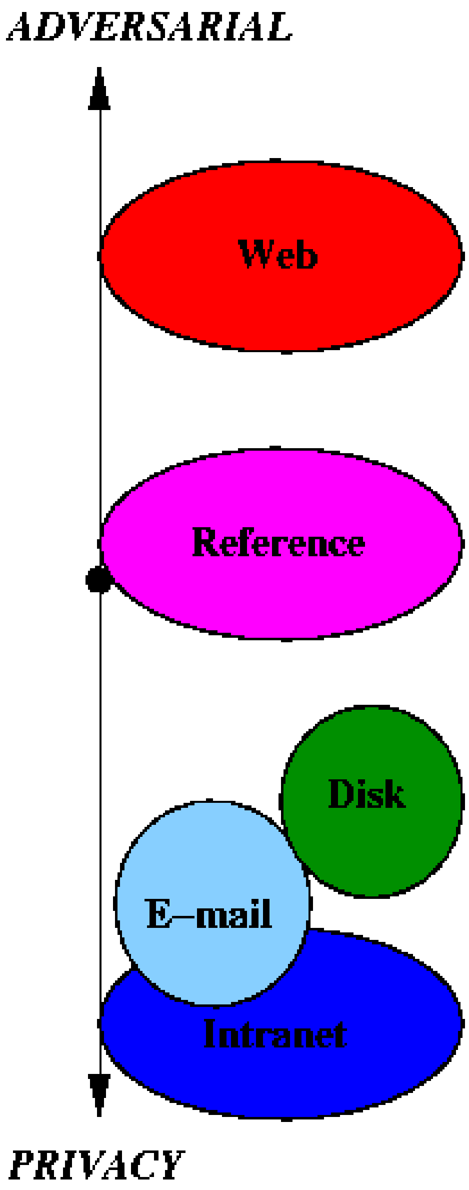
# Introduction and motivation

# 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 in 2010
- Internet traffic has increased 20 times in the last 5 years
- Today there are more than 170 million Web servers
- The Web is in practice unbounded
  - Dynamic pages are unbounded
  - Static pages over 20 billion?

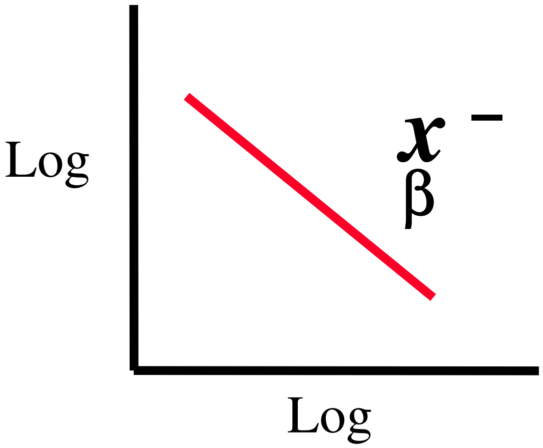
# Different Views on Data





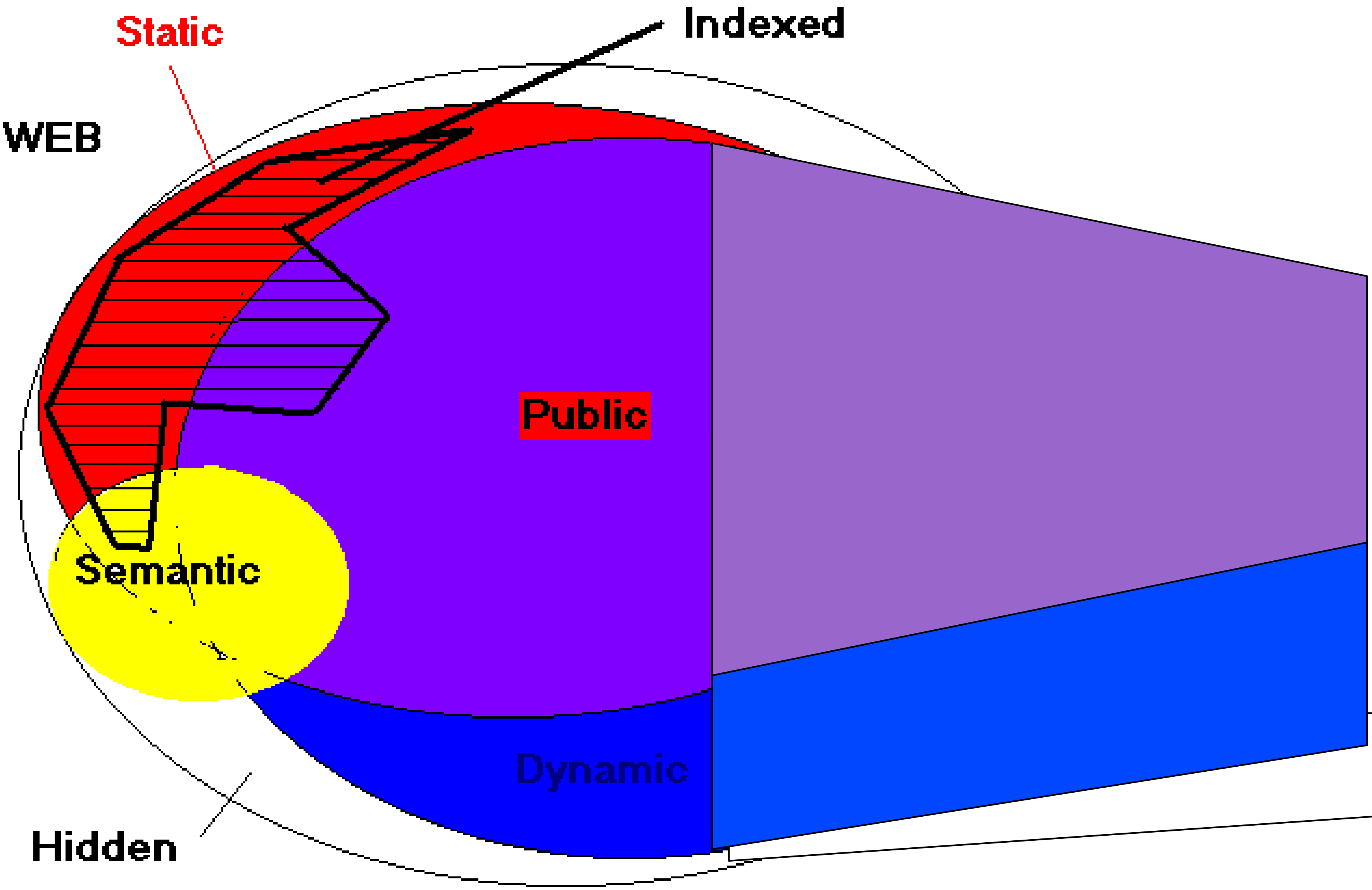
# The Web

- Largest public repository of *data* (more than 20 billion static pages?)
- Today, there are more than 170 million Web servers (Mar 08) and more than 540 million hosts (Jan 08)
- Well connected graph with out-link and in-link power law distributions



Self-similar &
Self-organizing

# Different facets of the Web



# Objectives of Web mining

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

# The Big challenge for search

Meet the diverse user needs
given
their poorly made queries
and
the size and heterogeneity of the Web corpus

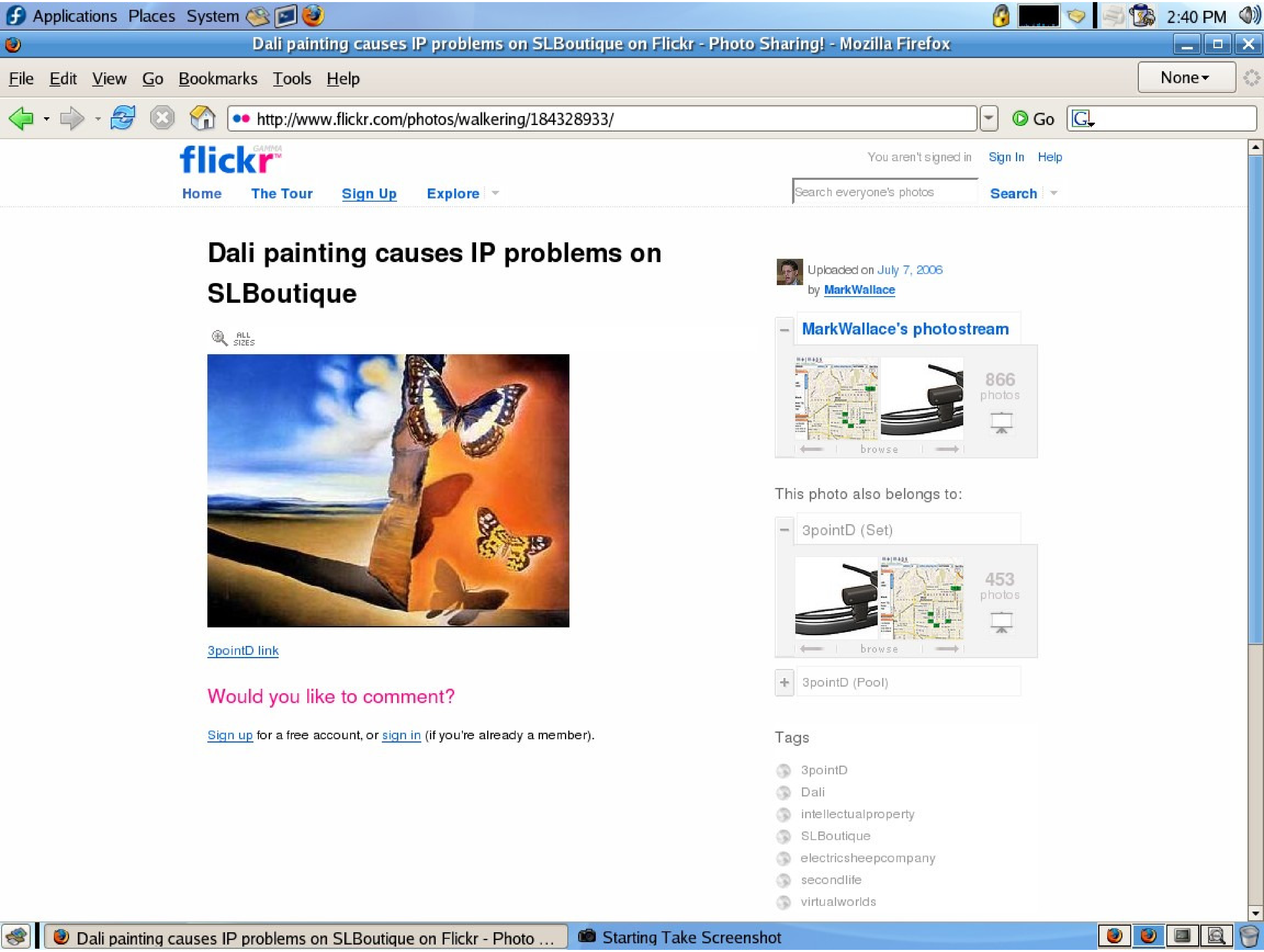
# Motivation for Web Mining

- 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

# The wisdom of crowds

- James Surowiecki, a *New Yorker* columnist, published this book in 2004
- Bottom line:

> *“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”.*



Applications Places System
2:40 PM
File Edit View Go Bookmarks Tools Help
None
Go
flickr GAMMA
You aren't signed in Sign In Help
Home The Tour Sign Up Explore
Search everyone's photos
Search
ALL SIZES
browse
Starting Take Screenshot

Dali painting causes IP problems on SLBoutique on Flickr - Photo Sharing! - Mozilla Firefox

http://www.flickr.com/photos/walkering/184328933/

## Dali painting causes IP problems on SLBoutique

3pointD link

Would you like to comment?

Sign up for a free account, or sign in (if you're already a member).

Uploaded on July 7, 2006
by MarkWallace

MarkWallace's photostream

866 photos

This photo also belongs to:

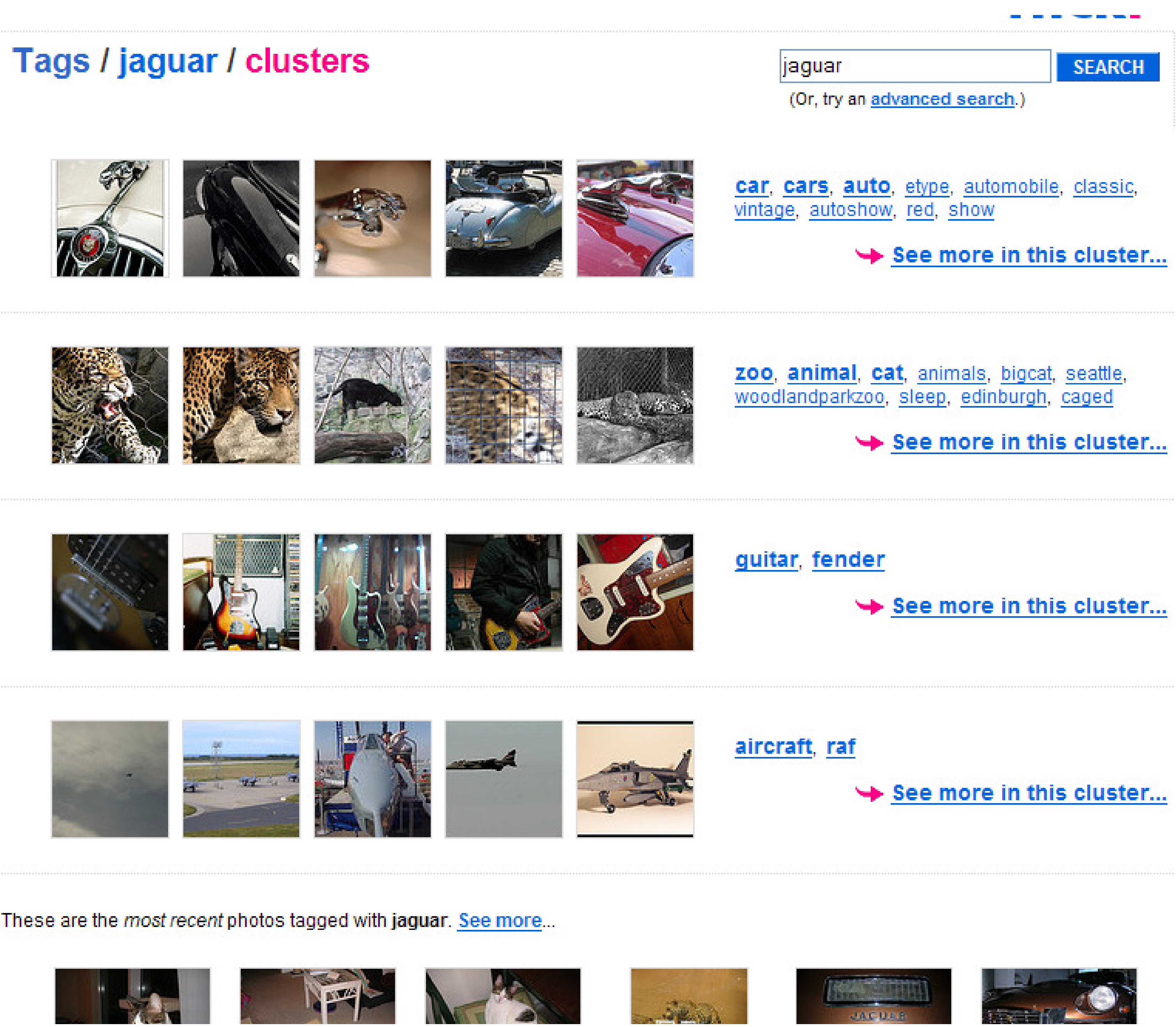
3pointD (Set)

453 photos

3pointD (Pool)

Tags

- 3pointD
- Dali
- intellectualproperty
- SLBoutique
- electricsheepcompany
- secondlife
- virtualworlds



# 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...

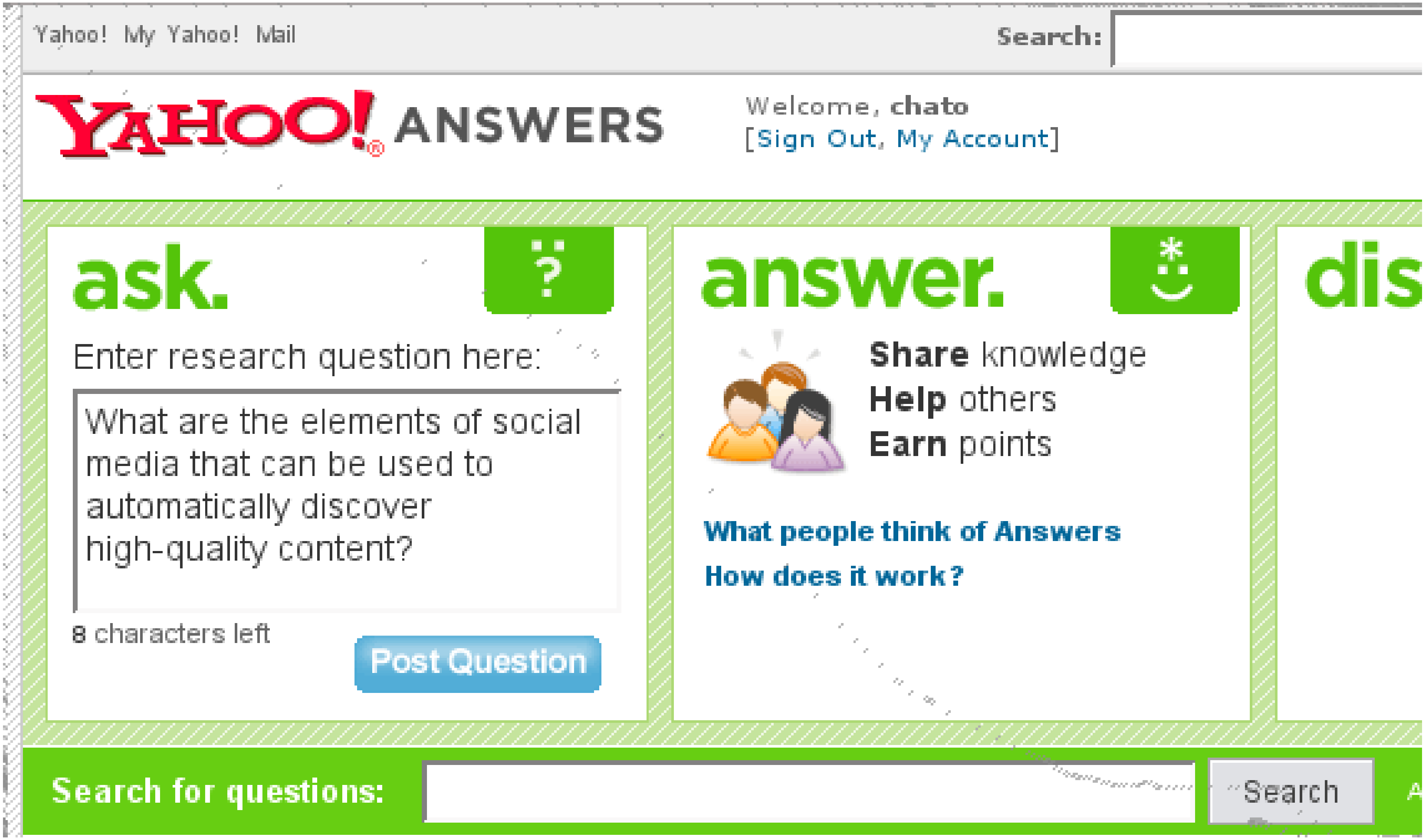
# The power of social media

- 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?

# 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!)

# Yahoo! answers



Welcome, chato
[Sign Out, My Account]

Enter research question here:

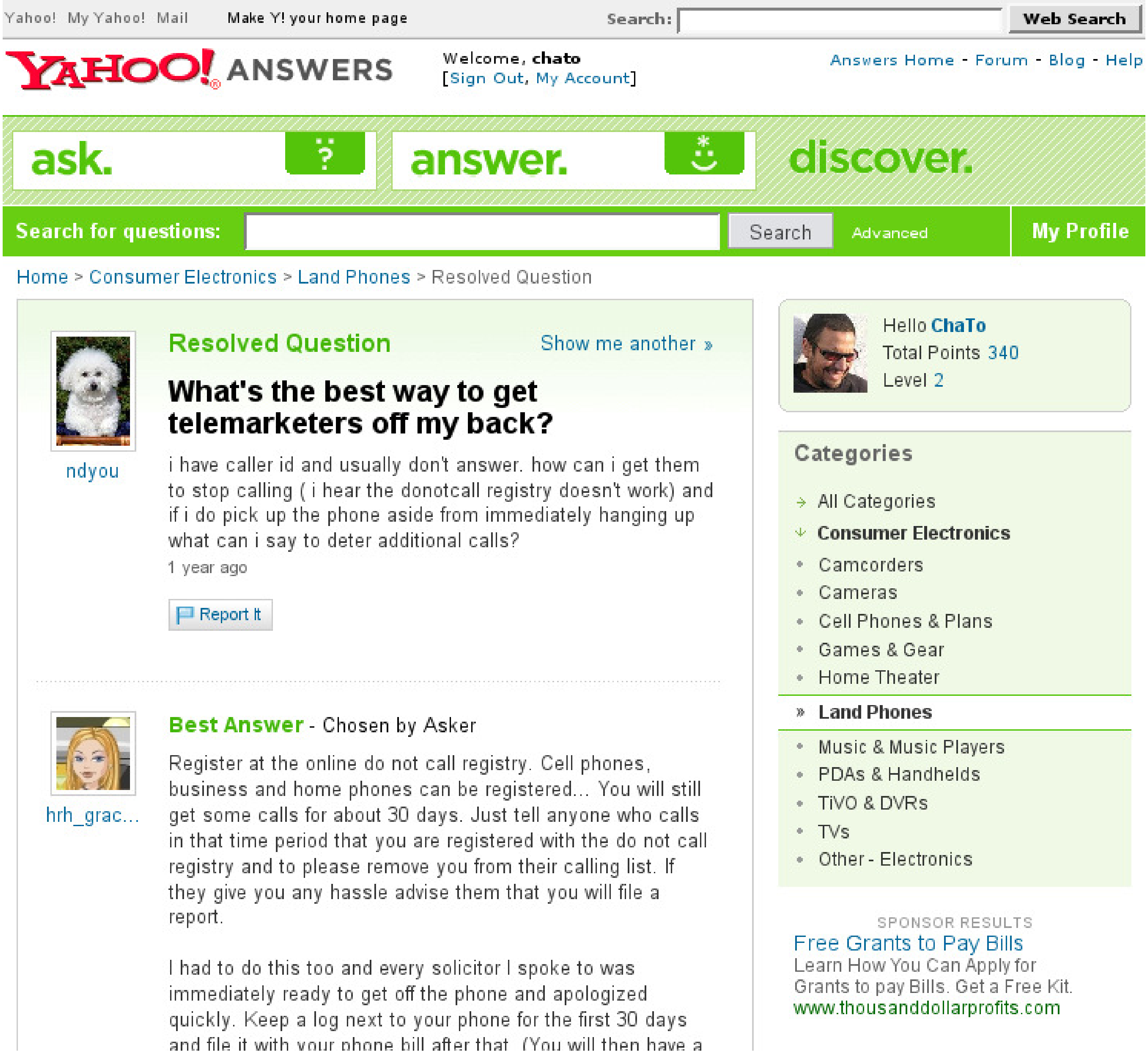
What are the elements of social media that can be used to automatically discover high-quality content?

8 characters left

**Share** knowledge
**Help** others
**Earn** points

**What people think of Answers**

**How does it work?**



## Resolved Question

### 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

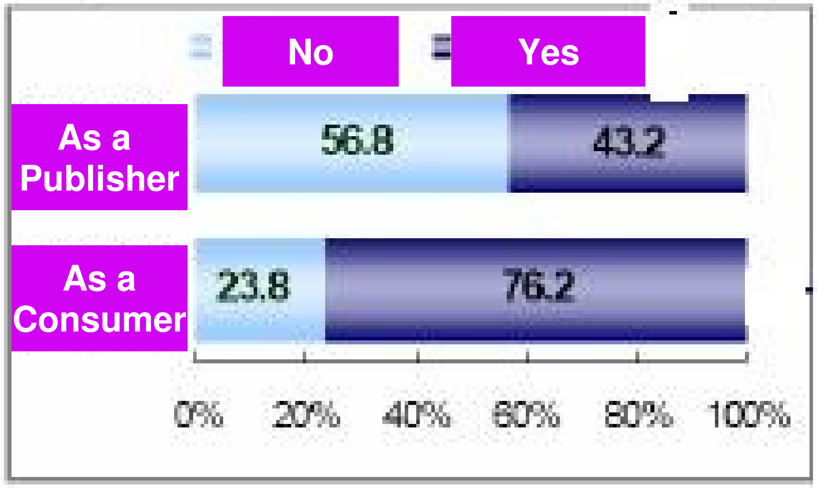
**Best Answer** - Chosen by Asker

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.

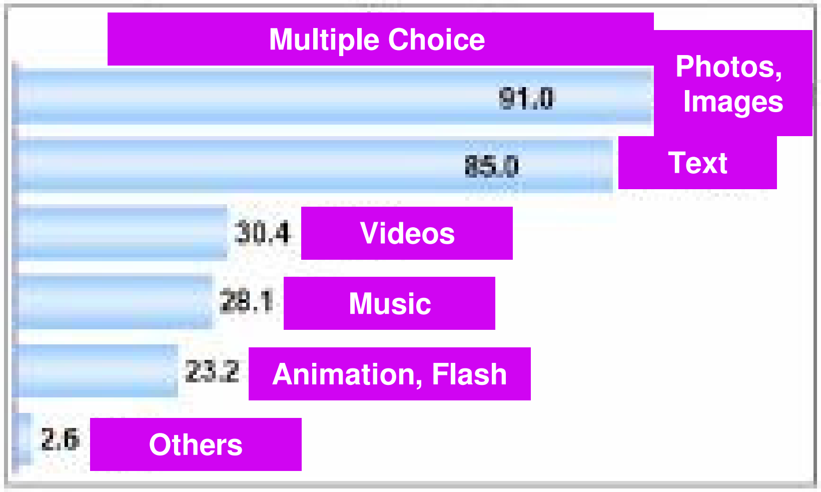
I had to do this too and every solicitor I spoke to was immediately ready to get off the phone and apologized quickly. Keep a log next to your phone for the first 30 days and file it with your phone bill after that. (You will then have a

# Internet UGC (User Generated Content)

Have you experienced UGC?

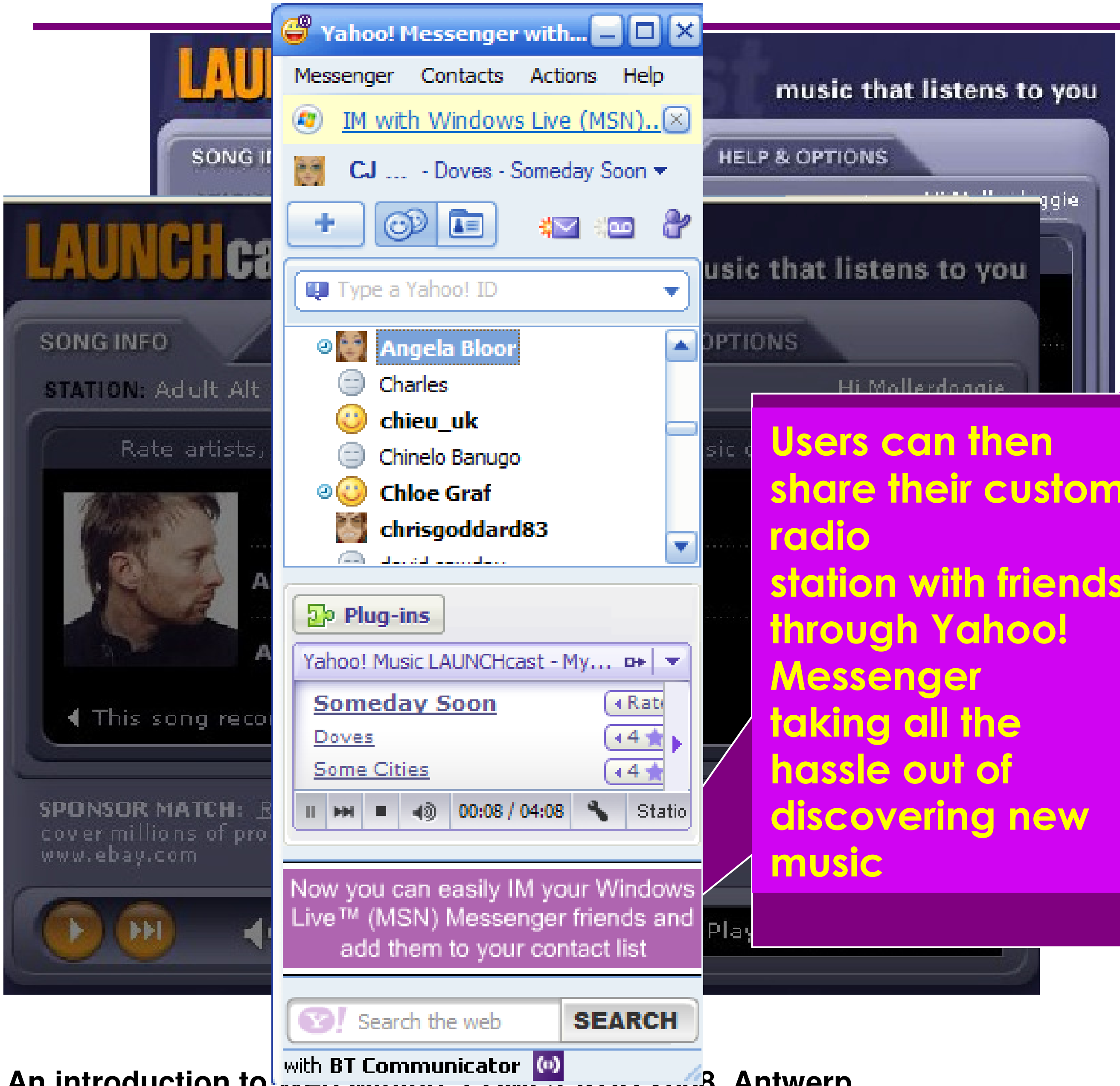


Types of Content



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

# Simple acts create value and opportunity



Now you can easily IM your Windows Live™ (MSN) Messenger friends and add them to your contact list

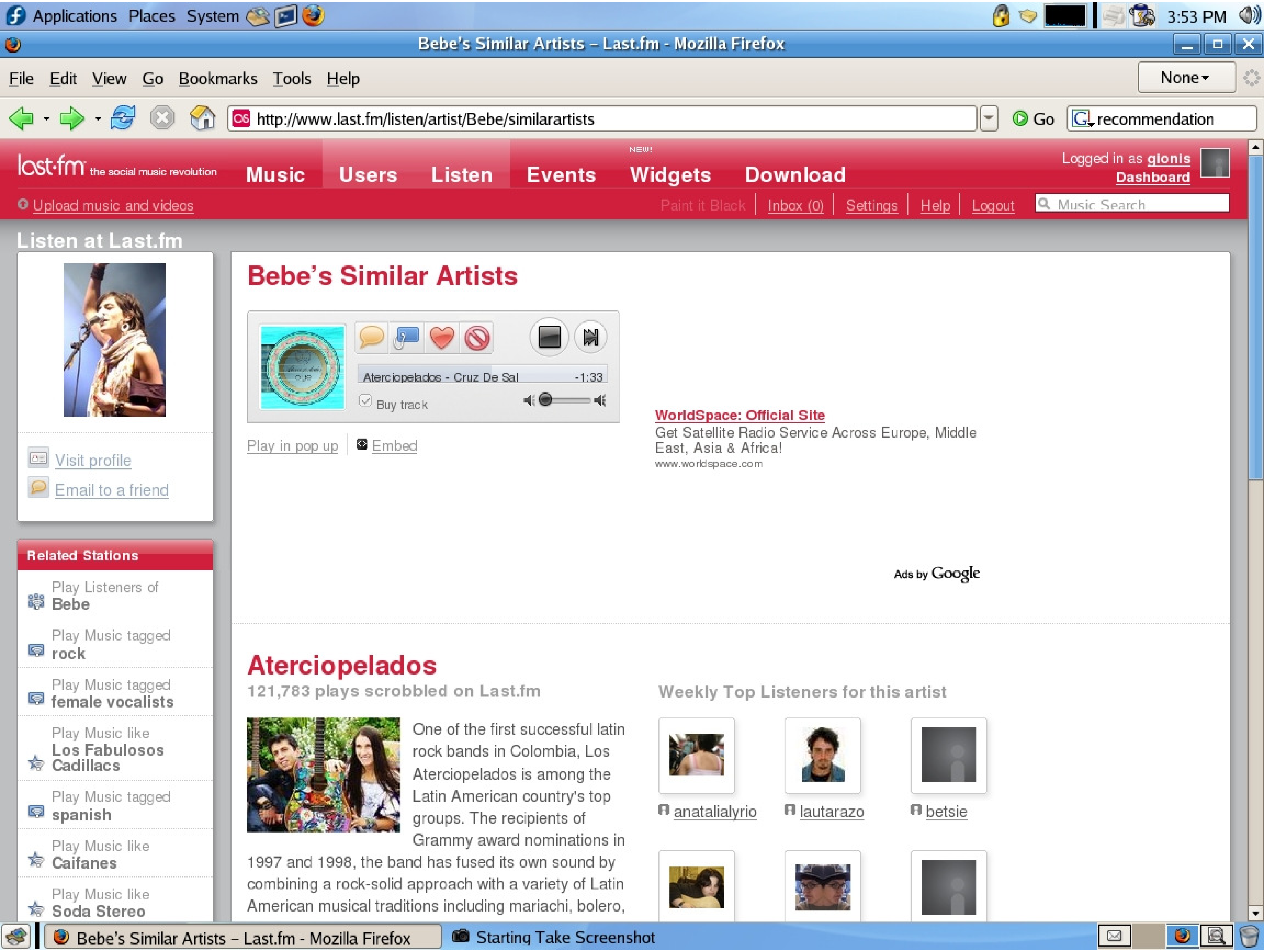
Using a system of user-assigned ratings, LAUNCHcast builds up a profile of preferences for each individual..

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



http://www.last.fm/listen/artist/Bebe/similarartists

## Listen at Last.fm

# Bebe's Similar Artists

Aterciopelados - Cruz De Sal -1:33

☑ Buy track

**WorldSpace: Official Site**
Get Satellite Radio Service Across Europe, Middle East, Asia & Africa!
www.worldspace.com

Ads by Google

### Related Stations

Play Listeners of **Bebe**

Play Music tagged **rock**

Play Music tagged **female vocalists**

Play Music like **Los Fabulosos Cadillacs**

Play Music tagged **spanish**

Play Music like **Caifanes**

Play Music like **Soda Stereo**

## Aterciopelados

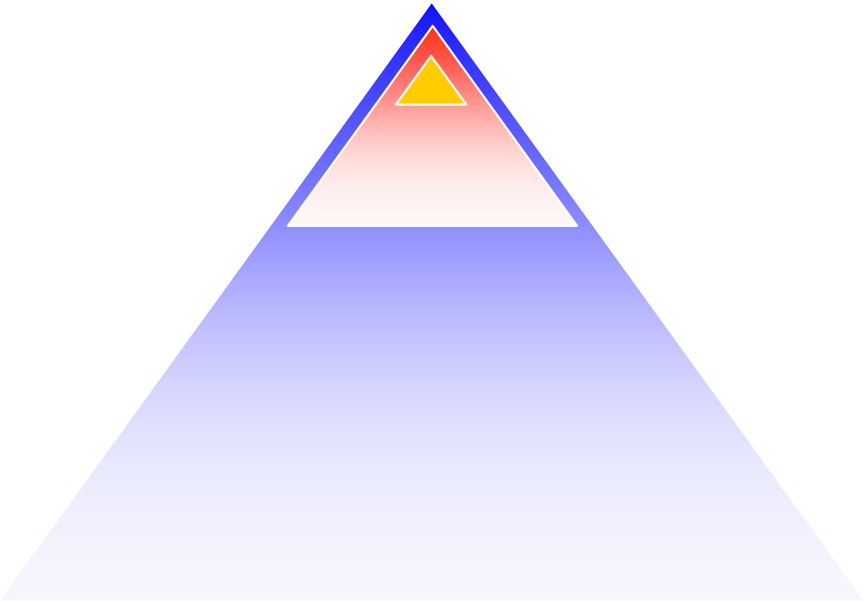
121,783 plays scrobbled on Last.fm

One of the first successful latin rock bands in Colombia, Los Aterciopelados is among the Latin American country's top groups. The recipients of Grammy award nominations in 1997 and 1998, the band has fused its own sound by combining a rock-solid approach with a variety of Latin American musical traditions including mariachi, bolero,

### Weekly Top Listeners for this artist

anatalialyrio
lautarazo
betsie

# Community dynamics



| | |
|---|---|
| **1** | **creators** |
| **10** | **synthesizers** |
| **100** | **consumers** |

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

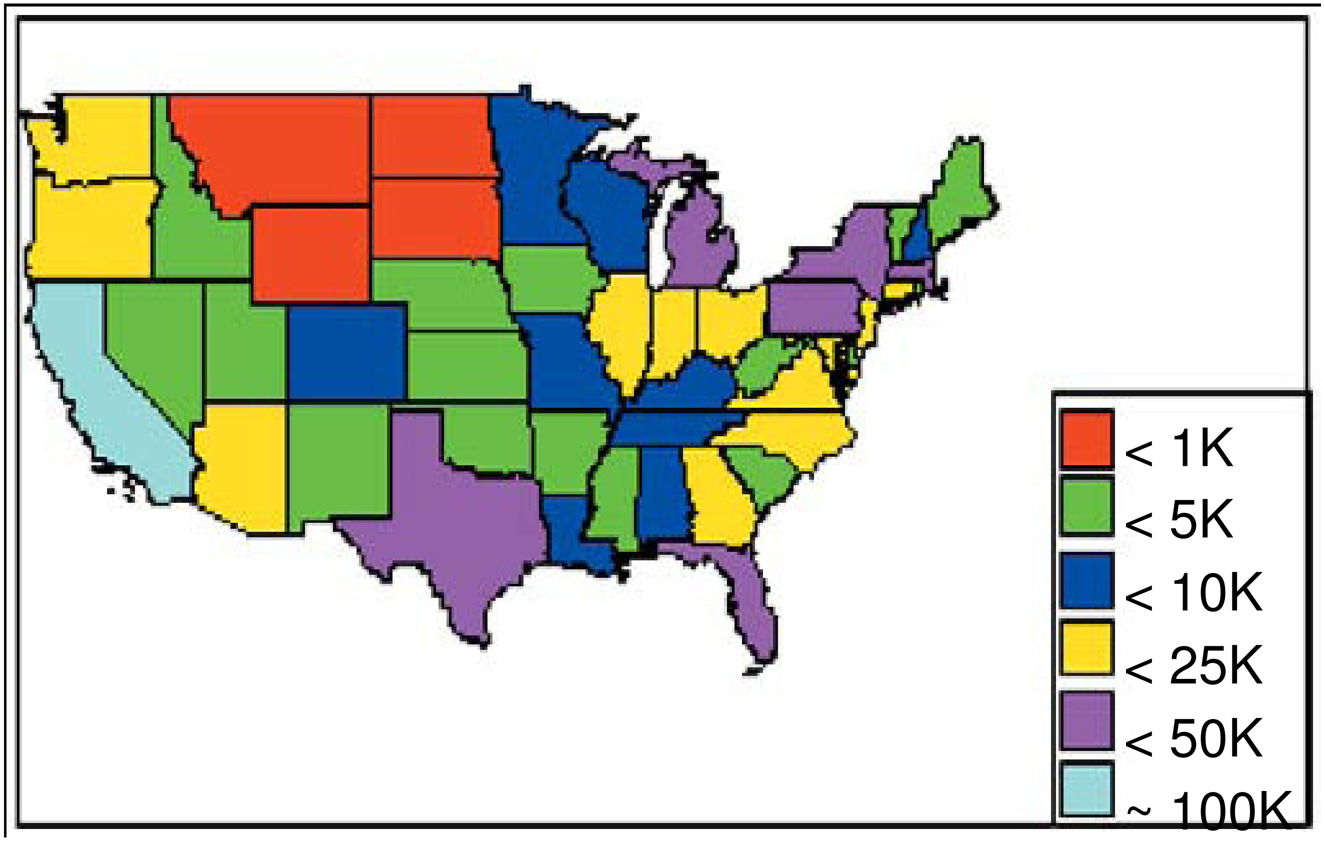
**Example: Launchcast**

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

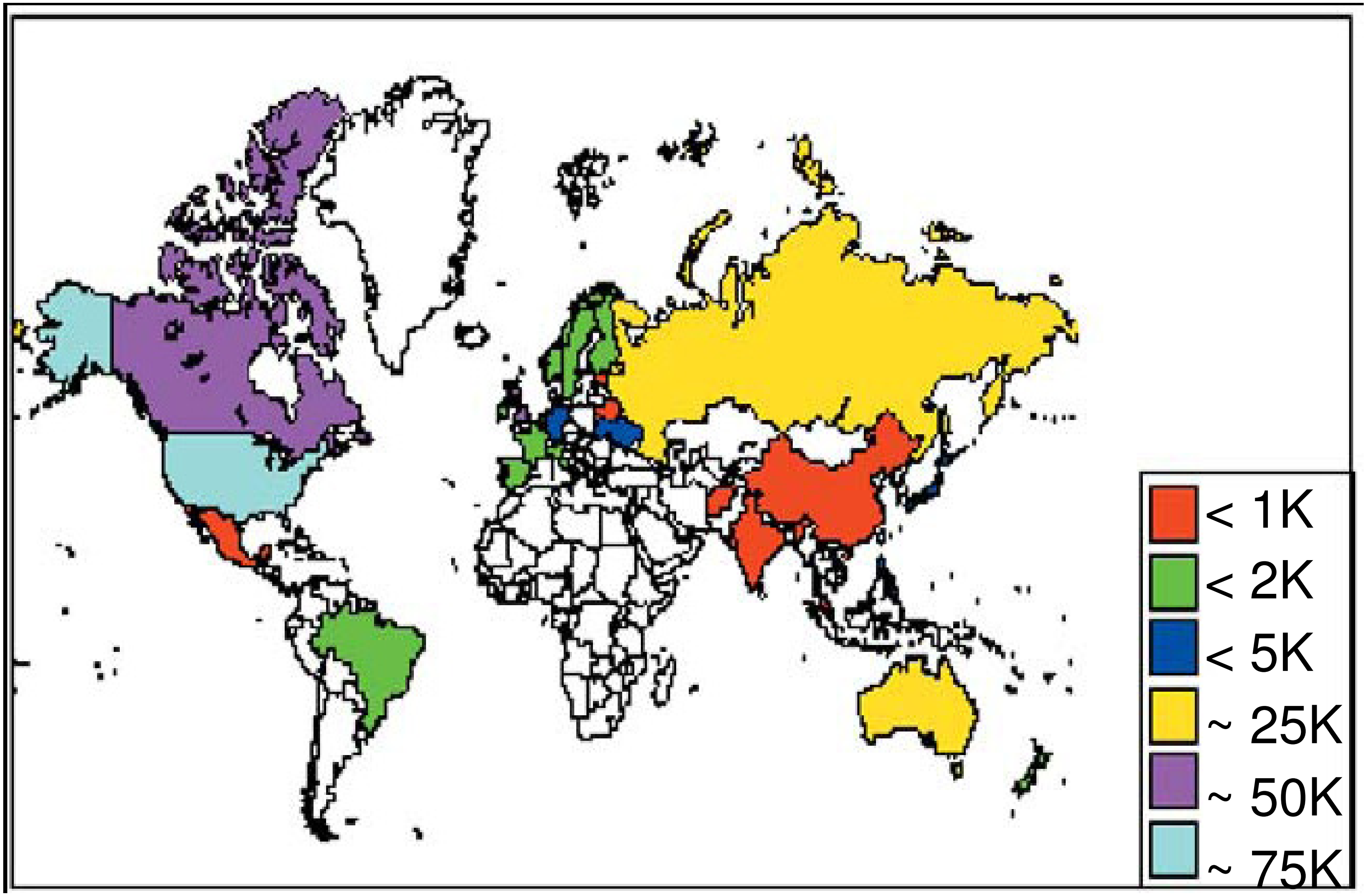
Listening itself implicitly creates a radio station…

# Community Geography:

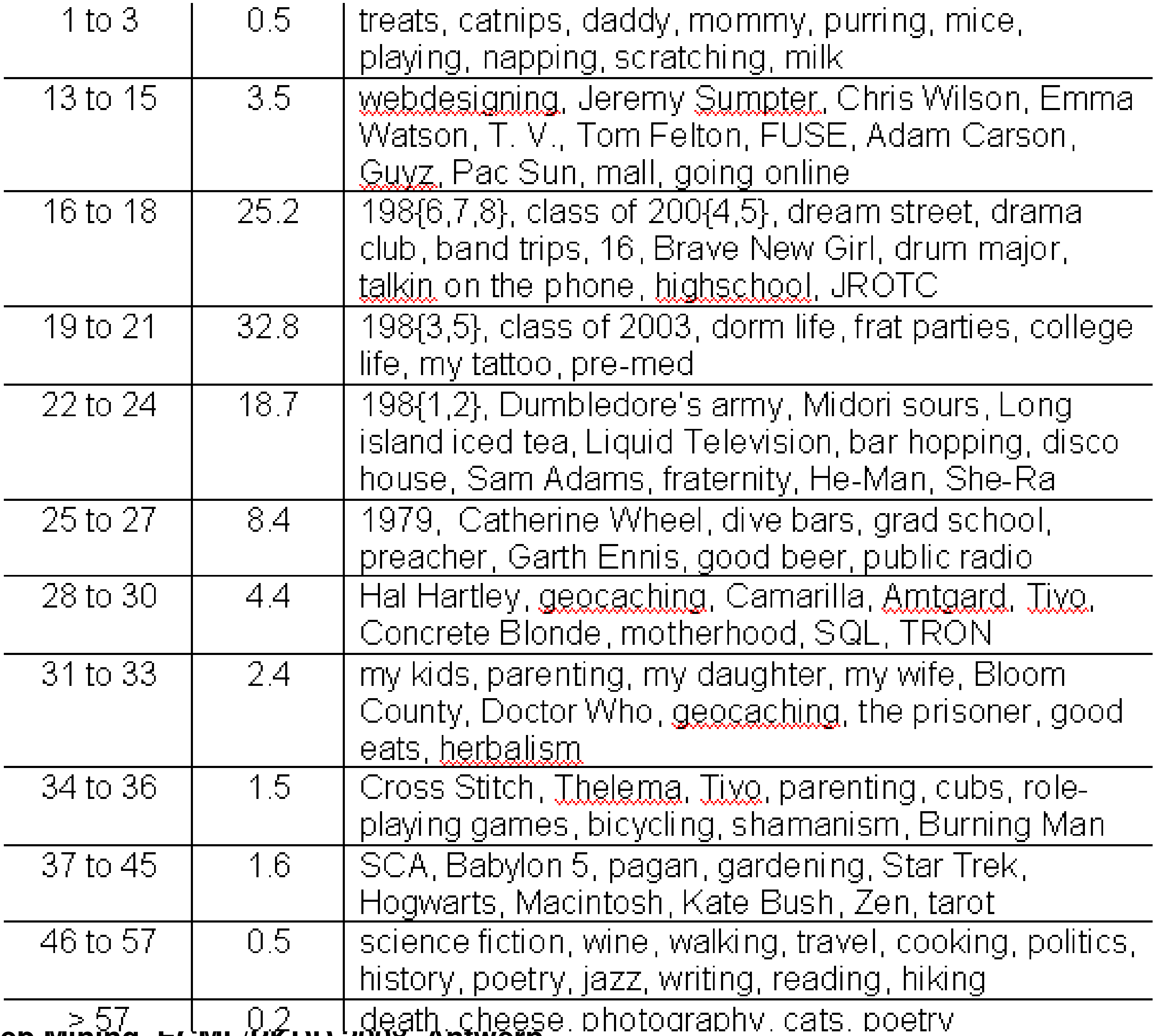
# LJ bloggers in US



# 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 |

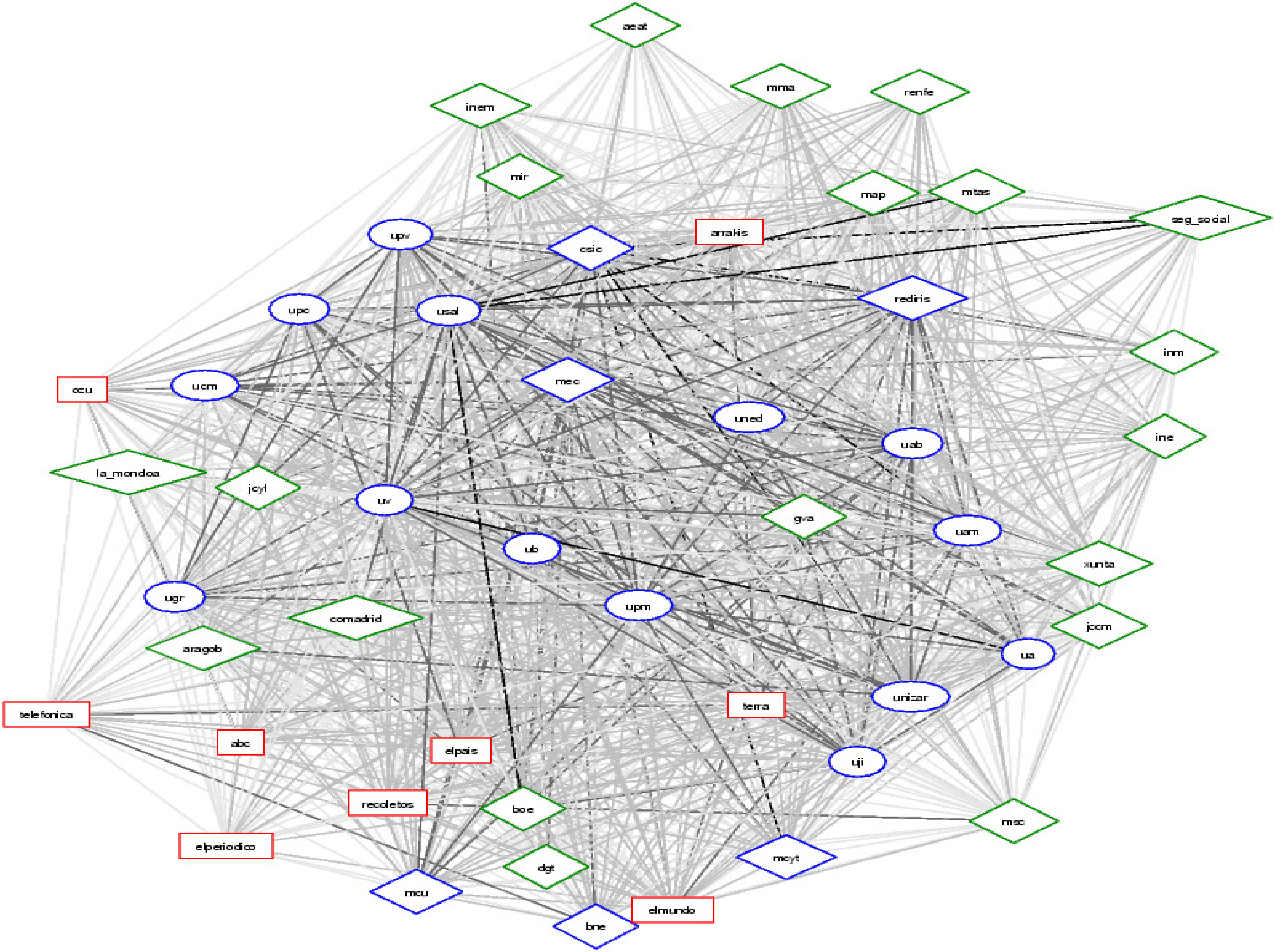
# What is in the Web?

# Web Mining

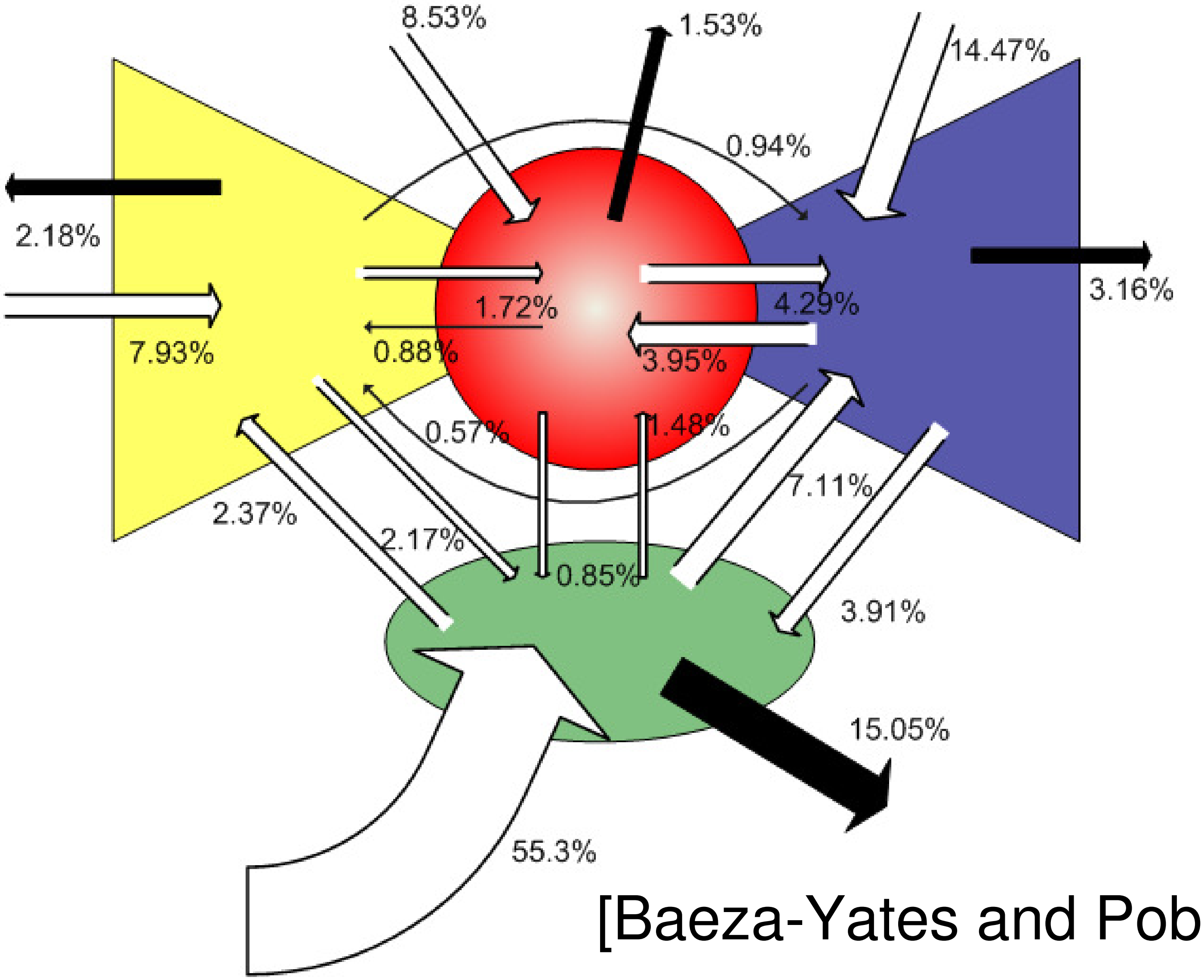
- **Content:** text & multimedia mining
- **Structure:** link analysis, graph mining
- **Usage:** log analysis, query mining
- **Relate all of the above**
  - – Web characterization
  - – Particular applications

# A Few Examples

- Web characterization of spain
- Link analysis
- Web dynamics
- User modeling

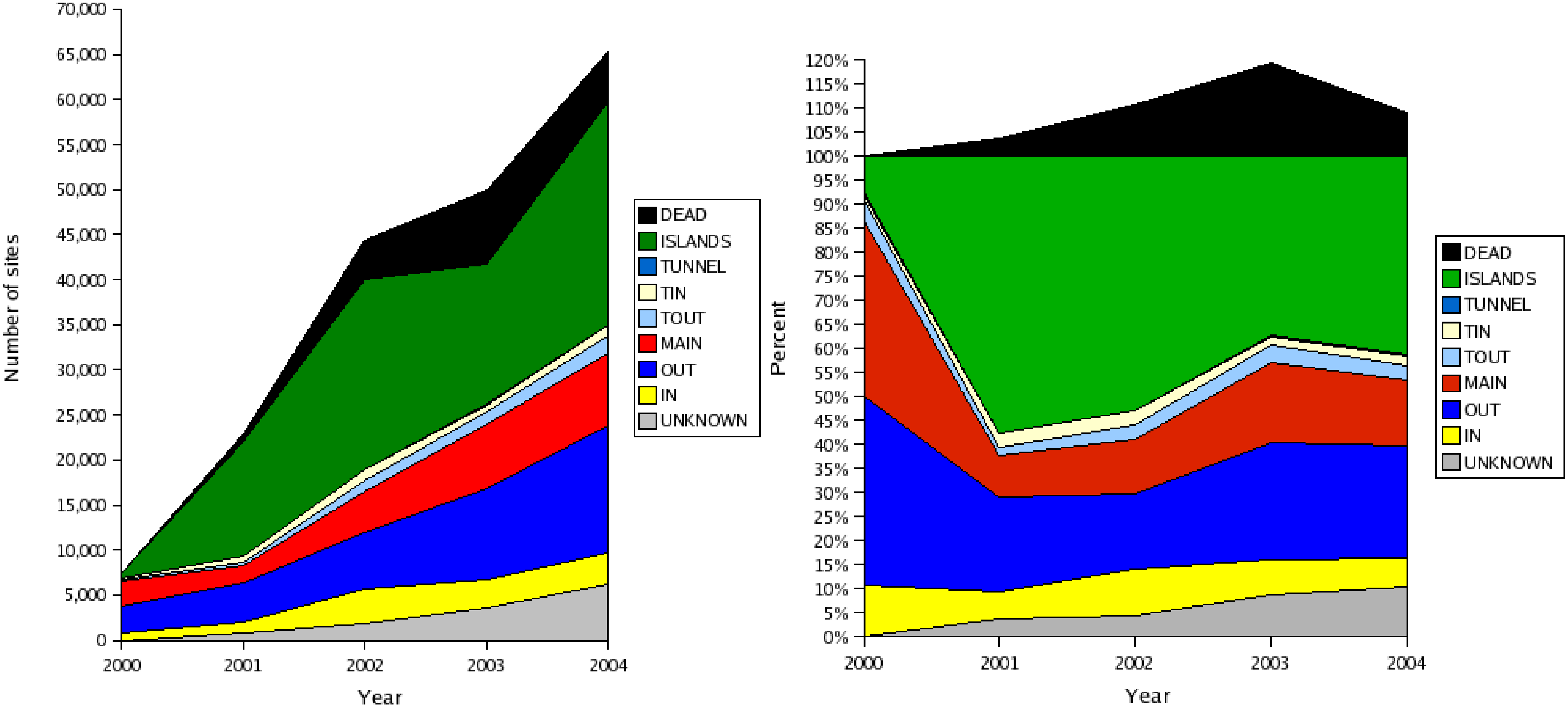

aeat
mma
renfe
inem
mir
map
mtas
seg_social
upv
csic
arrakis
upc
usal
rediris
inm
ocu
ucm
mec
uned
uab
ine
la_moncloa
jcyl
uv
gva
uam
ub
xunta
ugr
upm
comadrid
jccm
aragob
ua
unizar
terra
telefonica
abc
elpais
uji
recoletos
boe
msc
elperiodico
dgt
mcyt
mcu
elmundo
bne

# Structure Macro Dynamics

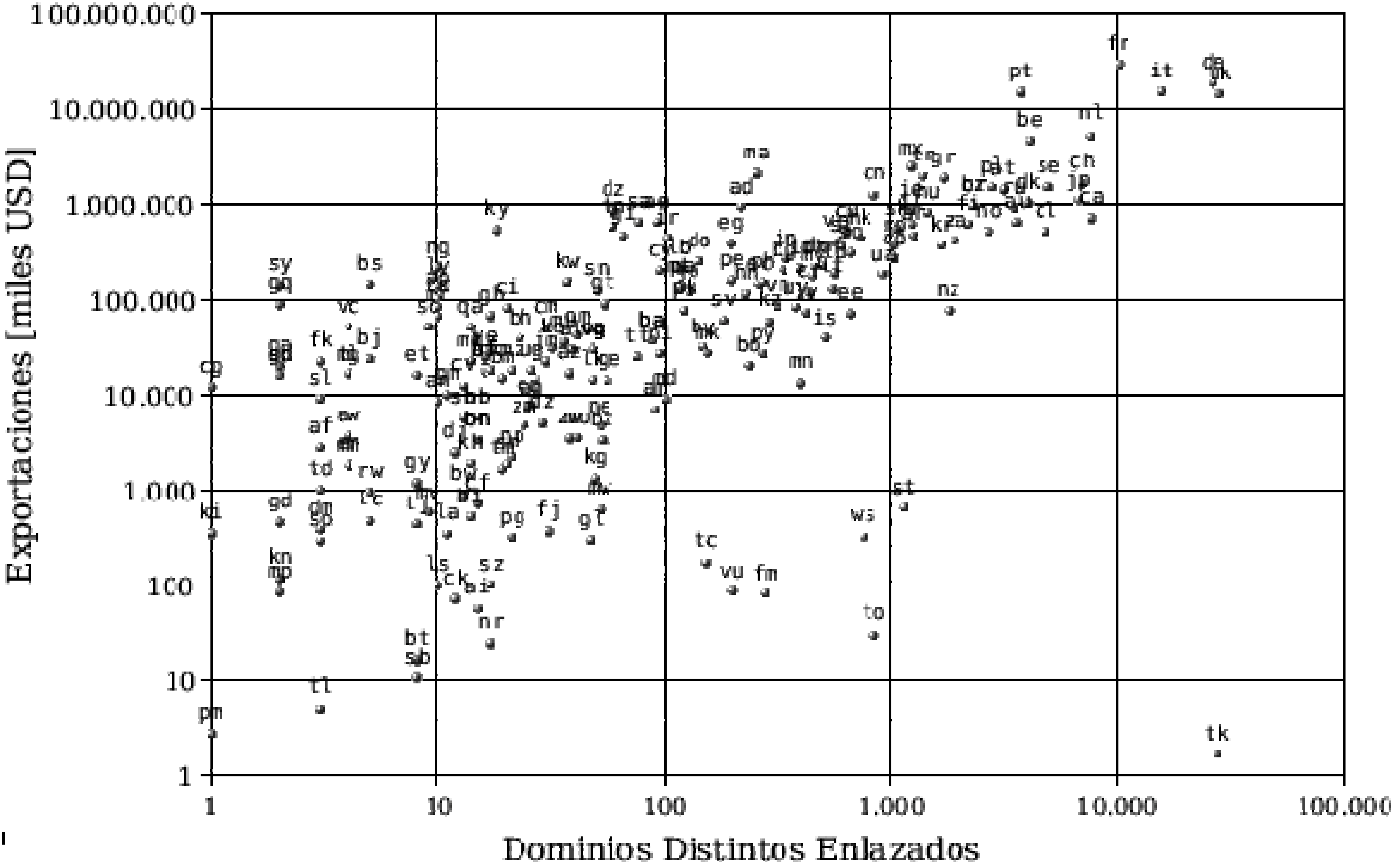


[Baeza-Yates and Poblete, 2006]

# Size Evolution

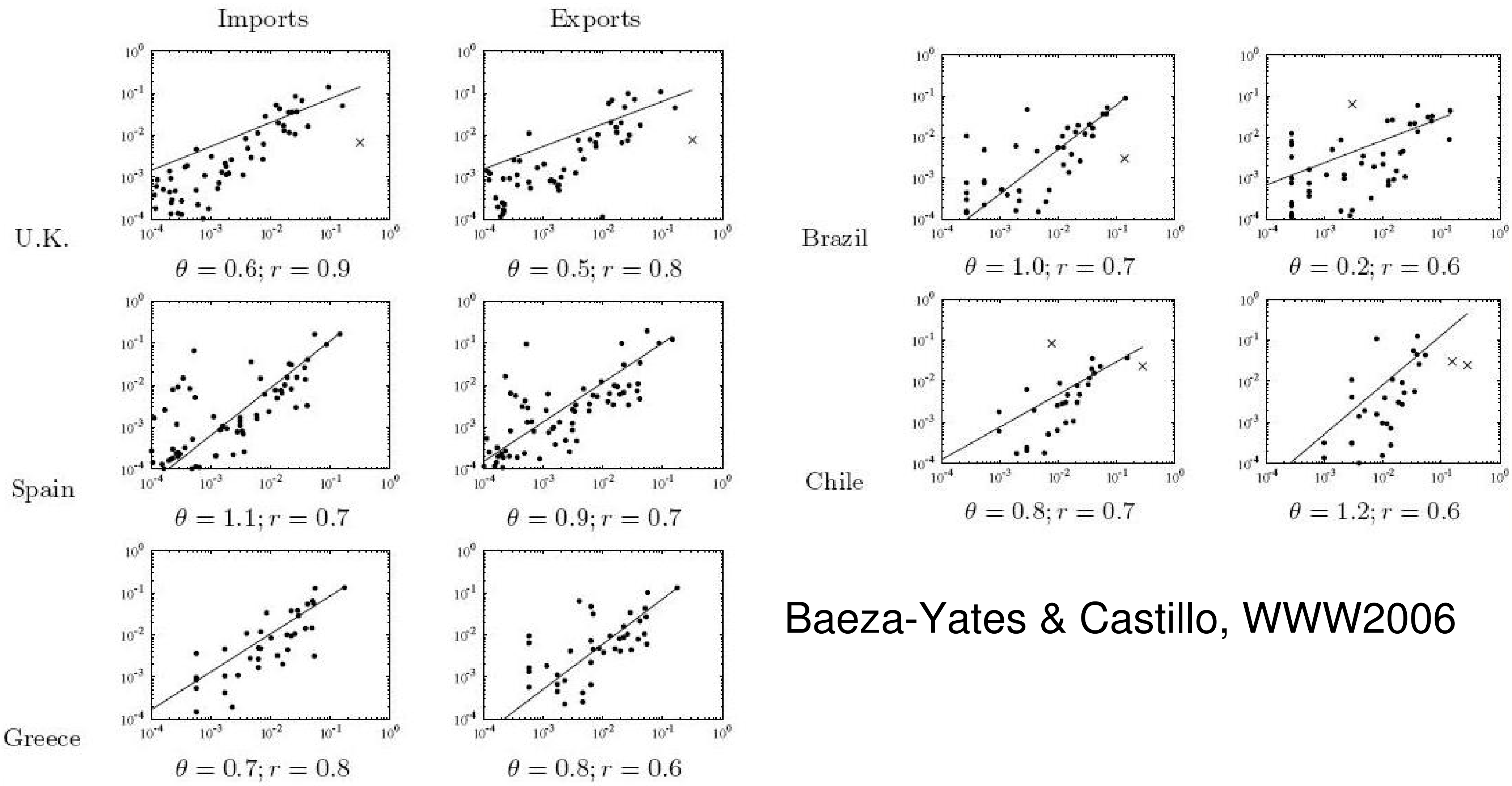


# Mirror of the Society



A

# Exports/Imports vs. Domain Links



U.K.: $\theta = 0.6; r = 0.9$ $\theta = 0.5; r = 0.8$

Spain: $\theta = 1.1; r = 0.7$ $\theta = 0.9; r = 0.7$

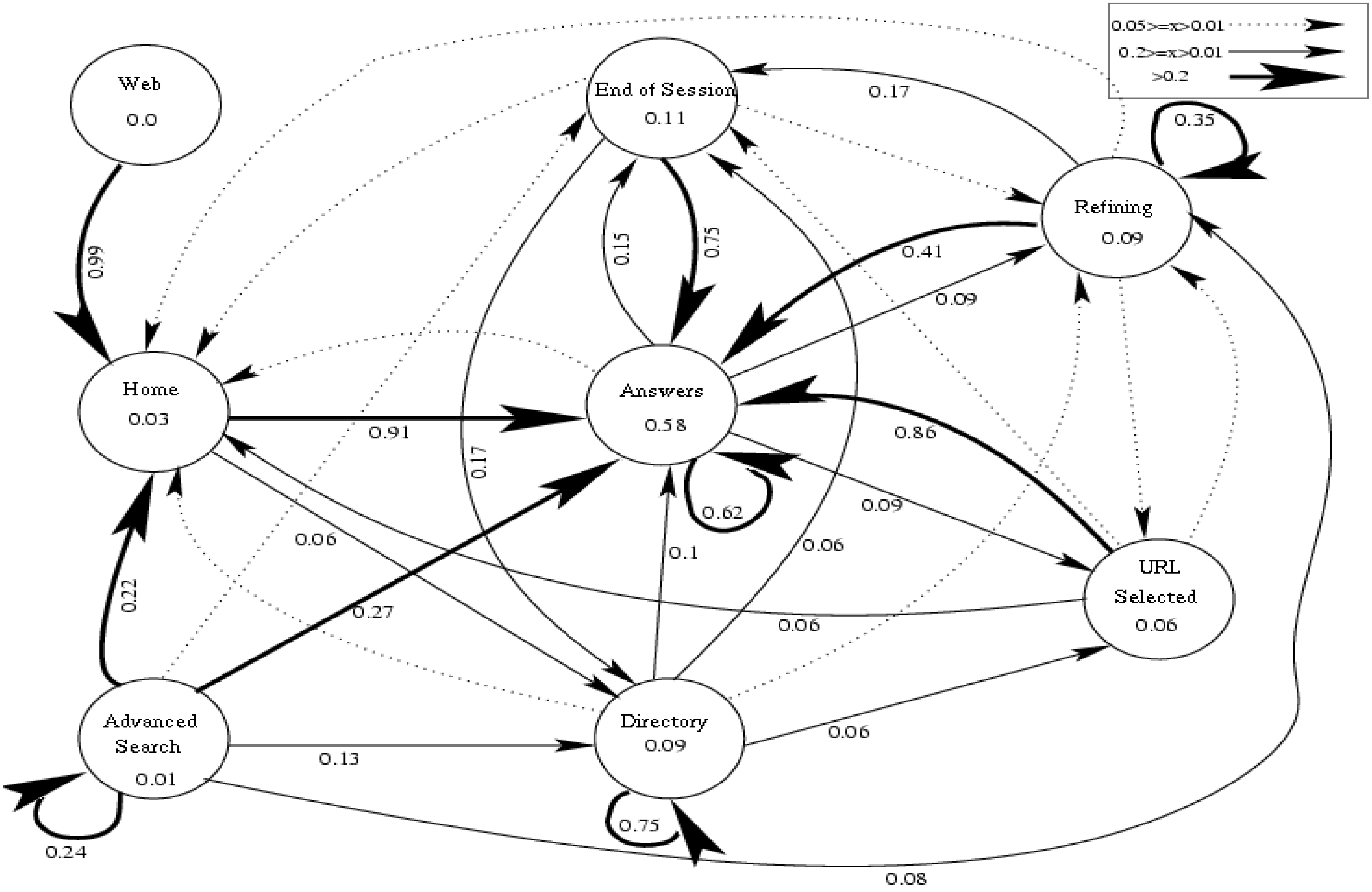
Greece: $\theta = 0.7; r = 0.8$ $\theta = 0.8; r = 0.6$

Brazil: $\theta = 1.0; r = 0.7$ $\theta = 0.2; r = 0.6$

Chile: $\theta = 0.8; r = 0.7$ $\theta = 1.2; r = 0.6$

Baeza-Yates & Castillo, WWW2006

# User modeling



# Data anonymization and data modeling

# Data anonymization

- The AOL query-log release
- American Online (AOL) query log released in August 2006
- Objective was to contribute to IR research
- Query log rough statistics
  - 20 million queries
  - 650 K users
  - from over 3 months
- Social security numbers, credit card numbers, driver license numbers, etc.
- Possible to uniquely identify many users by combining information from queries and yellow pages
- Big media scandal, big damage to AOL and the privacy of its users

# A typical query log

- Entries of the format:

  <cookie, query, rank, clickURL, timeStamp, IP, country,...>

# Anonymizing query logs

- [Adar 2007]
- Argue that anonymization is potentially possible

- Two main techniques:
  - Eliminate infrequent queries
  - Splitting personalities

- Additionally:
  - Eliminate identifying information (SSN, credit card numbers, etc.)

# Anonymizing query logs

- Eliminate infrequent queries:

- Keep only queries generated by a large number of users
- Computationally possible using counters
- How to do it on-the-fly?

# Online elimination of infrequent queries

- Background: How to split a secret among n people so that every coalition of *k* persons can access the secret?
- Answer: Let the secret be the coefficients of a *(k-1)*-degree polynomial $f(x) = a_{k-1}x^{k-1} + \ldots + a_1 x + a_0$
- For the *i*-th person, select a number $x_i$, and give to the person the pair $(x_i, f(x_i))$
- Any k persons can cooperate and recover the polynomial, while no *k-1* persons can recover it

# Online elimination of infrequent queries

- Straightforward application in eliminating infrequent queries

- A query $q$ is decoded as a $(k-1)$-degree polynomial $f_q$
- For a person $u_i$ who makes the query $q$, print $(u_i, f_q(u_i))$
- If $k$ or more people type the query $q$, it is possible to decrypt $q$!

# Split personalities

- Split the queries of the same user into sessions
- E.g., queries about food recipes, sport results, buying books, music, etc.
- Assign each of those sessions to a different virtual user
- Released query log can be still useful for many applications
- More difficult to identify users by combining queries
- Finding similar queries and finding query sessions is quite hard problem

# Anonymizing query logs: negative resuls

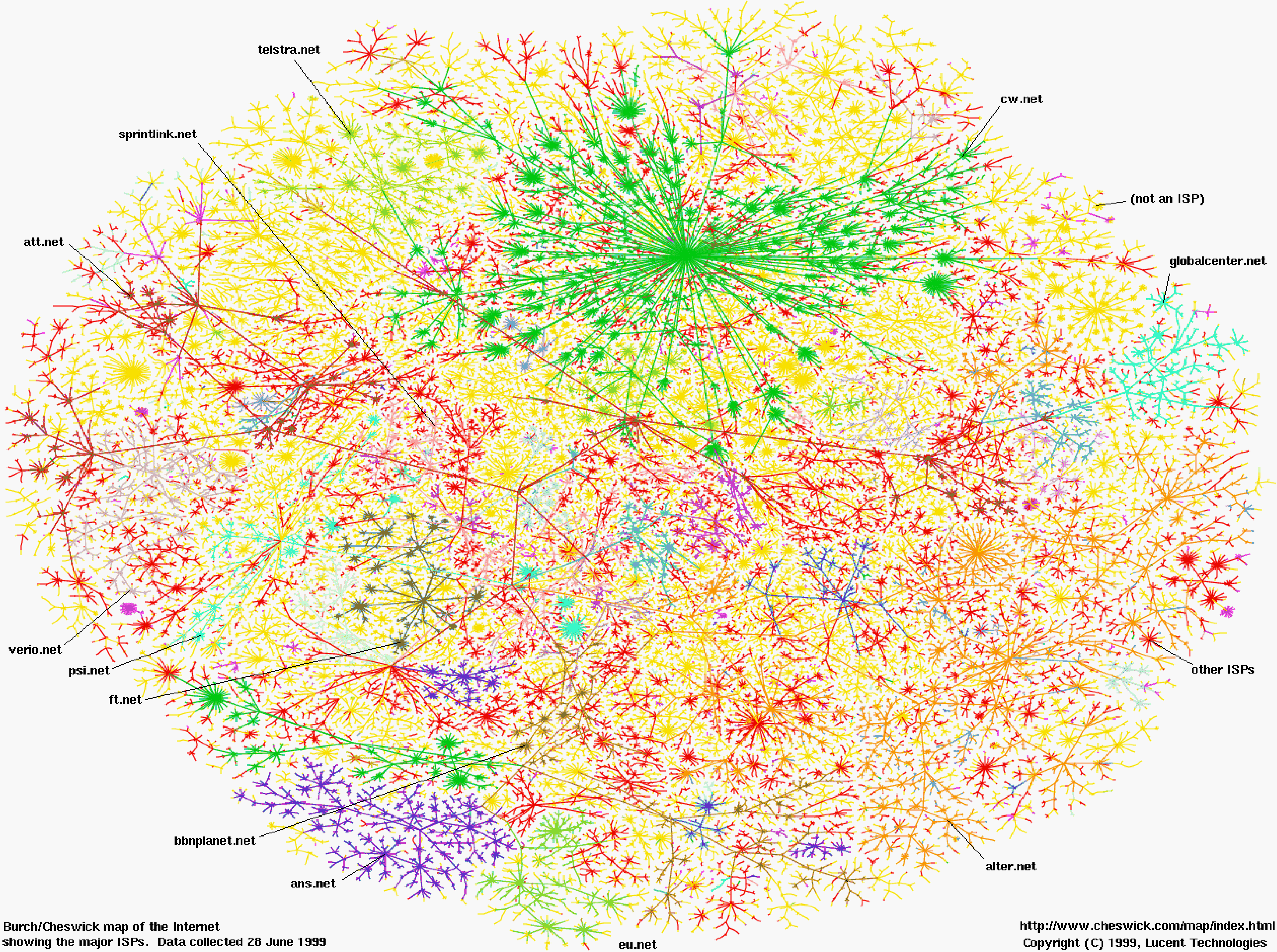
- [Kumar et al., 2007]
- Anonymization via token-based hashing:
- The query is split into terms and each term is hashed to a token
- Co-occurrence analysis and frequency analysis can be used to reveal the query terms
- Assume access to an unencrypted query log
- Query term statistics remain constant across different query logs
- Provide practical graph-matching algorithms and analysis of real query logs

# Anonymizing query logs: negative resuls

- [Jones et al., 2007]
- Simple classifiers can be used on the query log to identify gender, age, and location of the user issuing the queries
- Map a sequence of queries into a set of candidate users that is 300-600 times smaller than random chance would allow
- Identify person attacks: identify information for an acquaintance from speculated queries
- Releasing query logs has severe privacy risks

# Data statistics and data modeling

- Graph structures
- Degree distribution
- Community structure
- Diameter and other properties



Burch/Cheswick map of the Internet
showing the major ISPs.  Data collected 28 June 1999

http://www.cheswick.com/map/index.html

# Degree distribution

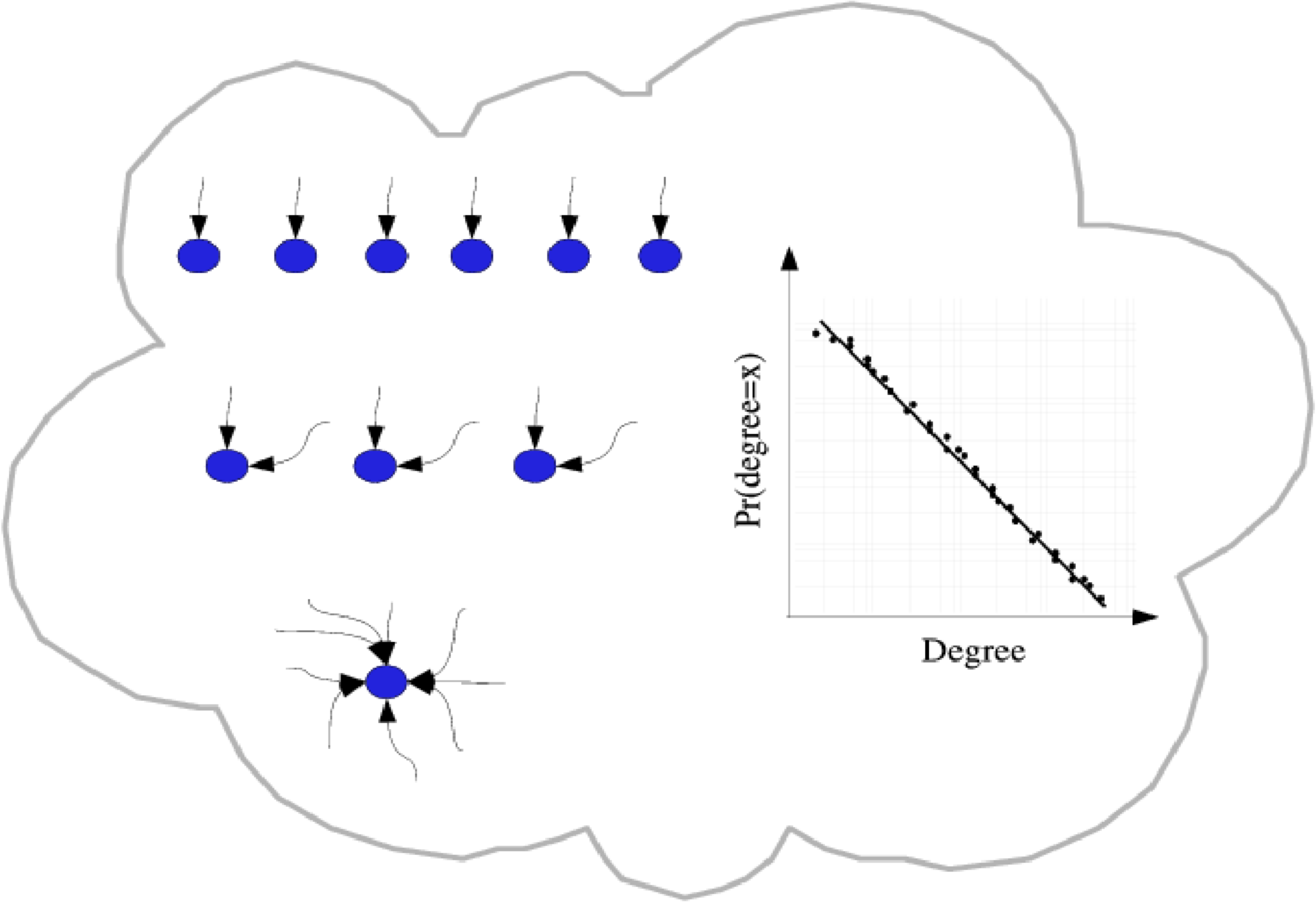
- Consider a graph $G=(V,E)$
- $C_k$ the number of vertices $u$ with degree $d(u) = k$

$$C_k = c\, k^{-a} \quad \text{with} \quad a>0$$

$$\log(C_k) = \log(c) - a \log(k)$$

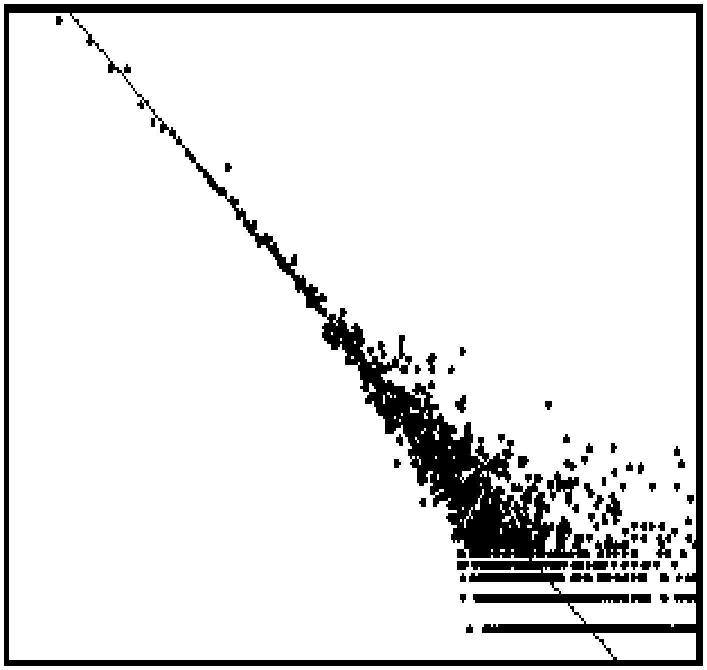
- So, plotting $\log(C_k)$ versus $\log(k)$ gives a straight line with slope $-a$
- Heavy-tail distribution: there is a non-negligible fraction of nodes that has very high degree (hubs)
- Scale-free: no characteristic scale, average is not informative

# Degree distribution

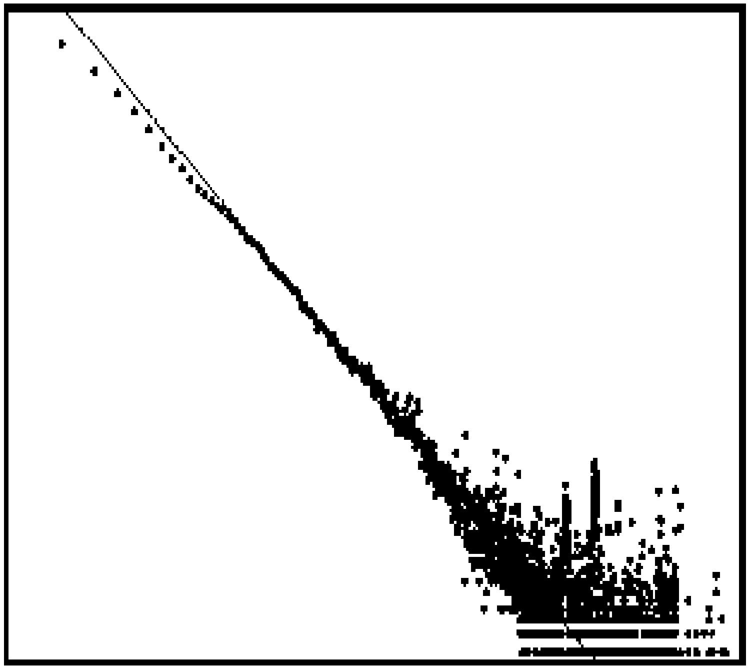


# Degree distribution

In-degree distributions of web graphs within national domains



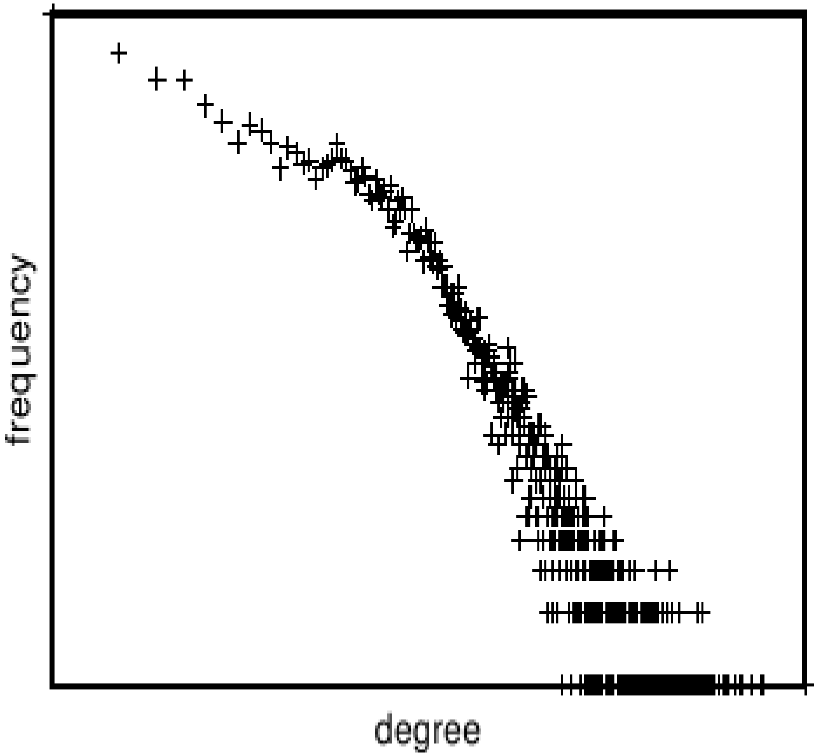
Greece



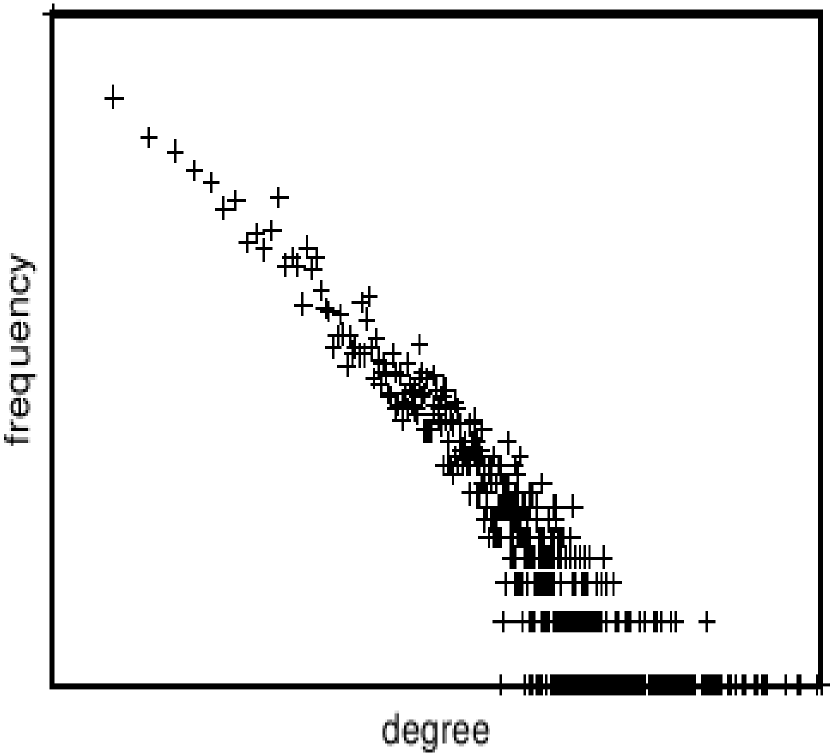
Spain

# 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:

$$C_2 = 1/n\ Sum_i\ C_i$$

- 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

# Other properties

- Degree correlations
- Distribution of sizes of connected components
- Resilience
- Eigenvalues
- Distribution of motifs
- ... all very different than predicted for random graphs

- Properties of evolving graphs [Leskovec et al., 05]
  - Densification power law
  - Diameter is shrinking

# Power-law distributions

- *“A brief history of generative models for power laws and log-normal distributions”* [Mitzenmacher, 04]
- A random variable $X$ has power-law distribution, if

$$Pr[X>x] = \quad cx^{-a} \text{ for } c > 0 \text{ and a } > 0$$

- A random variable $X$ has Pareto distribution, if

$$Pr[X>x] = (x/k)^{-a} \text{ for k } > 0,\ a > 0, \text{ and } X > k$$

- On a log-log plot straight line with slope -a

# A process that generates power-law

- Preferential attachment
- The main idea is that "the rich get richer"
  - First studied by [Yule, 1925] to suggest a model of why the number of species in genera follows a power-law
  - Generalized by [Simon, 1955]
    - applications in distribution of word frequencies, population of cities, income, etc.
  - Revisited in the 90s as a basis for Web-graph models [Barabasi and Albert, 1999, Broder et al., 2000, Kleinberg et al., 1999]

# Preferential attachement

- The basic theme:
  - Start with a single vertex, with a link to itself
  - At each time step a new vertex $u$ appears with out-degree $1$ and gets connected to an existing vertex $v$
  - With probability $p < 1$, vertex $v$ is chosen uniformly at random
  - With probability $1$–p, vertex $v$ is chosen with probability proportional to its degree
  - Process leads to power law for the in-degree distribution, with exponent $(2-p)/(1-p)$

# Log-normal distribution

- Random variable X has log-normal distribution, if Y=log(X) has normal distribution
- Always finite mean and variance
- But also appears as a straight line on a log-log plot (for small values of x)

- Multiplicative processes tend to give log-normal distributions:
  - The product of two log-normally distributed independent random variables follows a log-normal distribution

# Power law or log-normal?

- Distribution of income
- Start with some income $X_0$
- At time t, with probability 1/3 double the income, with probability 2/3 cut income at half
- Then income distribution is log-normal (multiplicative process)

- But... assume a “reflective barrier”:
  - At $X_0$ maintain same income with probability 2/3

- ... a power law!

# Usage mining

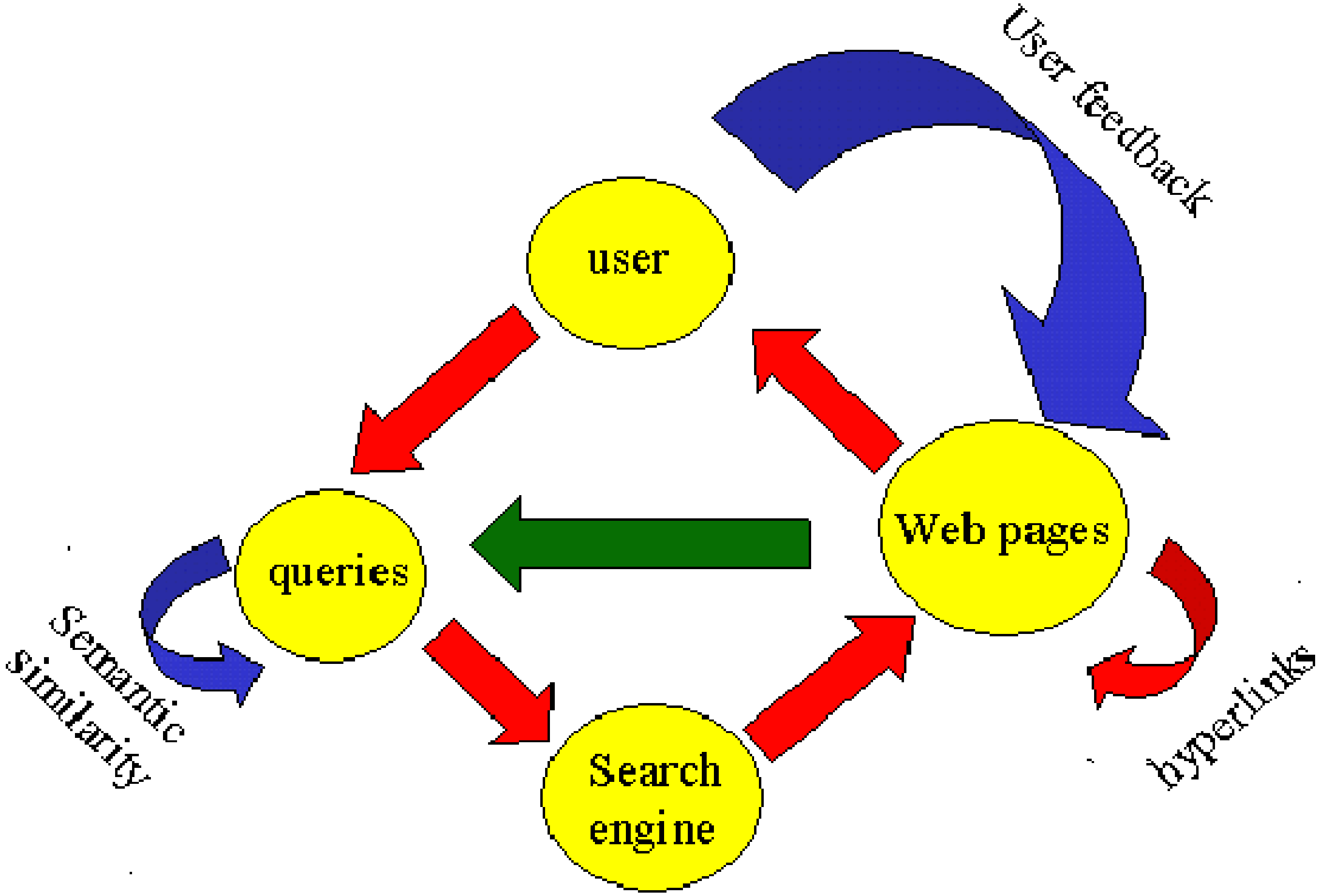
- Query log analysis

# 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?



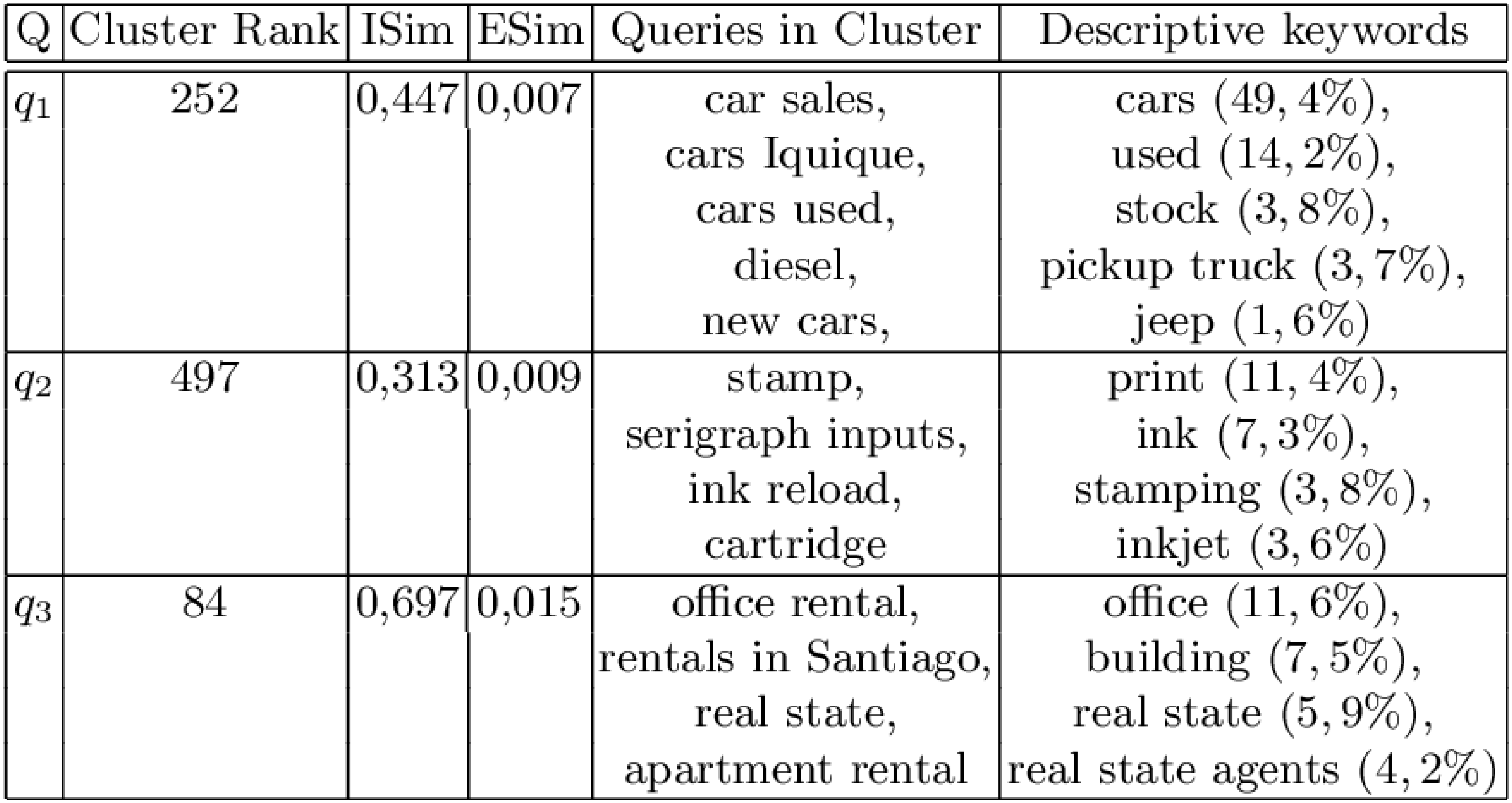
# Clustering queries

- 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,<br>cars Iquique,<br>cars used,<br>diesel,<br>new cars, | cars $(49, 4\%)$,<br>used $(14, 2\%)$,<br>stock $(3, 8\%)$,<br>pickup truck $(3, 7\%)$,<br>jeep $(1, 6\%)$ |
| $q_2$ | 497 | 0,313 | 0,009 | stamp,<br>serigraph inputs,<br>ink reload,<br>cartridge | print $(11, 4\%)$,<br>ink $(7, 3\%)$,<br>stamping $(3, 8\%)$,<br>inkjet $(3, 6\%)$ |
| $q_3$ | 84 | 0,697 | 0,015 | office rental,<br>rentals in Santiago,<br>real state,<br>apartment rental | office $(11, 6\%)$,<br>building $(7, 5\%)$,<br>real state $(5, 9\%)$,<br>real state agents $(4, 2\%)$ |

# Using the Clusters

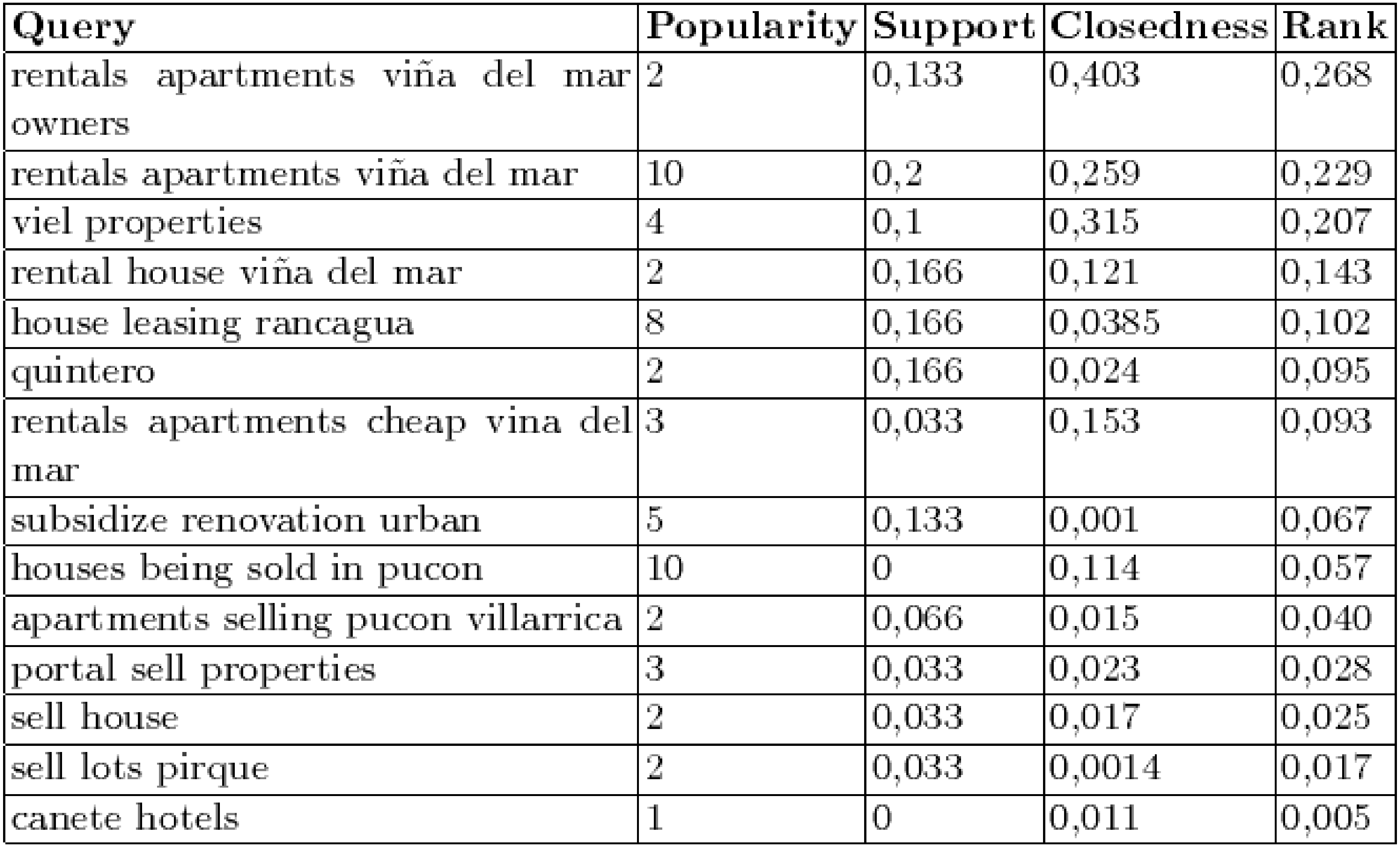
- Improved ranking
- Word classification
  - Synonyms & related terms are in the same cluster
  - Homonyms (polysemy) are in different clusters
- Query recommendation (ranking queries!)
  - Real queries, not query expansion

**Baeza-Yates, Hurtado & Mendoza**

**Journal of ASIST 2007**

$$\mathtt{Rank}(q) = \gamma \times \mathtt{Sup}(q, q_{ini}) + (1 - \gamma) \times \mathtt{Clos}(q)$$

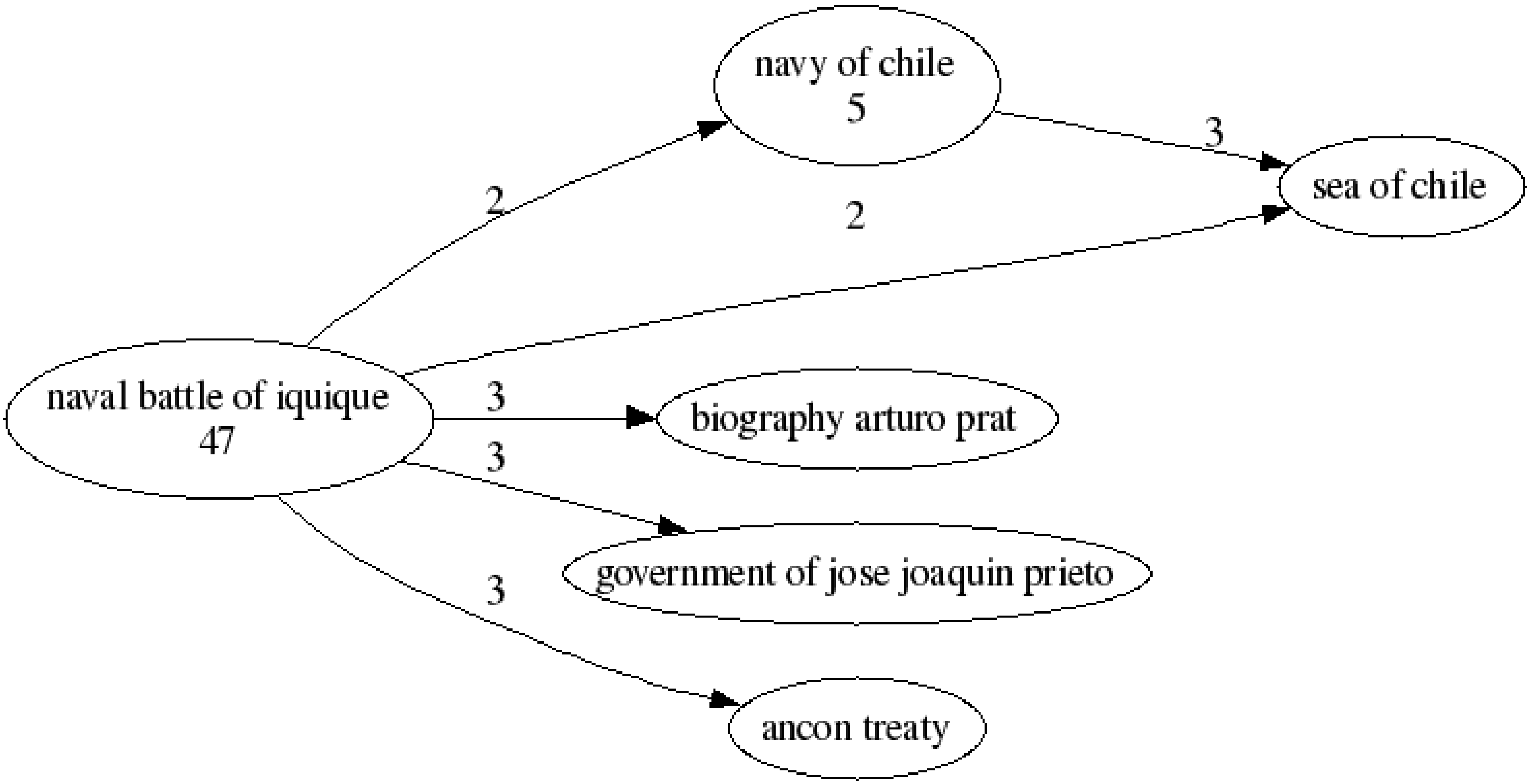
# Query Recommendation

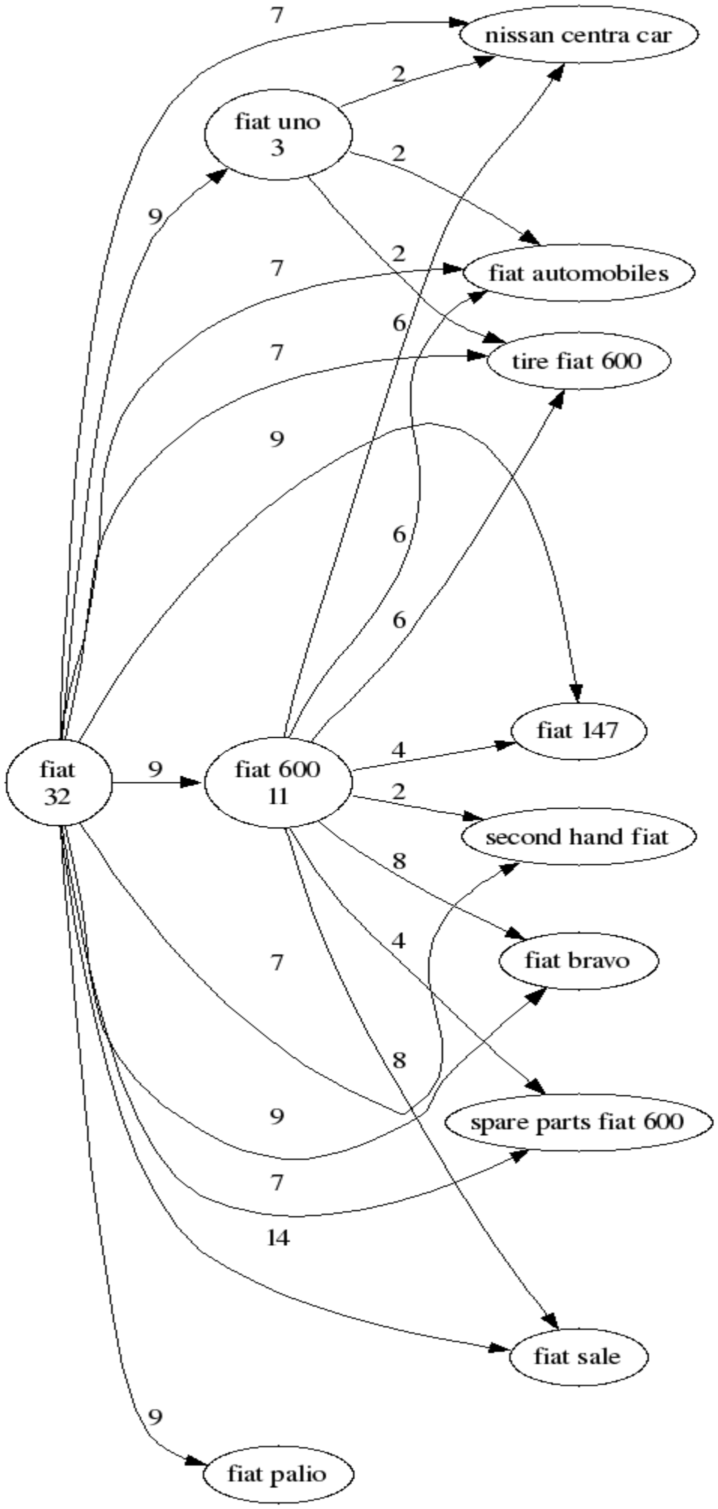


| Query | Popularity | Support | Closedness | Rank |
|---|---|---|---|---|
| rentals apartments viña del mar owners | 2 | 0,133 | 0,403 | 0,268 |
| rentals apartments viña del mar | 10 | 0,2 | 0,259 | 0,229 |
| viel properties | 4 | 0,1 | 0,315 | 0,207 |
| rental house viña del mar | 2 | 0,166 | 0,121 | 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 |

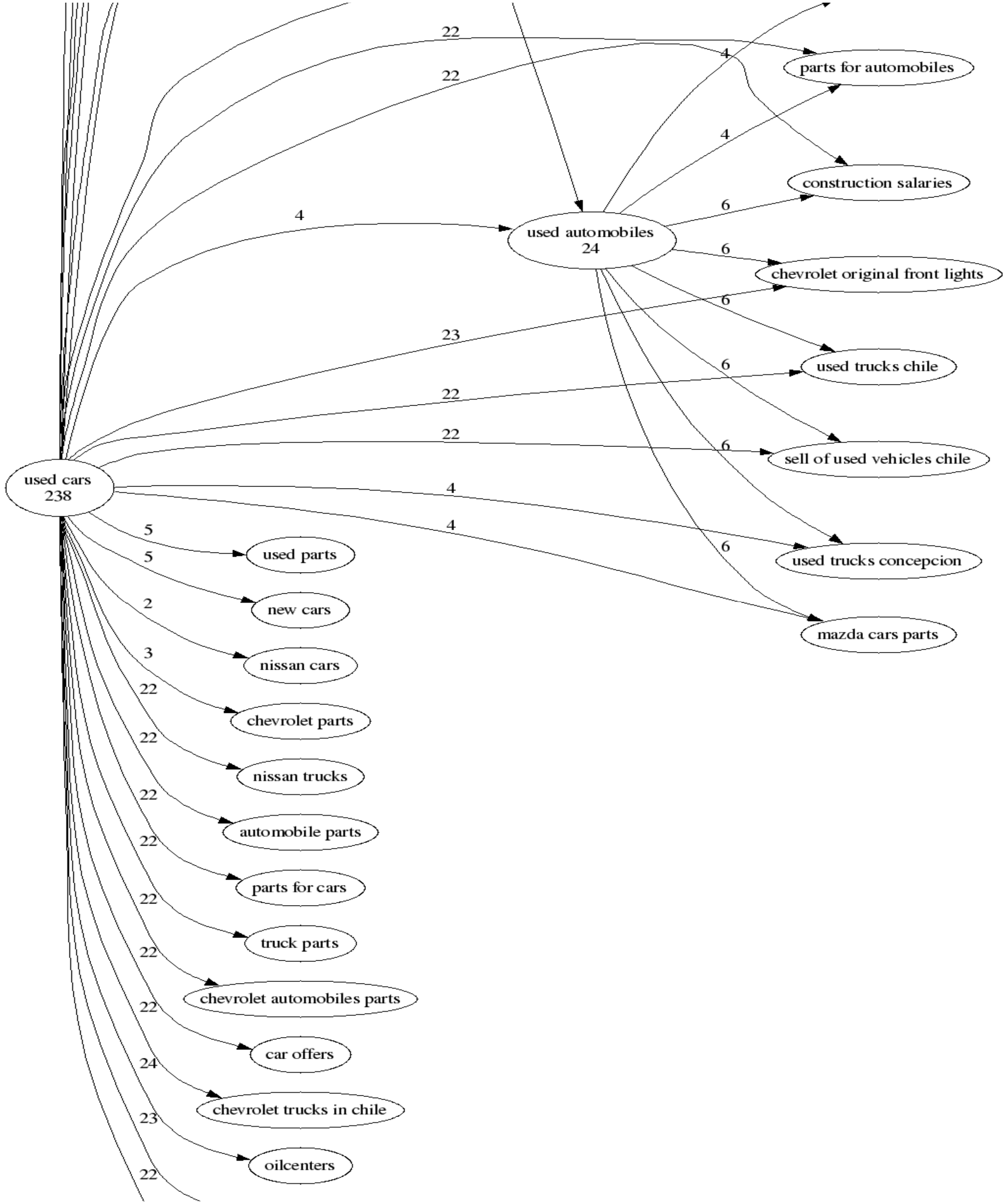
# Simple Related Terms

Query dominance based on clicked pages



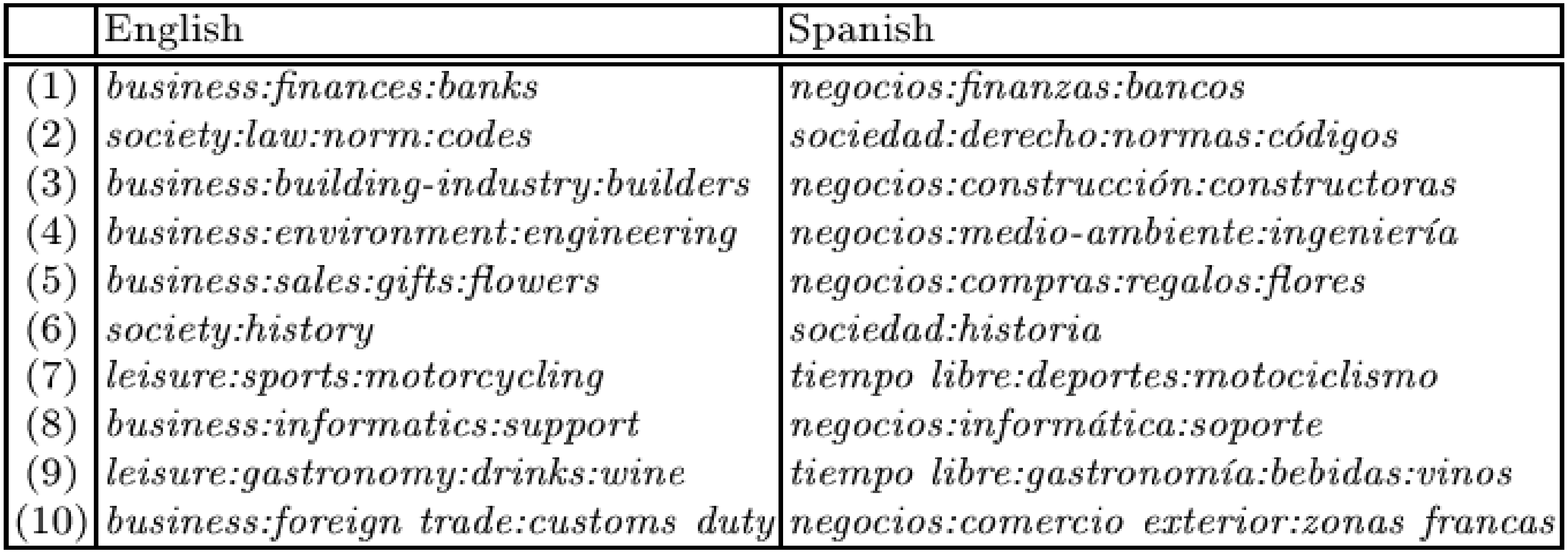

fiat
32
fiat uno
3
fiat 600
11
nissan centra car
fiat automobiles
tire fiat 600
fiat 147
second hand fiat
fiat bravo
spare parts fiat 600
fiat sale
fiat palio
7
2
2
9
2
7
6
7
9
6
6
9
4
2
8
4
7
8
9
7
14
9

DD

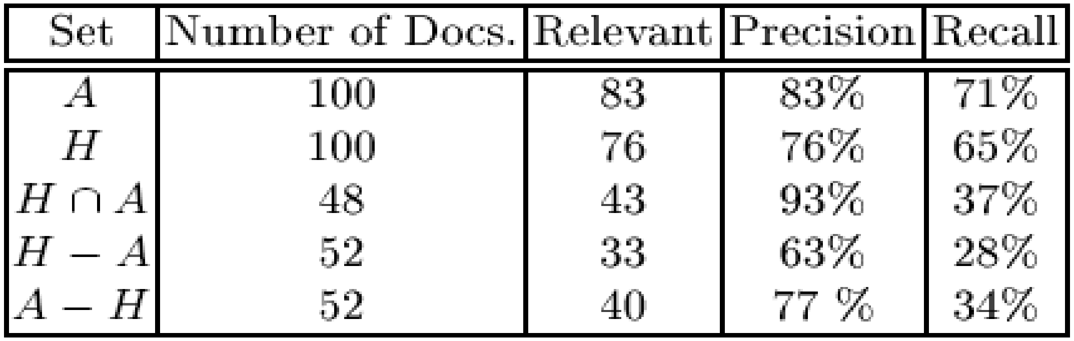

used cars
238
used automobiles
24
parts for automobiles
construction salaries
chevrolet original front lights
used trucks chile
sell of used vehicles chile
used trucks concepcion
mazda cars parts
used parts
new cars
nissan cars
chevrolet parts
nissan trucks
automobile parts
parts for cars
truck parts
chevrolet automobiles parts
car offers
chevrolet trucks in chile
oilcenters
22
22
4
4
4
6
6
6
6
6
6
23
22
22
4
4
5
5
2
3
22
22
22
22
22
22
22
24
23
22

# Taxonomies

## Infer topics from queries that imply



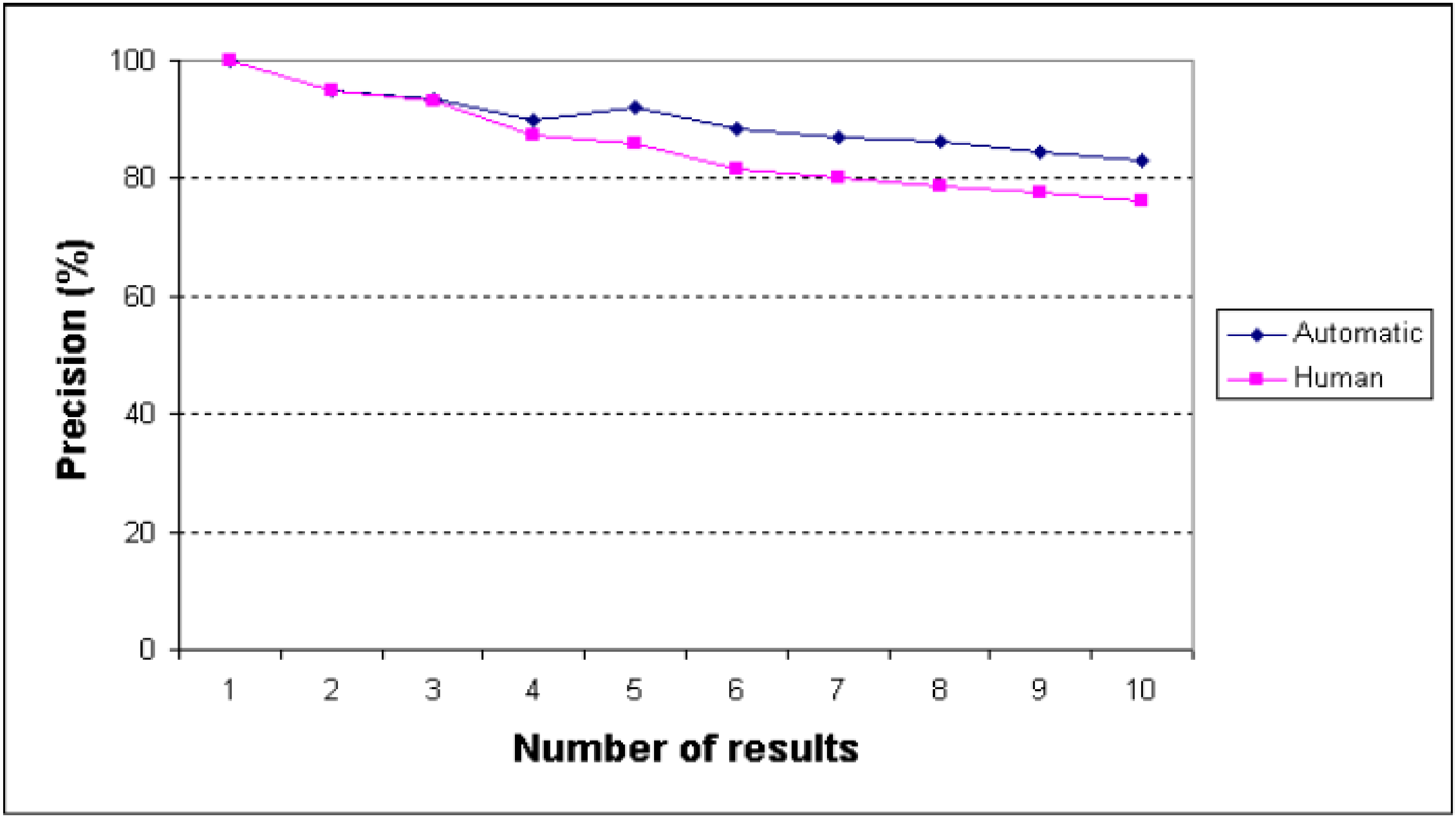
| | English | Spanish |
|---|---|---|
| (1) | *business:finances:banks* | *negocios:finanzas:bancos* |
| (2) | *society:law:norm:codes* | *sociedad:derecho:normas:códigos* |
| (3) | *business:building-industry:builders* | *negocios:construcción:constructoras* |
| (4) | *business:environment:engineering* | *negocios:medio-ambiente:ingeniería* |
| (5) | *business:sales:gifts:flowers* | *negocios:compras:regalos:flores* |
| (6) | *society:history* | *sociedad:historia* |
| (7) | *leisure:sports:motorcycling* | *tiempo libre:deportes:motociclismo* |
| (8) | *business:informatics:support* | *negocios:informática:soporte* |
| (9) | *leisure:gastronomy:drinks:wine* | *tiempo libre:gastronomía:bebidas:vinos* |
| (10) | *business:foreign trade:customs duty* | *negocios:comercio exterior:zonas francas* |



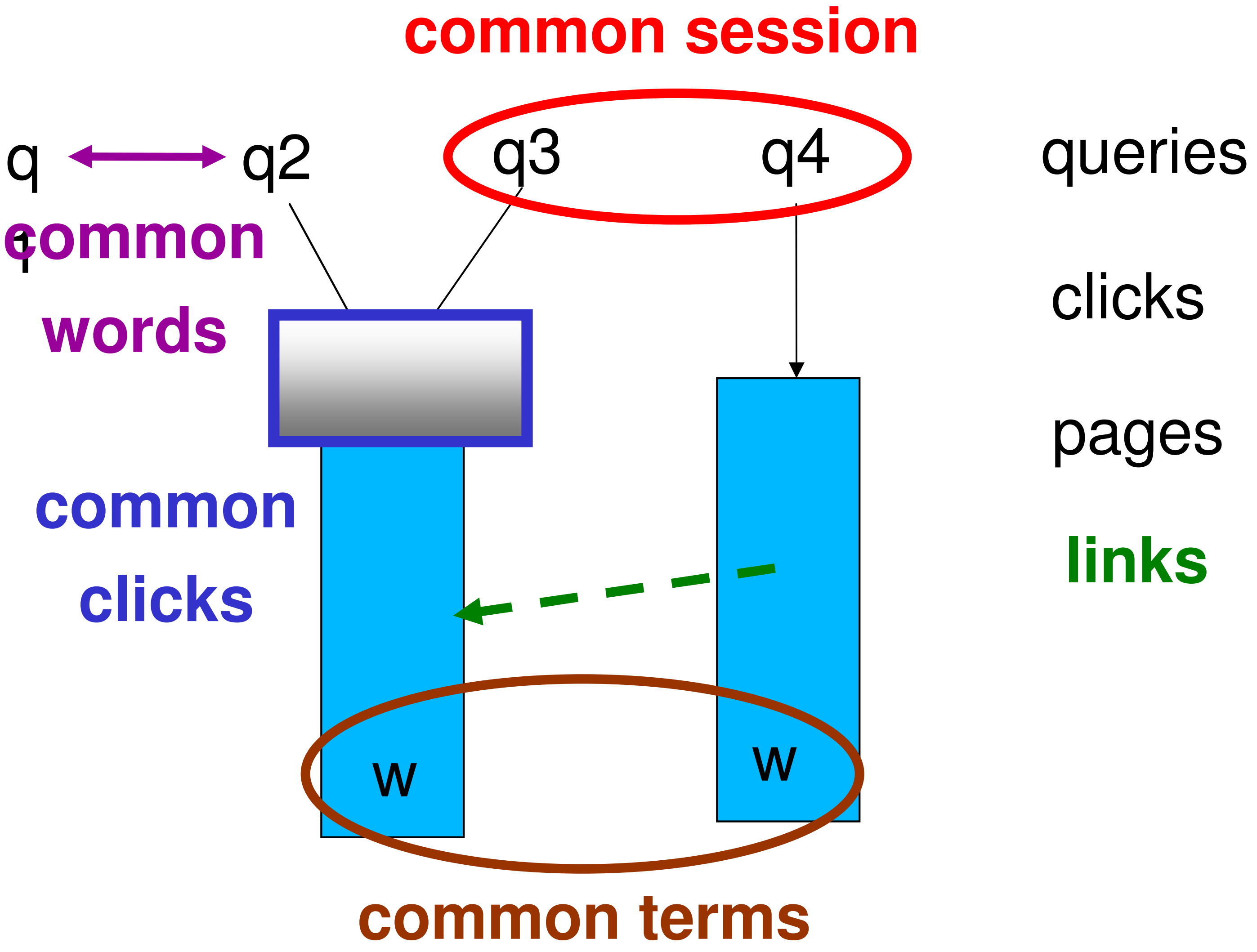
| Set | Number of Docs. | Relevant | Precision | Recall |
|---|---|---|---|---|
| $A$ | 100 | 83 | 83% | 71% |
| $H$ | 100 | 76 | 76% | 65% |
| $H \cap A$ | 48 | 43 | 93% | 37% |
| $H - A$ | 52 | 33 | 63% | 28% |
| $A - H$ | 52 | 40 | 77 % | 34% |

# Results better than humans!

## Quality of answers



# Relating Queries (Baeza-Yates, 2007)



common session

q1 q2 q3 q4

queries

common words

clicks

pages

common clicks

w w

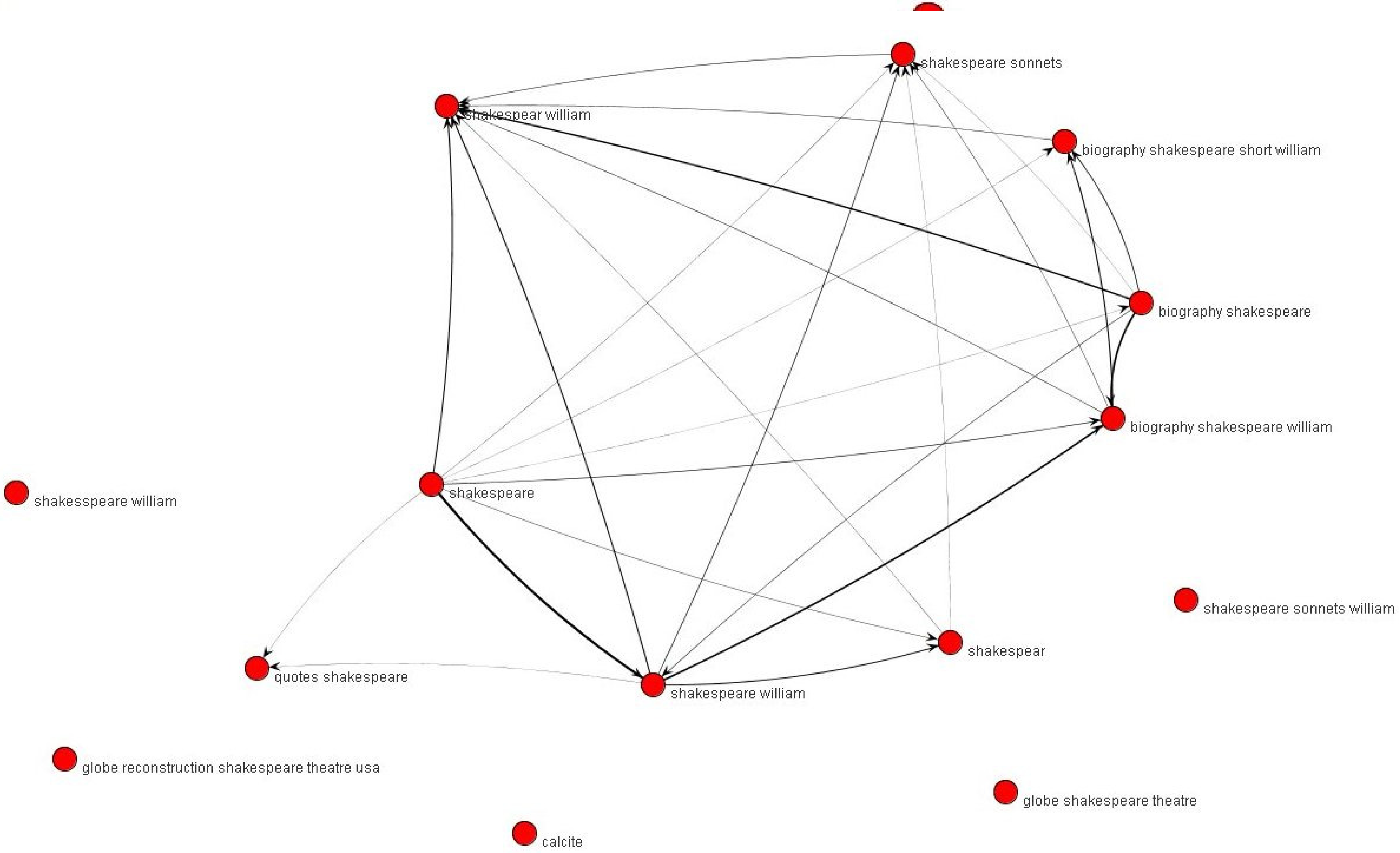
links

common terms

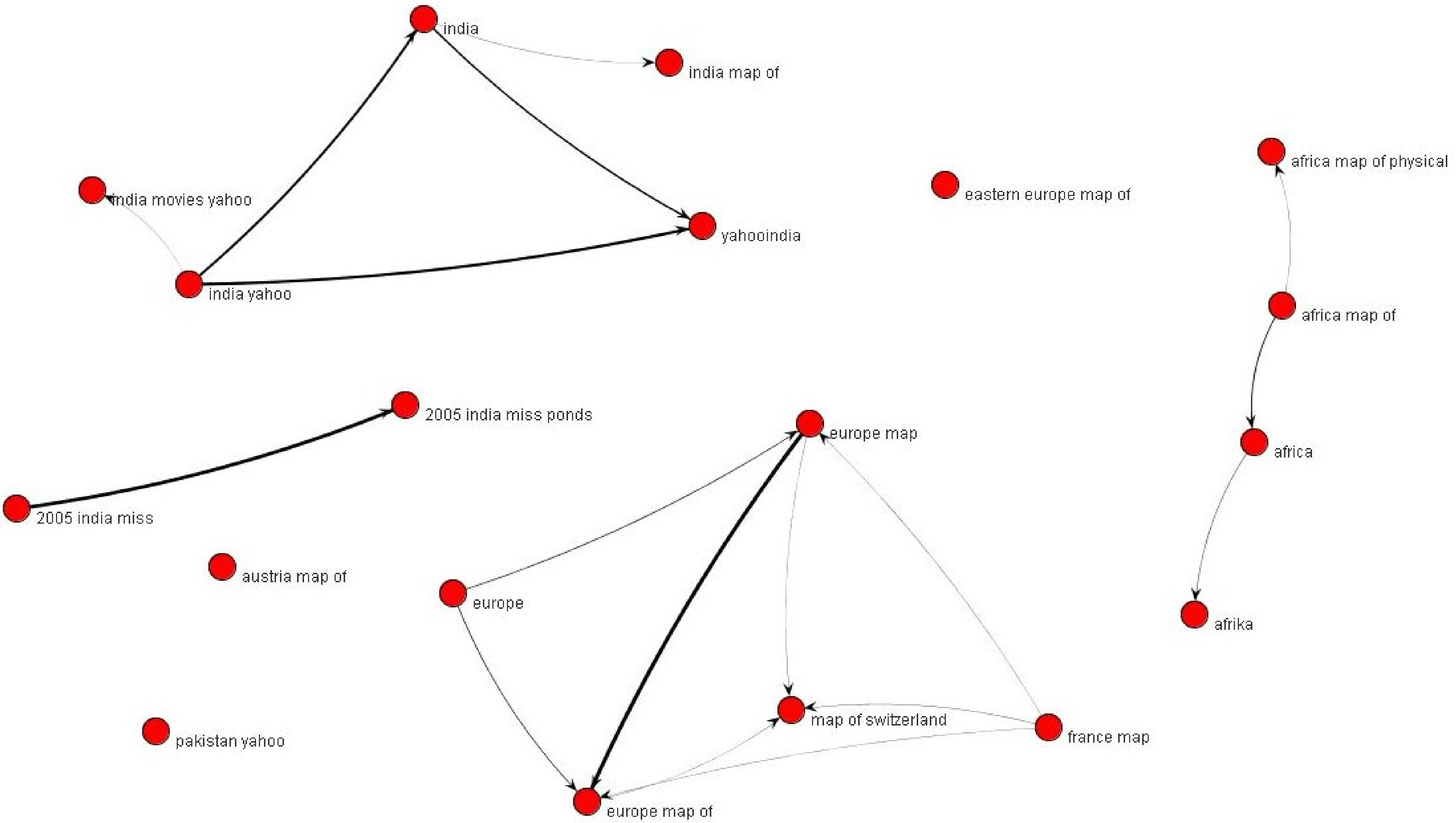
# Qualitative Analysis

| Graph | Strength | Sparsity | Noise |
|---|---|---|---|
| Word | Medium | High | Polysemy |
| Session | Medium | High | Physical sessions |
| Click | High | Medium | Multitopic pages<br>Click spam |
| Link | Weak | Medium | Link spam |
| Term | Medium | Low | Term spam |

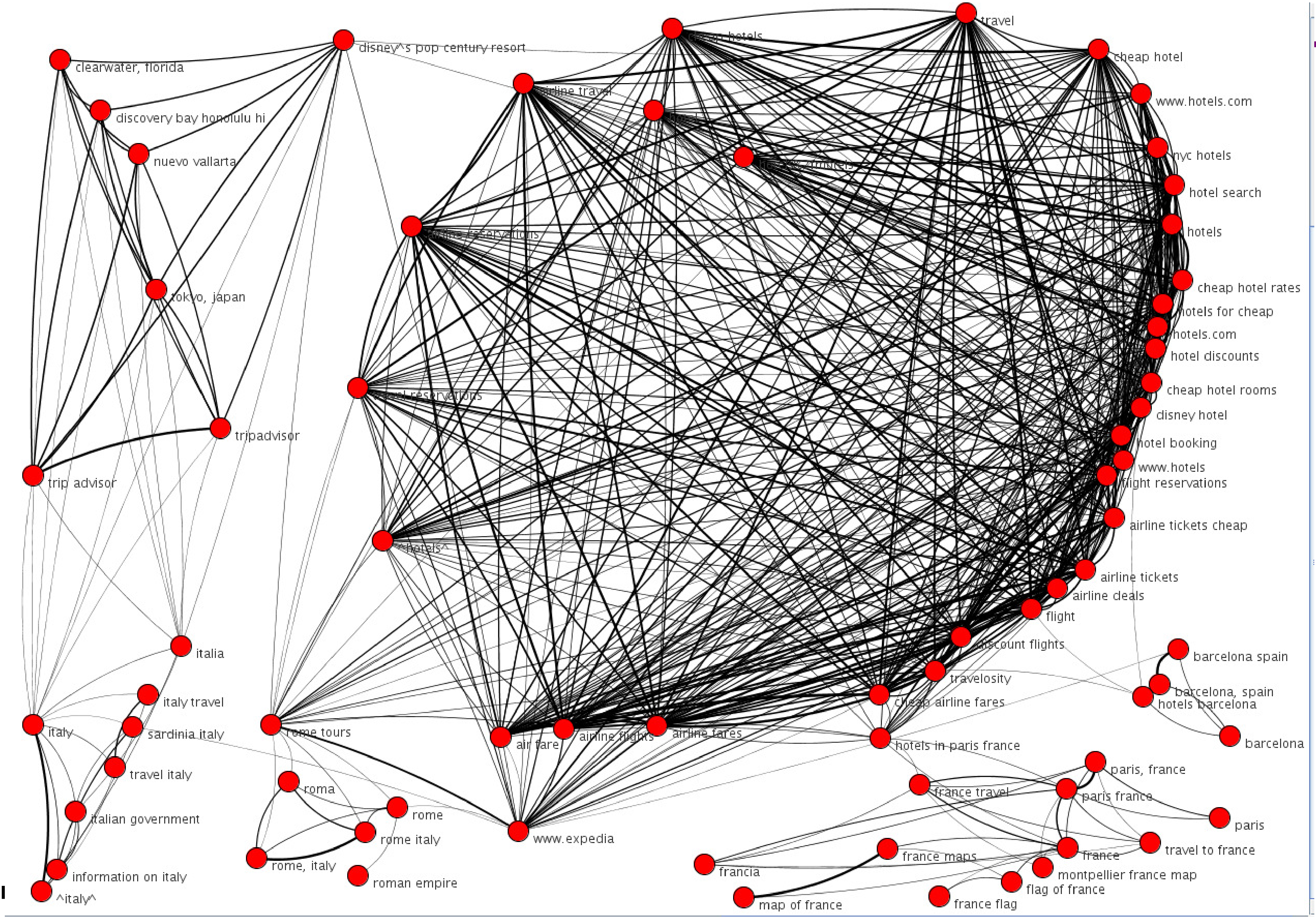
# Words, Sessions and Clicks


shakespeare sonnets
shakespear william
biography shakespeare short william
biography shakespeare
biography shakespeare william
shakesspeare william
shakespeare
shakespeare sonnets william
shakespear
quotes shakespeare
shakespeare william
globe reconstruction shakespeare theatre usa
globe shakespeare theatre
calcite

# Words, Sessions and Clicks



# Click Graph



Aı

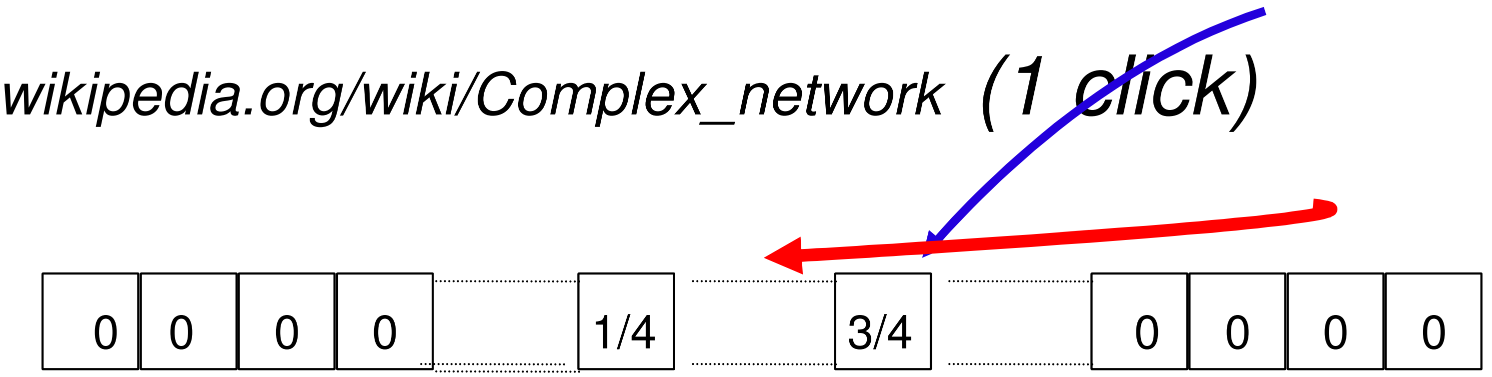
# Formal definition

- 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

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

# 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”

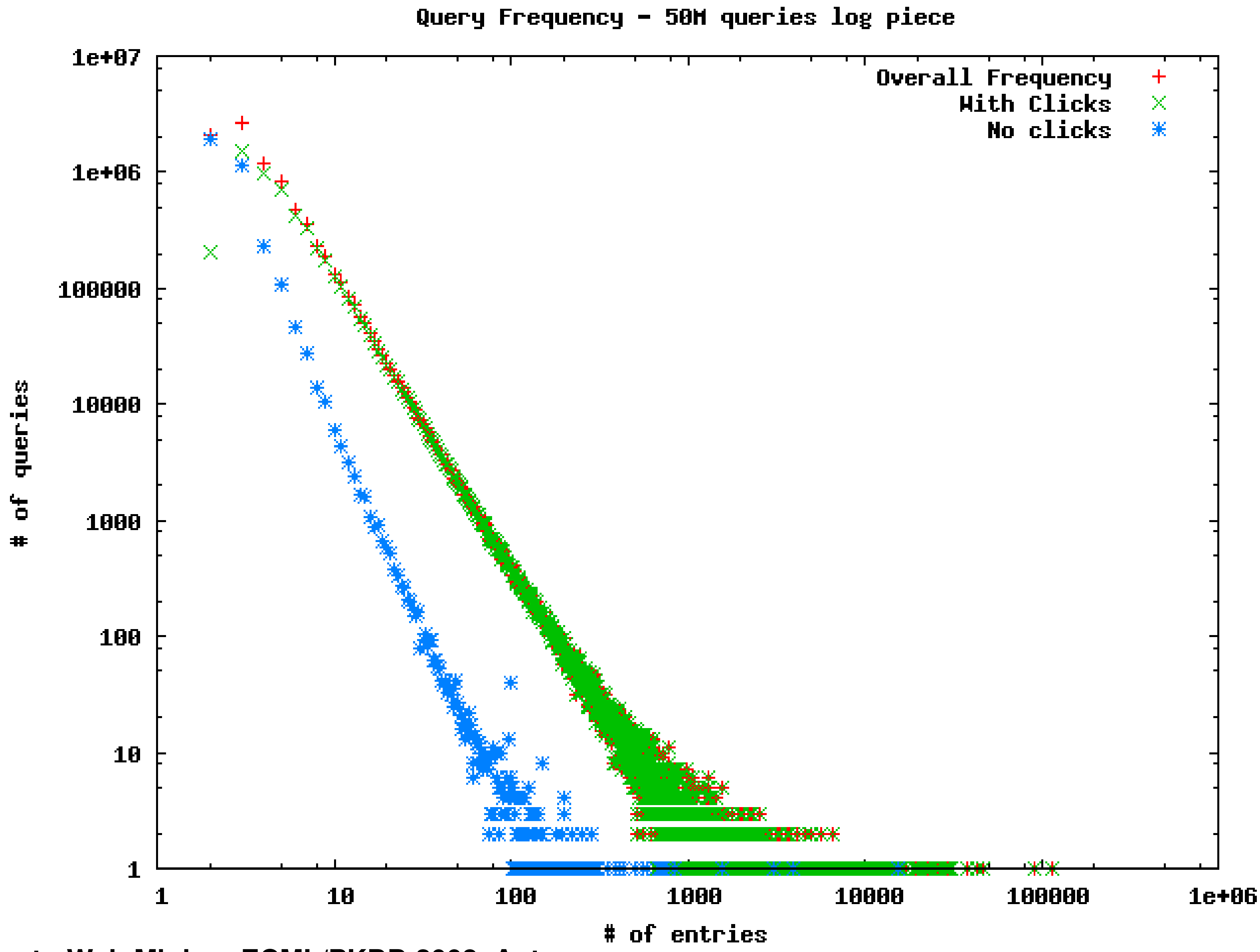
# Building the Graph

- 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*

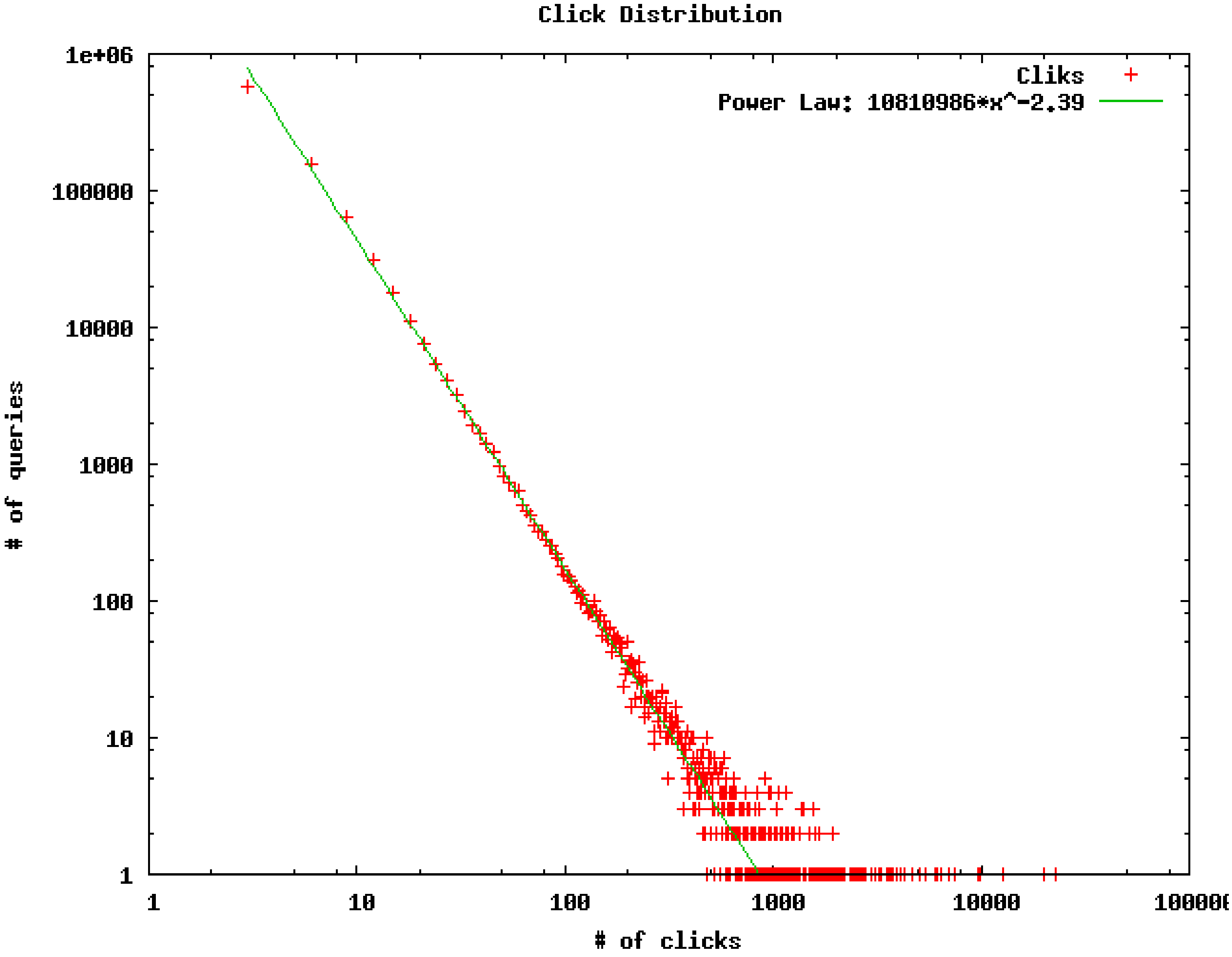
# Anatomy of a Click Graph

- 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

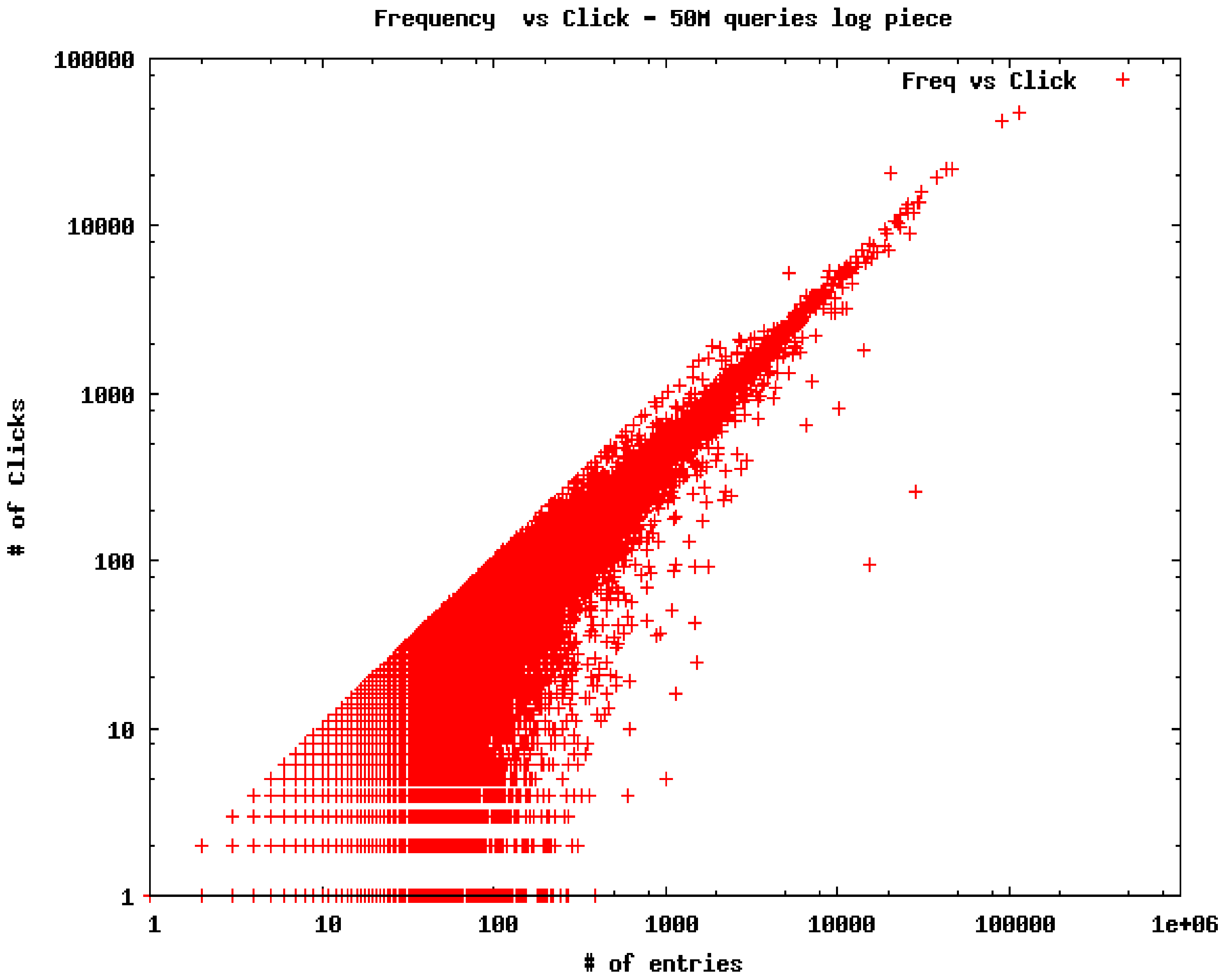
# Query Frequency



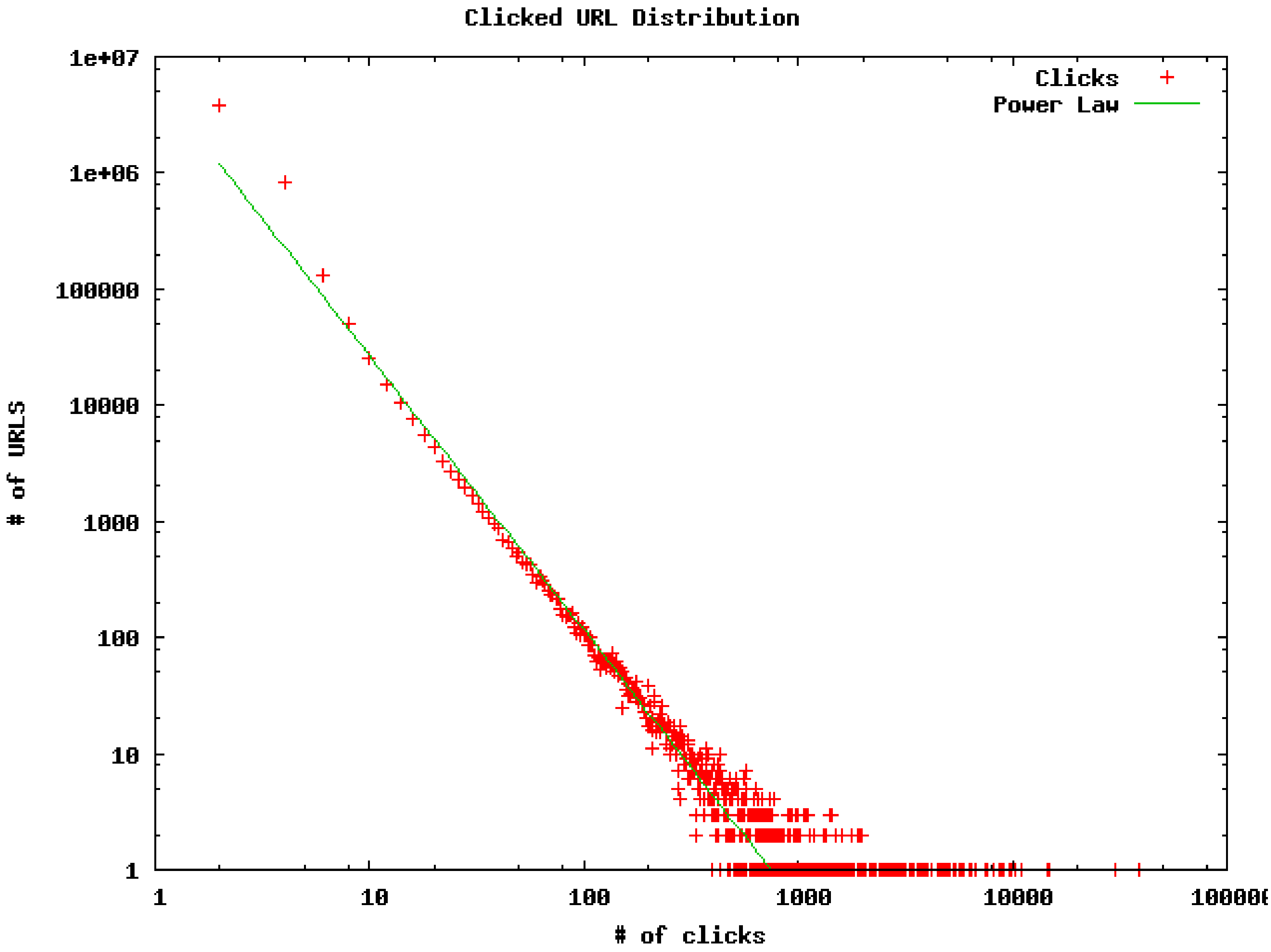
# Click Distribution



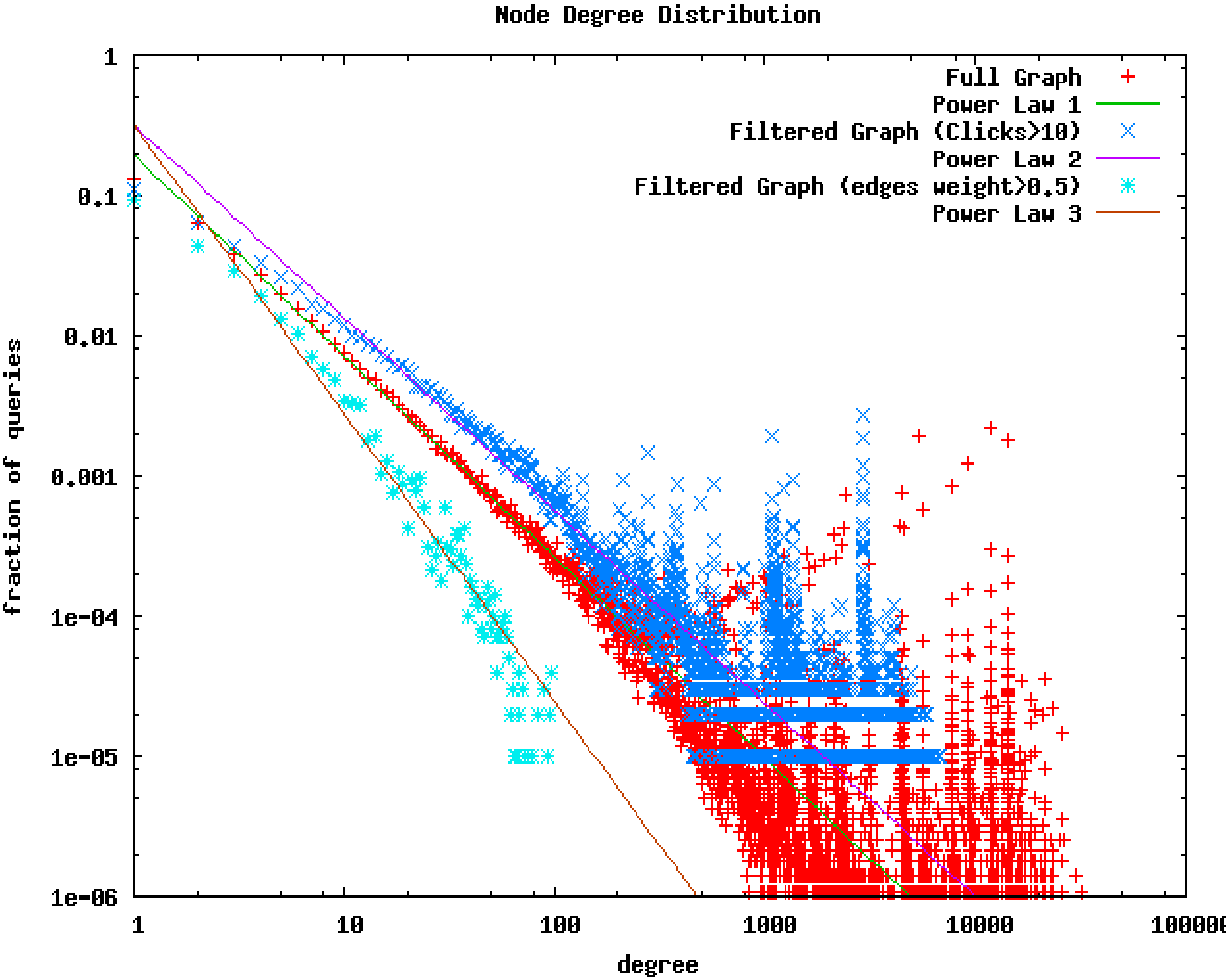
# Query Frequency vs. Clicks



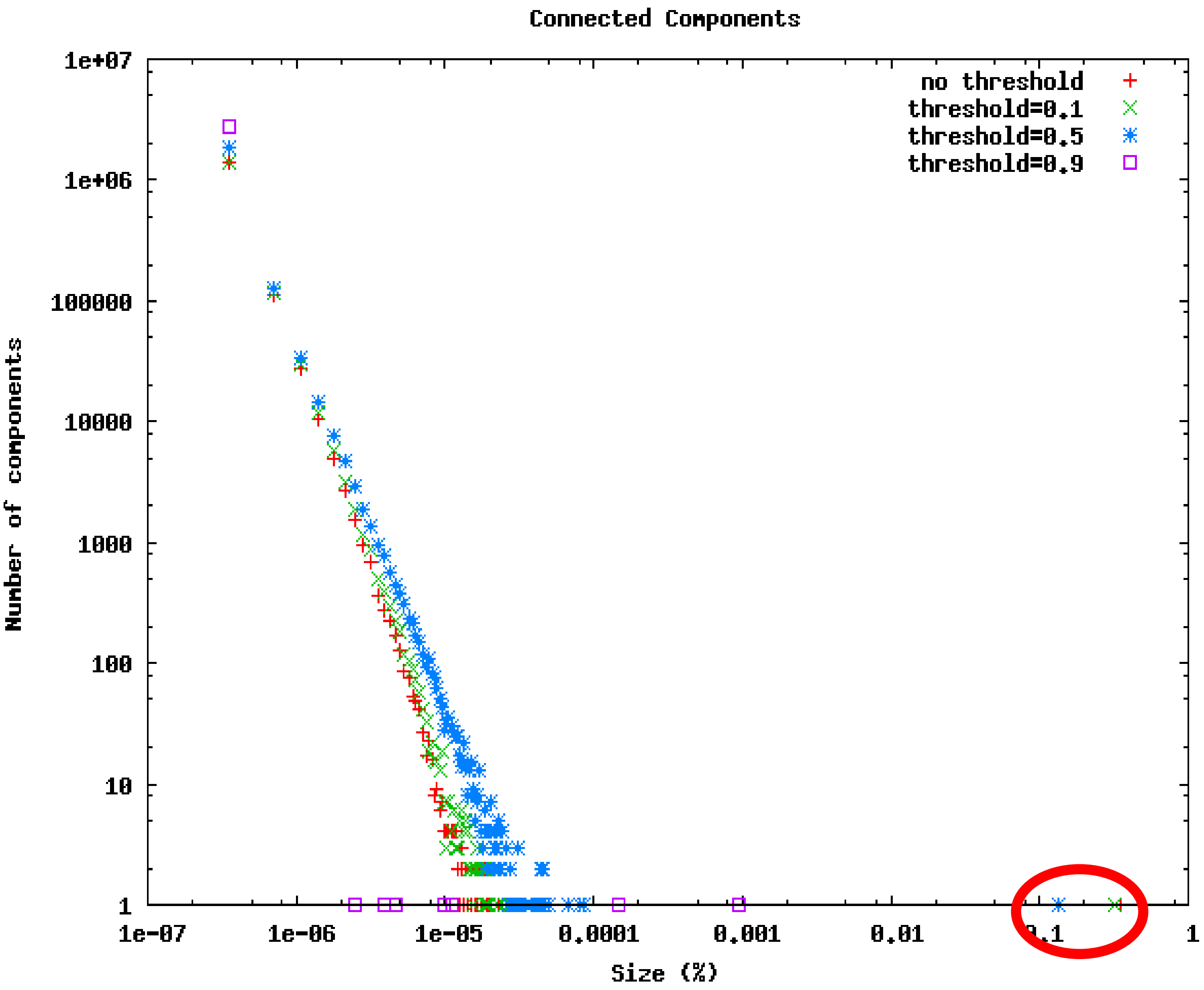
# Clicked URL DIstribution



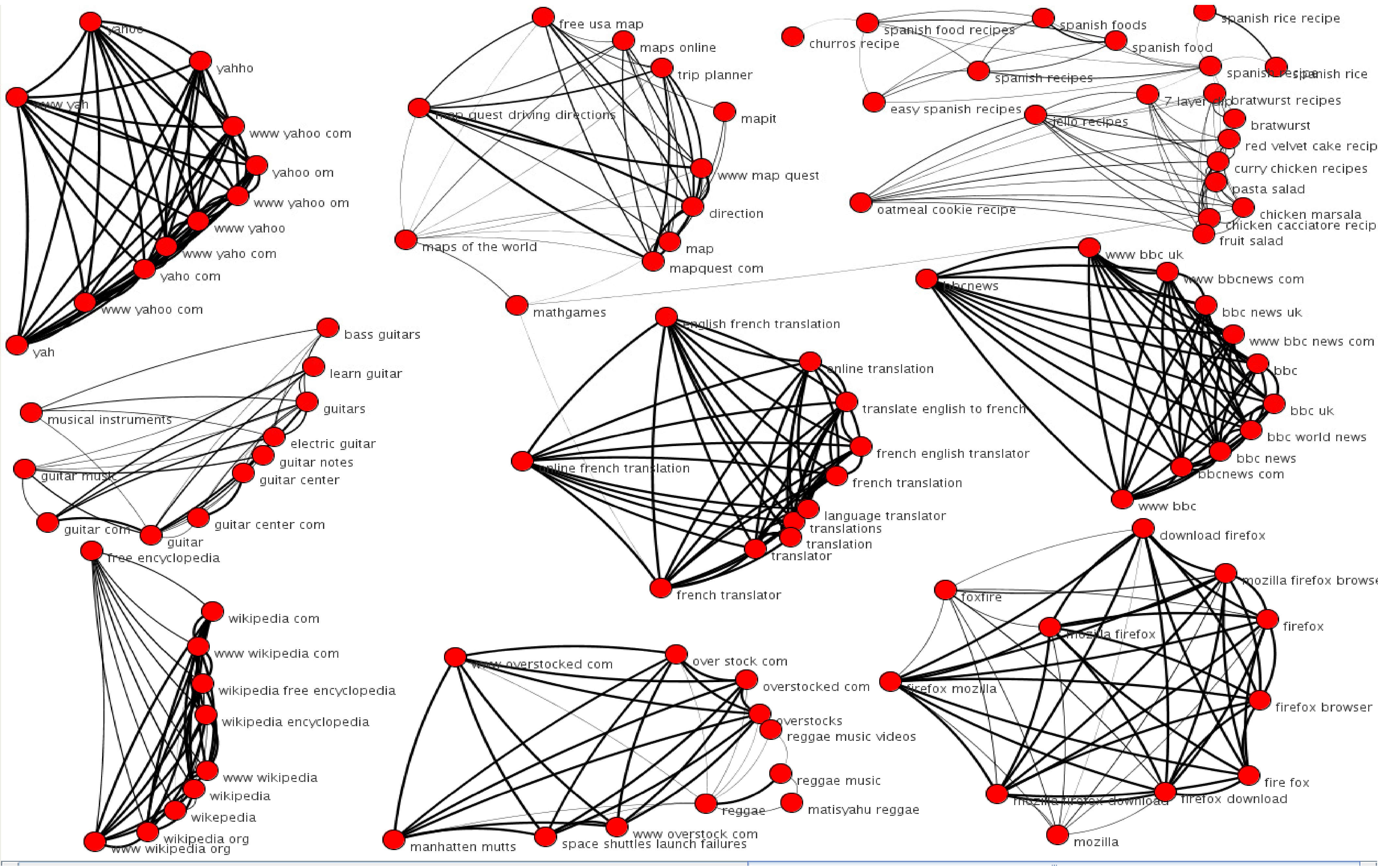
# Node Degree Distribution



# Connected Components



# Implicit Folksonomy?



# Set Relations and Graph Mining

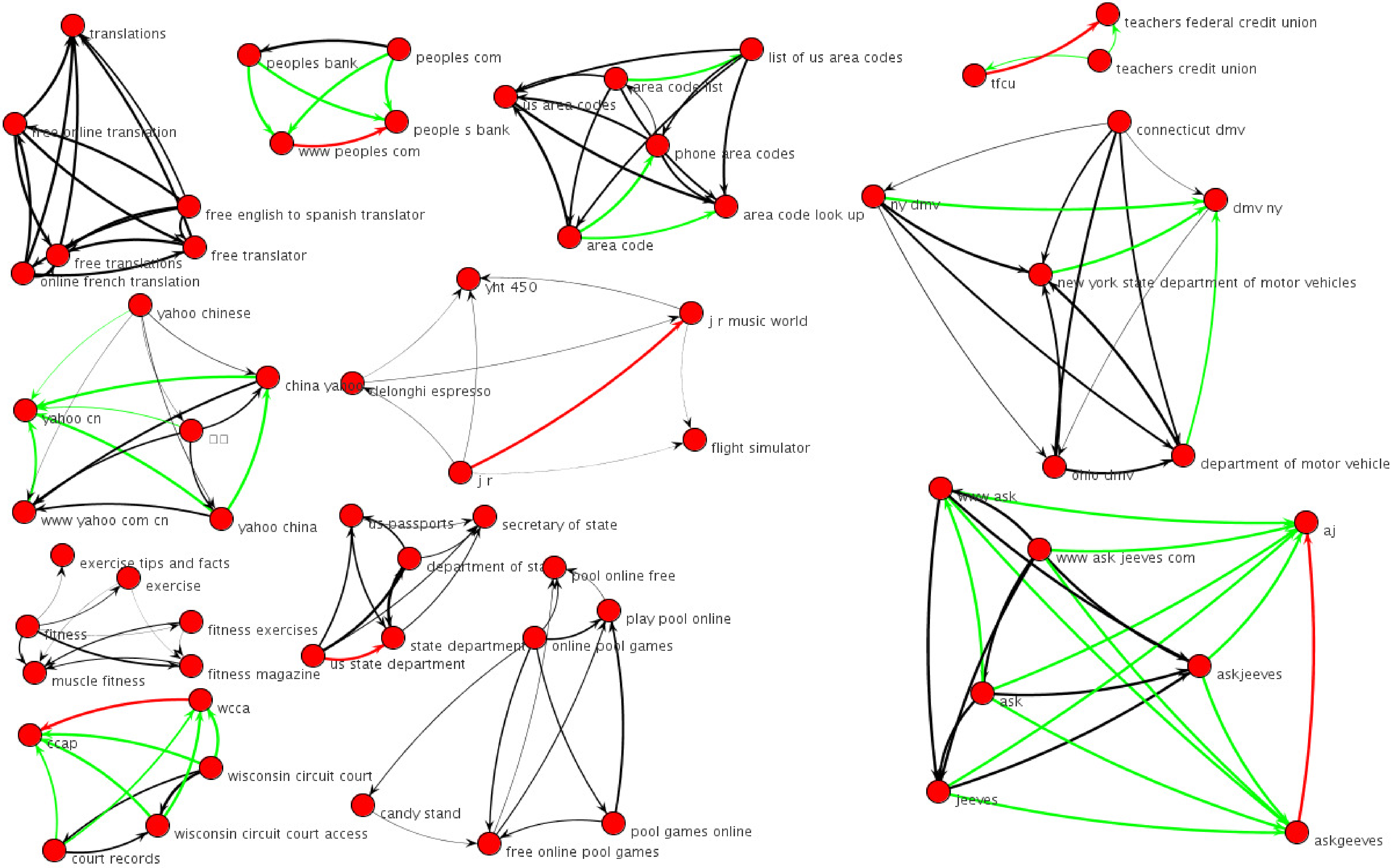
- Identical sets: equivalence
- Subsets: specificity
  - directed edges

Baeza-Yates & Tiberi

ACM KDD 2007

- Non empty intersections (with threshold)
  - degree of relation
- Dual graph: URLs related by queries
  - High degree: multi-topical URLs

# 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) | i,j=1,..,K }*

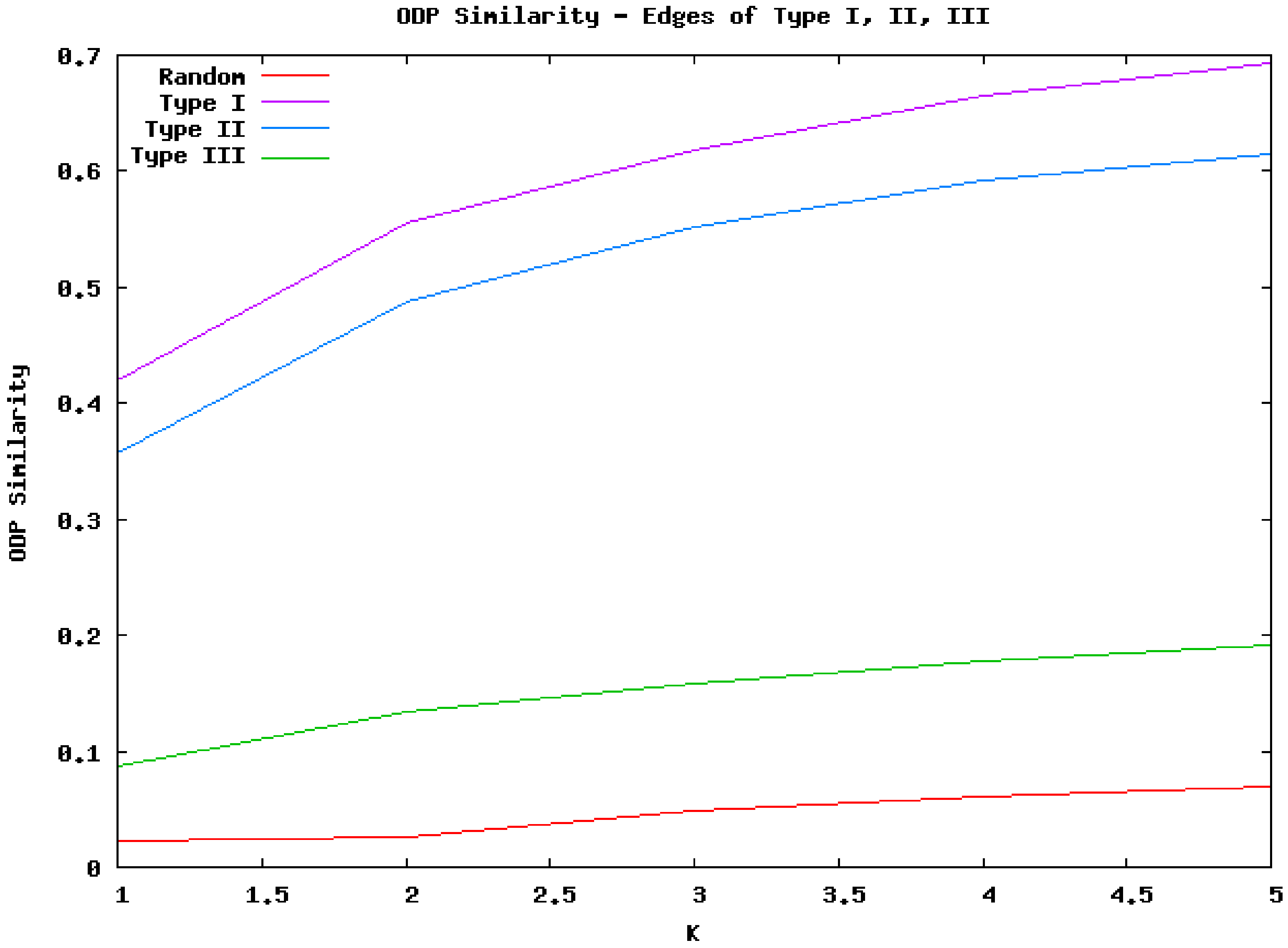
# ODP Similarity

- 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

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

# Experimental Evaluation



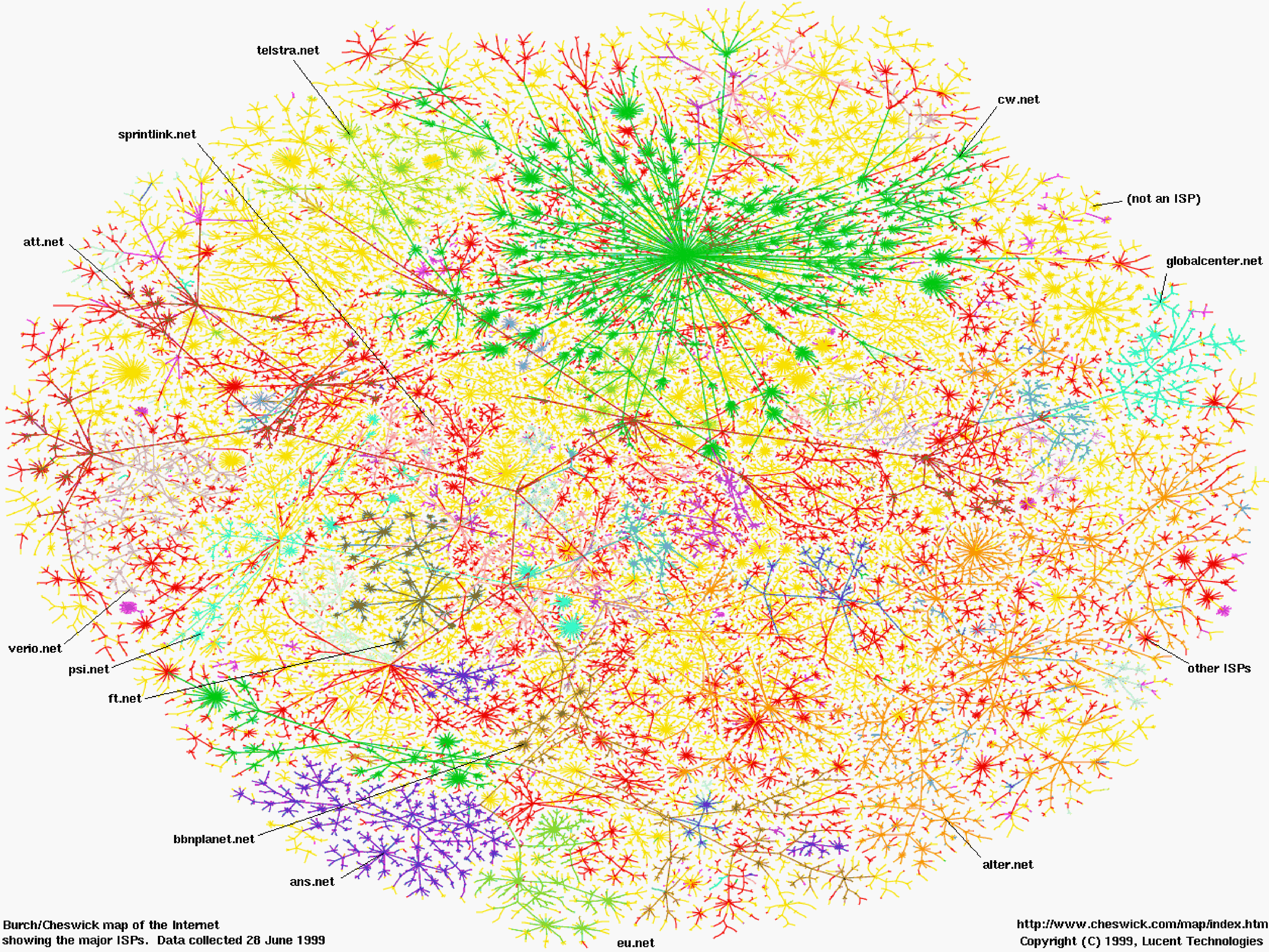
# Open Issues

- 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

# Link analysis

- Infer properties of Web entities based on their connectivity / link structure of graph structures they belong to
- Such properties can be importance of nodes or similarity between nodes
- Mostly focused on Web pages, but ideas apply to many domains: social networks, query logs, etc.

- Prestige, centrality, co-citation, PageRank, HITS



Burch/Cheswick map of the Internet
showing the major ISPs.  Data collected 28 June 1999

http://www.cheswick.com/map/index.html

# Social sciences and bibliometry

*“...we are involved in an 'infinite regress': [an actor's status] is a function of the status of those who choose him; and their [status] is a function of those who choose them, and so ad infinitum”*

[Seeley, 1949]

# Prestige

- Consider a graph $G=(V,E)$
- $E[u,v] = 1$ if there is a link from $u$ to $v$
- $E[u,v] = 0$ otherwise
- $p$ a prestige vector: $p[u]$ the prestige score of node $u$

$$p' = E^T p$$

because

$$p[u] = Sum_v\, E[v,u]\, p[u] = Sum_v\, E^T[u,v]\, p[u]$$

- After each iteration normalize by setting $||p|| = 1$
- $p$ converges to the principal eigenvector of $E^T$

# Centrality

- Importance notion based on centrality
- Used by epidimiology, social-network analysis, etc.: removing a central node disconnects the graph to a big extend

- $d(u,v)$ the shortest-path distance between *u* and *v*
- $r(u) = \max_v d(u,v)$ radius of node u
- $\arg\min_u r(u)$ center of the graph

- Various other notions of centrality in the literature

# Co-citation

- Measure of similarity between nodes
- If nodes *v* and *w* are both linked by node *u*, then they are co-cited
- If E is the adjacency matrix of the graph, the number of nodes that co-cite both v and w is

$$p[u] = Sum_u E[u,v] E[u,w] = Sum_u E^T[v,u] E[u,w] = (E^T E)[v,w]$$

- Thus similarity is captured in the entries of matrix $E^T E$

# PageRank

- [Brin and Page, 1998]
- Algorithm suggested for ranking results in web search
- An authority score is assigned to each Web page
- Authority scores independent of the query

- Authority scores corresponds to the stationary distribution of a random walk on the graph:
  - With probability a follow a link in the graph
  - With probability 1-a go to a node chosen uniformly at random (teleportation)

- Random walk also known as random surfer model

# PageRank

- Let E be the adjacency matrix of the graph, and L the row-stochastic version of E
- Each row of E is normalized so that it sums to 1
- Authority score defined by

$$p_{(i+1)} = L^T p_{(i)}$$

- problematic if the graph is not strongly connected, So:

$$p_{(i+1)} = a\, L^T p_{(i)} + (1\text{-}a) 1/n\ I$$

- where I is the matrix with all entries equal to 1
- and a in [0,1], common value a = 0.85

# PageRank variants and enchancements

- Personalized PageRank
  - Teleportation to a set of pages defining the preferences of a particular user
- Topic-sensitive PageRank [Haveliwala 02]
  - Teleportation to a set of pages defining a particular topic
- TrustRank [Gyöngyi 04]
  - Teleportation to “trustworthy” pages

- Many papers on analyzing PageRank and numerical methods for efficient computation

# HITS

- [Kleinberg 1998]
- Exploit the intuition that there are:
  - pages that contain high-quality information (authorities)
  - pages with good navigational properties (hubs)

*Good hubs point to good authorities and good authorities are pointed by good hubs*

# HITS algorithm

- Given a query *q*
- Use a standard wen IR system to find a set of pages *R* relevant to *q* (*root set*)
- Expand to the set of pages connected to *R* (*expanded set*) and form the graph *G=(V,E)*
- *a authority vector: a[u] the authority score of node u*
- *h hub vector: h[u] the hub score of node u*

$$a = E^T h$$

$$h = E a$$

- *a* converges to the principal eigenvector of $E^T E$
- *h converges to the principal eigenvector of* $EE^T$

# HITS

- HITS is related to SVD on the graph matrix E
- non-principal eigenvectors provide different topics
- HITS sensitive to local-topology
- PageRank is more stable – due to trandom jump step
- Researchers attempted to make HITS more stable
  - SALSA stochastic algorithm for link analysis [Lempel and Moran, 01]:
  - A random surfer model in which the surfer follows alternatively random inlinks and outlinks
  - [Ng et al. 01] introduce a random jump step in the HITS model

# Discussion

- HITS introduces the notion of hub, which does not exist in PageRank
- HITS is query sensitive
- PageRank does not depend on the query; thus the authority scores can be pre-computed

- Nepotism, two-host nepotism, and clique attacks

# Algorithmic tools

- Keep an eye on efficiency
- Web graphs are huge and any computation on them should be very efficient
- Data stream algorithms for
  - Computing the clustering coefficient
  - Counting the number of triangles
  - Estimating the diameter of a graph

# Clustering coefficient

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

- How to compute it?
- How to compute the number of triangles in a graph?
- Assume that the graph is very large, stored on disk

# 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 Chernoff bound, to get an eps-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 large compared to $|T_3|$

# Counting triangles

- SampleTriangle Algorithm [Buriol et al., 2006]
- Incidence stream model – all edges incident on the same edge are consecutive on the disk

- Three pass algorithm:
- Pass 1: Count the number of paths of length 2
- Pass 2: Choose one path (a,u,b) uniformly at random
- Pass 3: If (a,b)in E return 1 o/w return 0

# Counting triangles

- The previous idea can be also applied to:
  - Count triangles when edges are stored in arbitrary order
  - Obtain one-pass algorithm
  - Count other minors

# 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

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

- With N(h) can 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
    $M(u, h) = M(u, h-1)$ for all $u$ in $V$
    for each edge $(u, v)$ do
        $M(u, h) = M(u, h)$ union $M(v, h-1)$
```

- 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

# Finding communities

- A set of related Web pages
- A group of scientists collaborating with each other
- A set of blog posts discussing a specific topic
- A set of related queries

- Can be used for improving relevance of search, recommendations, propagating an idea, advertising a product, etc.

- Usually formulated as a graph clustering problem

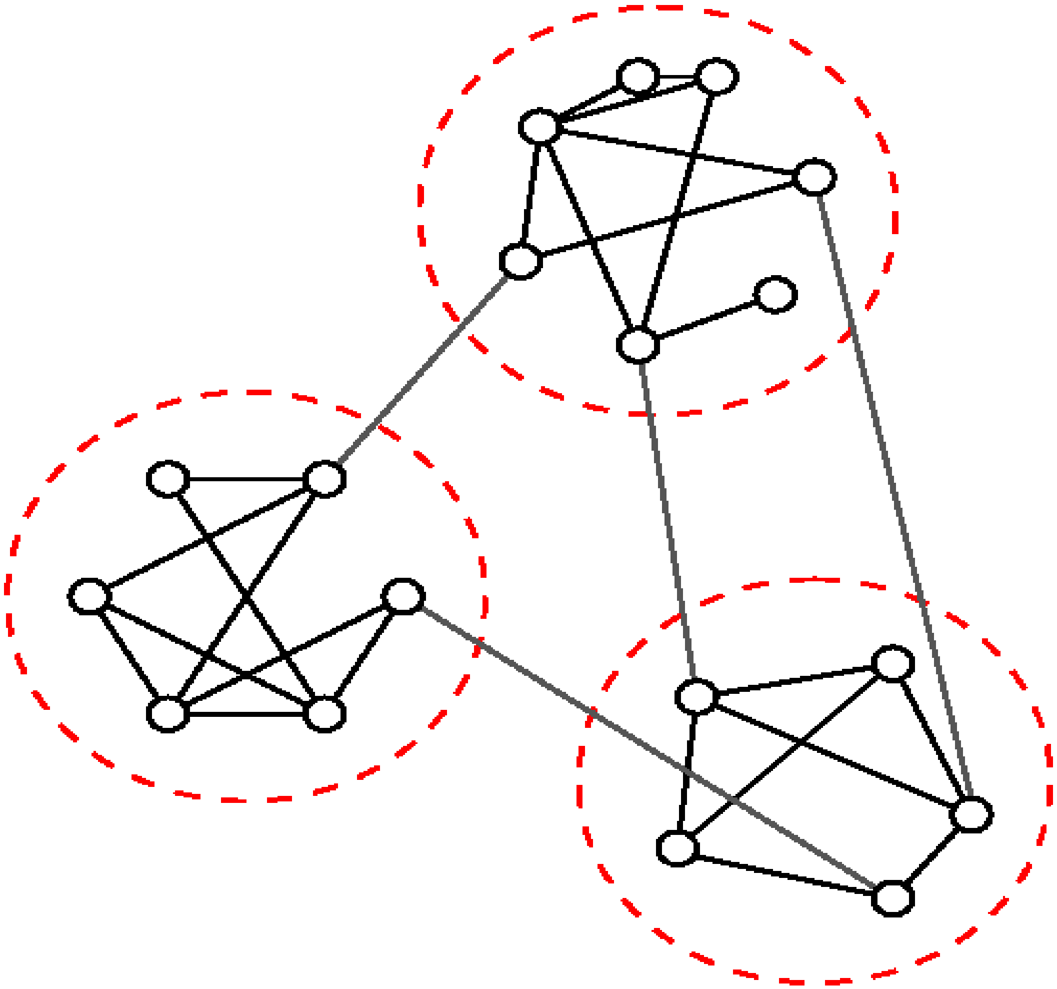
# Graph clustering

- Graph $G = (V, E)$
- Edge $(u, v)$ denotes similarity between $u$ and $v$
  - weighted edges can be used to denote degree of similarity

- We want to partition the vertices in clusters so that:
  - vertices within clusters are well connected, and
  - vertices across clusters are sparsely connected

- Most graph partitioning problems are NP hard

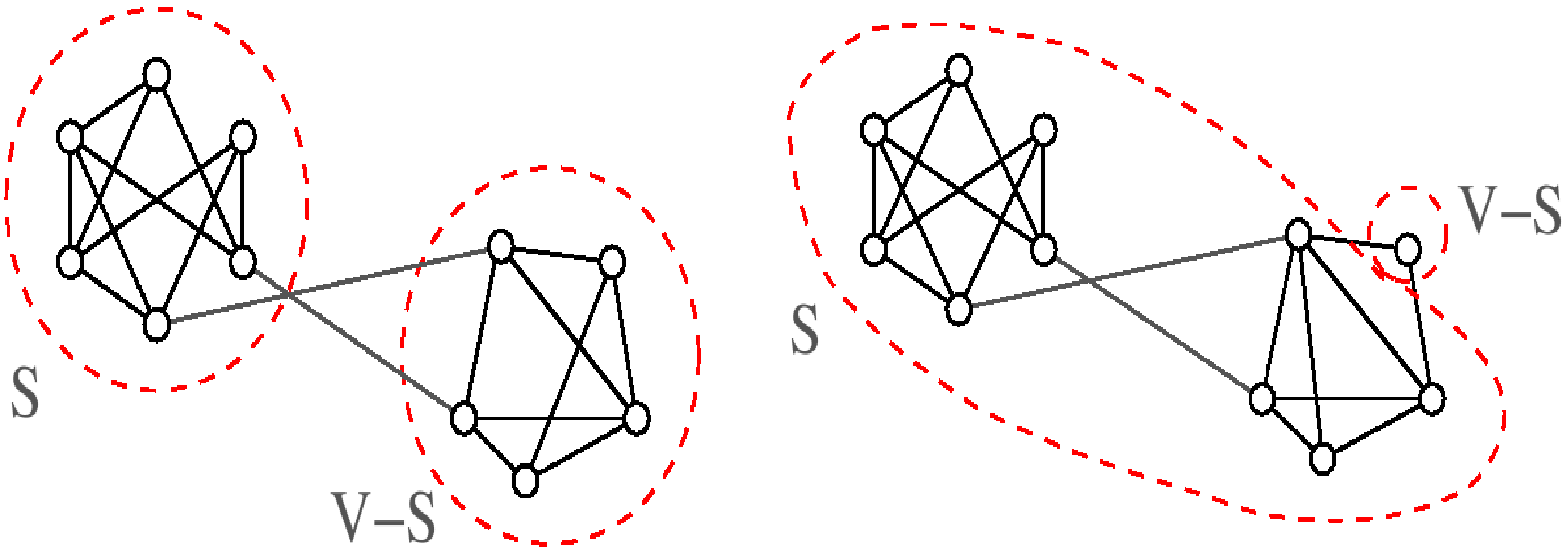
# Graph clustering



# Measuring connectivity

- Minimum cut: The minimum number of edges whose removal disconnects the graph

$$c(S) = min_{S\ in V} |\{(u,v)\ in\ E\ s.t.\ u\ in\ S\ and\ v\ in\ V\text{-}S\ \}|$$



# Graph expansion

- Normalize the cut by the size of the smallest component
- Define cut ratio

$$\alpha(G,S) = \frac{c(S)}{\min\{|S|, |V-S|\}}$$

- And graph expansion

$$\alpha(G) = \min_{S} \frac{c(S)}{\min\{|S|, |V-S|\}}$$

- Other similar normalized criteria have been proposed
- Related to the eigenvalues of the adjacency matrix of the graph, thus with the expansion properties of the graph

# Spectral analysis

- Let $A$ be the adjacency matrix of the graph $G$
- Define the Laplacian matrix of $A$ as

$$L = D - A,$$

- $D = diag(d_1, . . . , d_n)$, a diagonal matrix
- $d_i$ the degree of vertex $i$

$$L_{ij} = \begin{cases} d_i & \text{if } i = j \\ -1 & \text{if } (i,j) \in E, i \neq j \\ 0 & \text{if } (i,j) \notin E, i \neq j \end{cases}$$

- $L$ is symmetric positive semidefinite
- The smallest eigenvalue of $L$ is $lambda_1 = 0$, with
- corresponding eigenvector $w_1 = (1, 1, ... , 1)^T$

# Spectral analysis

- For the second smallest eigenvector *lambda₂* of L

$$\lambda_2 = \min_{\substack{\mathbf{x}^T\mathbf{w}_1=0 \\ ||\mathbf{x}||=1}} \mathbf{x}^T L\mathbf{x} = \min_{\sum x_i=0} \frac{\sum_{(i,j)\in E}(x_i - x_j)^2}{\sum_i x_i^2}$$

- Corresponding eigenvector $w_2$ is called Fielder vector
- The ordering according to the values of $w_2$ will group similar (connected) vertices together
- Physical interpretation: The stable state of springs placed on the edges of the graph, when graph is forced to 1 dimension

# Spectral partition

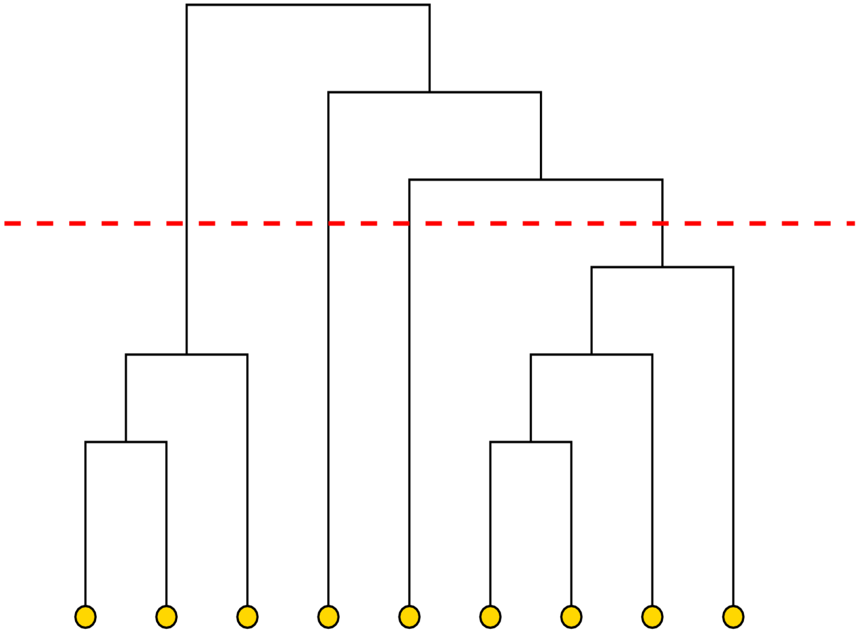
- Partition the nodes according to the ordering induced by the Fielder vector

- Some partitioning rules:
- Bisection: use the median value in $w_2$
- Cut ratio: find the partition that minimizes
- Sign: Separate positive and negative values
- Gap: Separate according to the largest gap in the values of $w_2$

- Spectral partition works very well in practice
- However, not scalable

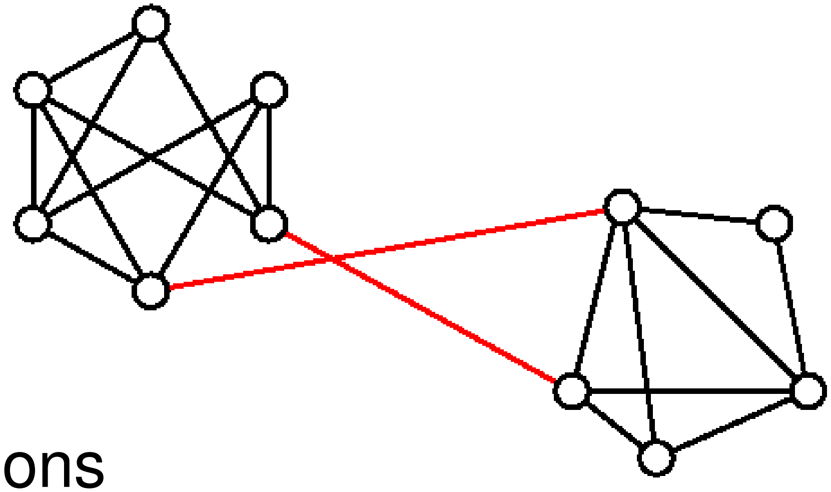
# Top down algorithms

- [Newman and Girvan, 2004]
- A set of algorithms based on removing edges from the graph, one at a time
- The graph gets progressively disconnected, creating a hierarchy of communities



# Top down algorithms

- *Select  edge to remove based on ”betweenenss”*

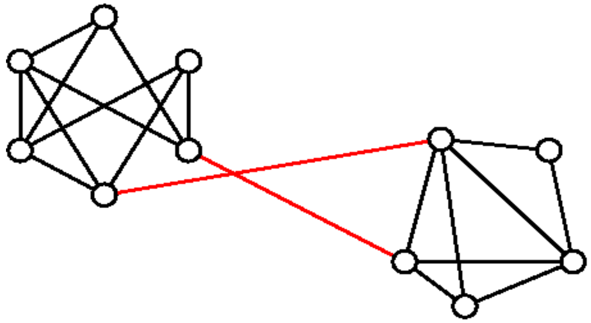


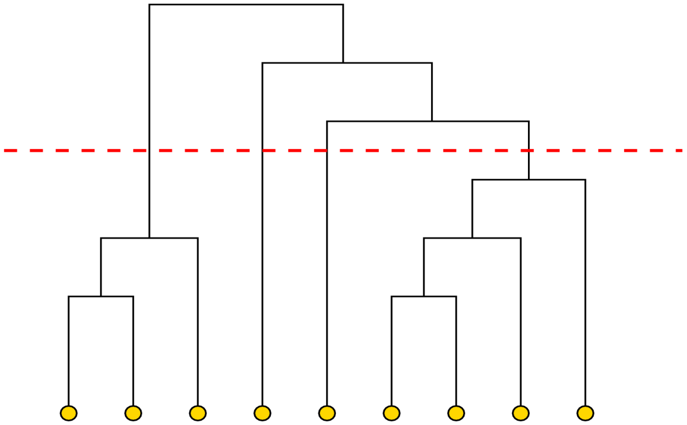
- Three definitions
- Shortest-path betweeness: Number of shortest paths that the edge belongs to
- Random-walk betweeness: Expected number of paths for a random walk from u to v
- Current-flow betweeness: Resistance derived from considering the graph as an electric circuit

# Generic top-down algorithm

- Top down

- Compute betweeness value of all edges
- [Recompute betweeness vlaue of all remaining edges]
- Remove the edge with the highest betweeness
- Repeat until no edges left

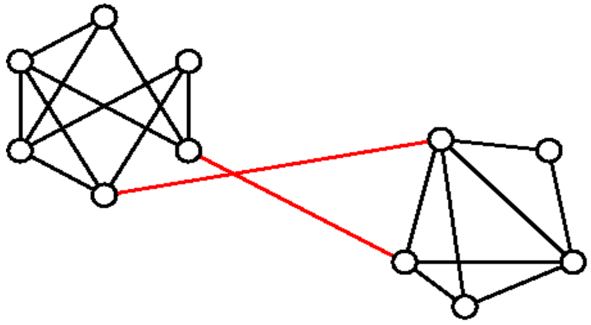


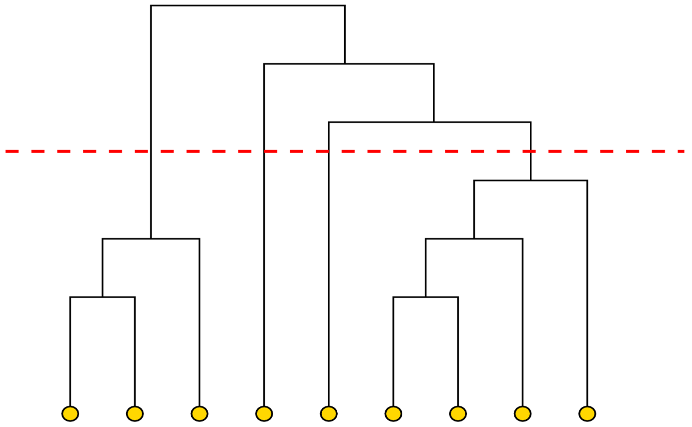


# Modularity measure

- How to pick the right clustering from the whole hierarchy?
- Modularity measure [Newman and Girvan, 2004]
- Compared with a “random clustering”

- Direct optimization of modularity measure by
  - Agglomerative [Newman and Girvan, 2004]
  - Spectral [White and Smyth, 2005]

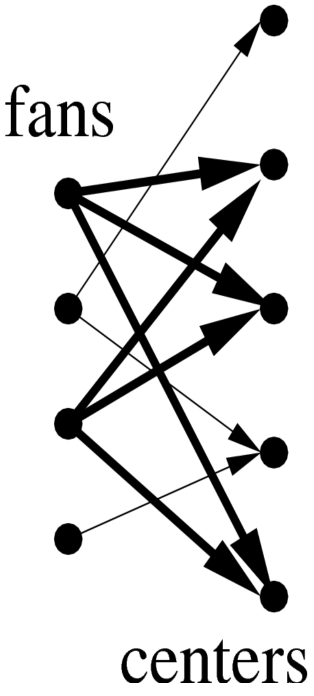




# Scaling up

- How to find communities on a large graph, say, the Web?
- Web communities are characterized by dense directed bipartite graphs [Kumar et al., 1999]
- Idea similar to hubs and authorities
- Example: Pages of sport cars (Lotus, Ferrari, Lamborghini) and enthusiastic fans
- Bipartite cores: Complete bipartite cliques contained in a community
- Support from random graph theory: If $G = (U, V, E)$ is a dense bipartite graph, then w.h.p. there is a $K_{i,j}$, for some $i$ and $j$

# Detecting communities by trawling



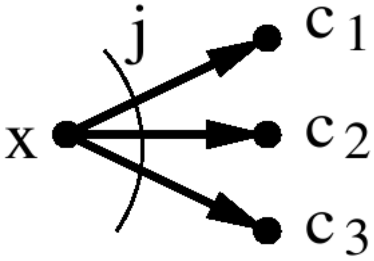
- Many pruning phases

- Heuristic pruning (quality consideration)
  - Fans should point to at least 6 different hosts
  - Centers should be pointed by at most 50 fans

- Degree-based pruning
  - For a fan to participate in a $K_{i,j}$, it should have out-degree at least *j*
  - For a center to participate in a $K_{i,j}$, it should have in-degree at least *i*
  - Prune iteratively fans and centers
  - Can be done efficiently by sorting edges:
  - Sort edges by src to prune fans
  - Sort edges by dst to prune centers

# Detecting communities by trawling

- Inclusion-exclusion pruning
  - Either a core is output or a vertex is pruned
- Computation is organized so that pruning is done with successive passes on the data



- A-priori pruning
  - Cores satisfy monotonicity
  - If *(X,Y)* is a $K_{i,j}$ then every *(X',Y)* with $X' \subseteq X$ is a $K_{i',j}$
  - A-priori algorithm: start with *(1,j), (2,j), ...*
  - Most computationally demanding phase, but the graph is already heavily pruned

# Conclusions (communities)

- Finding communities
- What is the right objective?
- Designing scalable algorithms is challenging
- How to evaluate the results?
- Studying dynamics and evolution of communities