Navigating the Web graph

Workshop on Networks and Navigation
Santa Fe Institute, August 2008

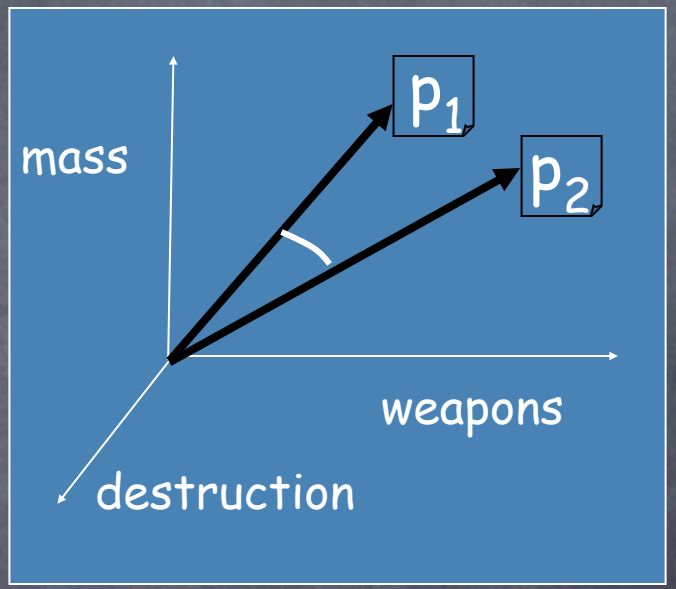
Filippo Menczer
Informatics & Computer Science
Indiana University, Bloomington

Outline

- Topical locality: Content, link, and semantic topologies
- Implications for growth models and navigation
- Applications
  - Topical Web crawlers
  - Distributed collaborative peer search

The Web as a text corpus

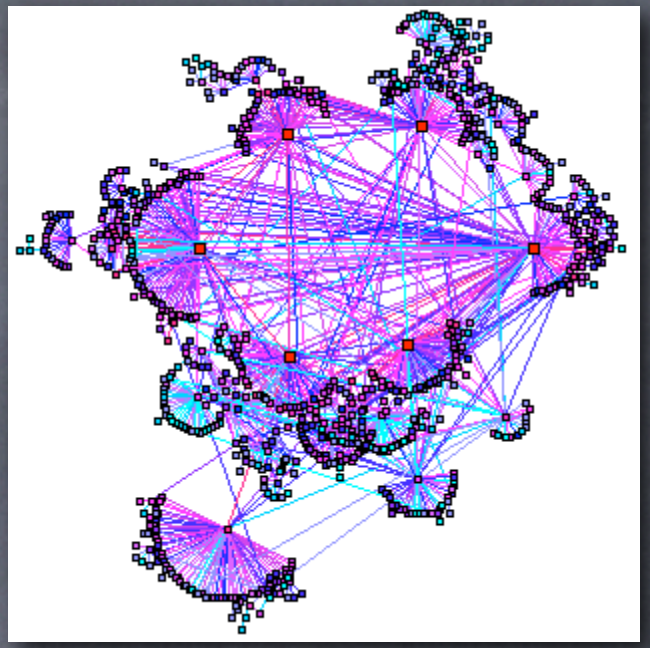
- Pages close in word vector space tend to be related
  - Cluster hypothesis (van Rijsbergen 1979)
  - The WebCrawler (Pinkerton 1994)
  - The whole first generation of search engines



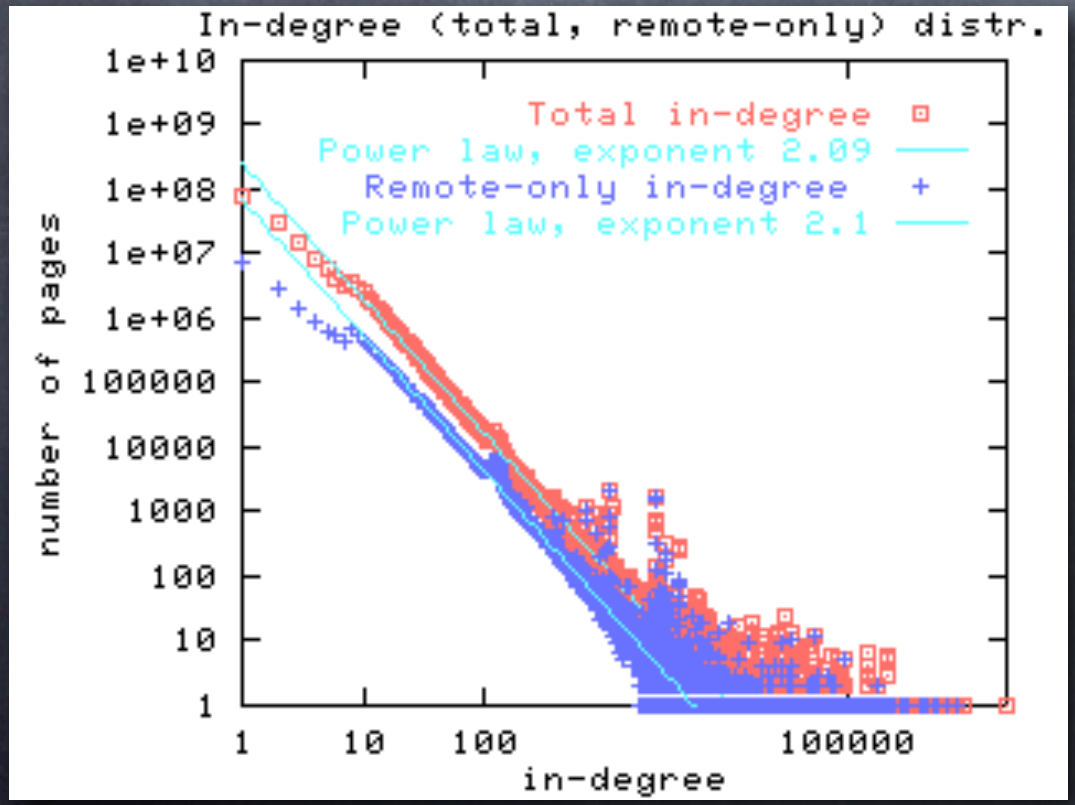
Enter the Web's link structure

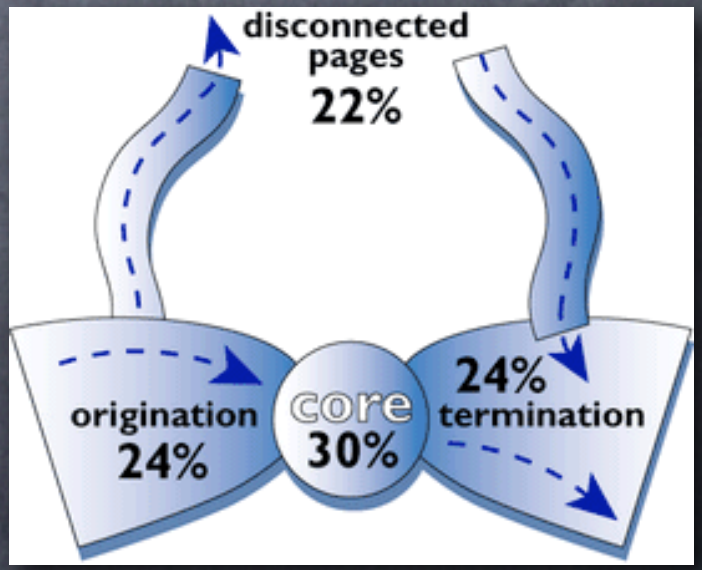
$$p(i) = \frac{\alpha}{N} + (1-\alpha) \sum_{j:j\to i} \frac{p(j)}{|\ell : j \to \ell|}$$

Brin & Page 1998



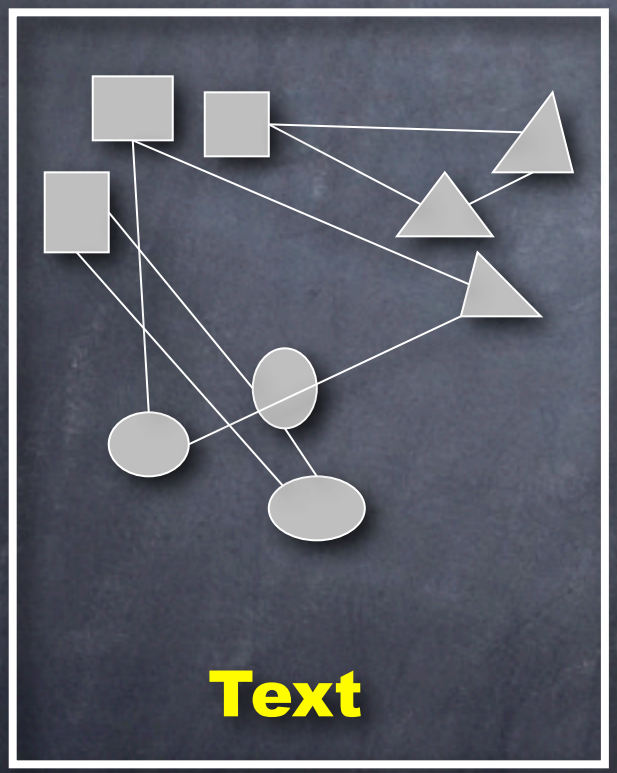
Barabasi & Albert 1999

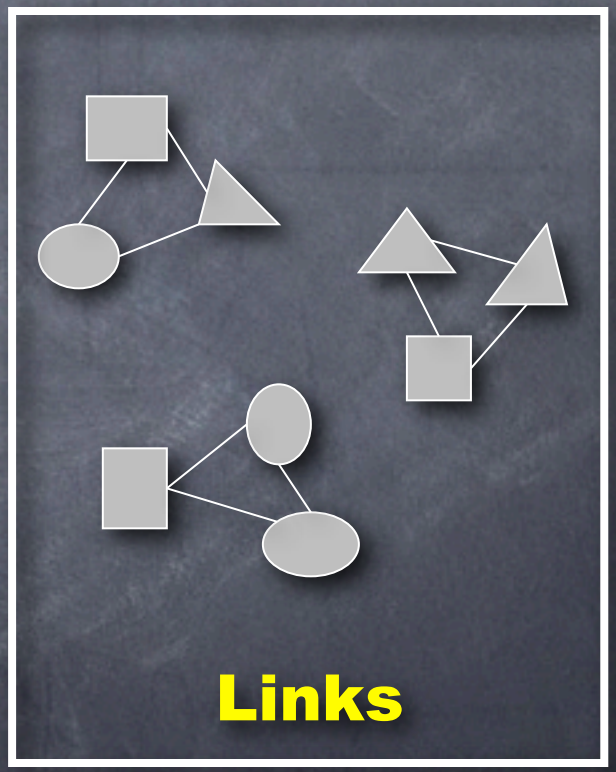




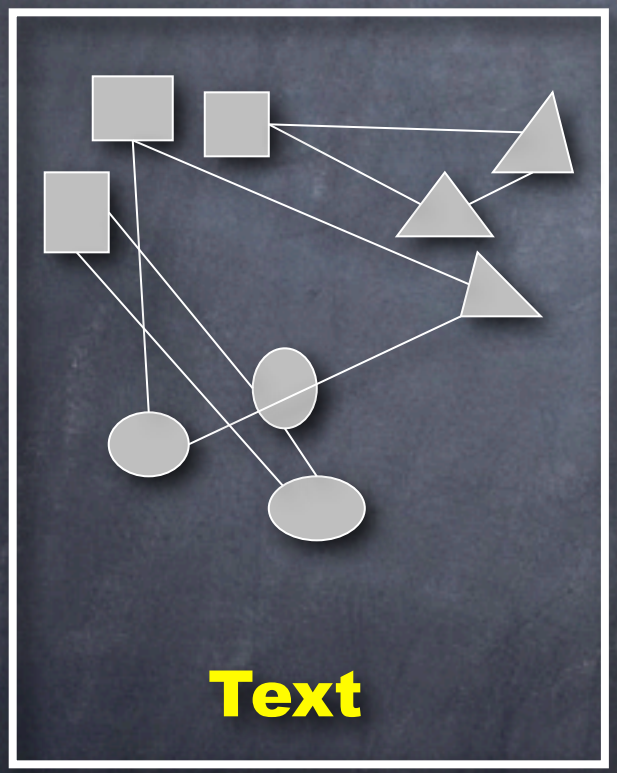
Broder & al. 2000

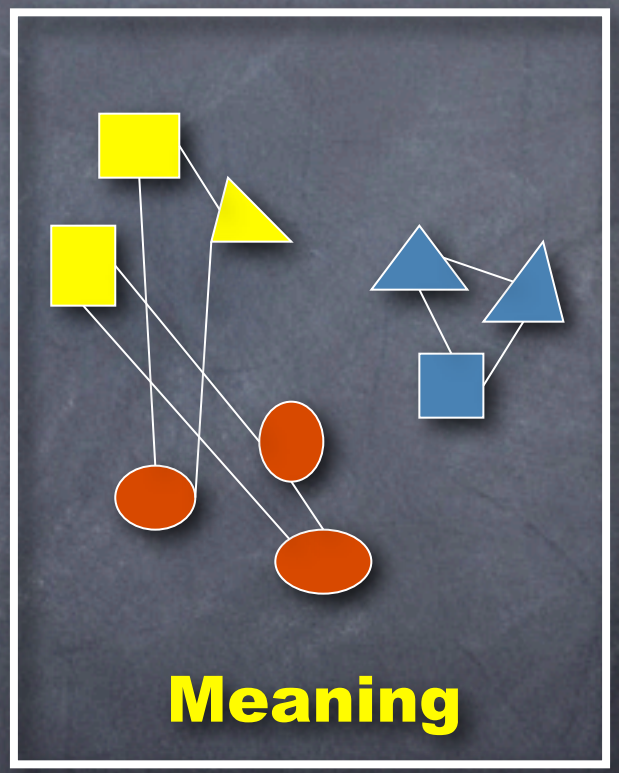
Three network topologies

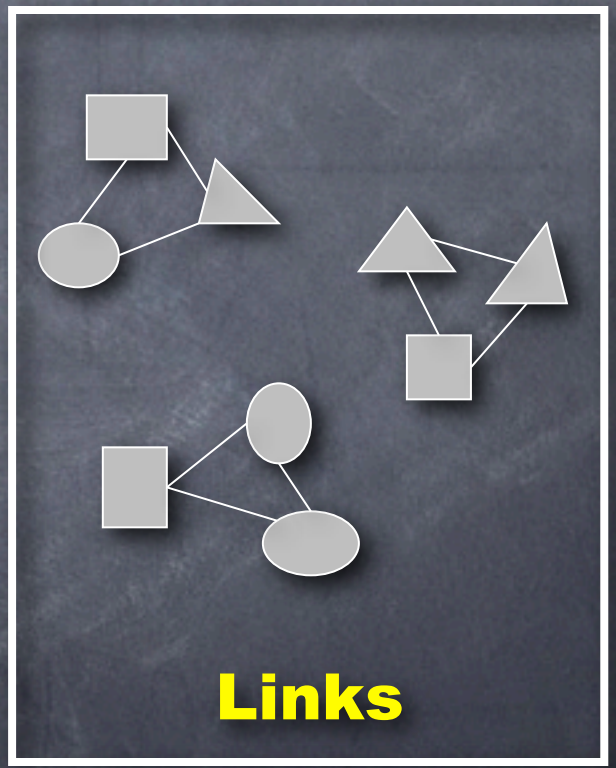




Three network topologies

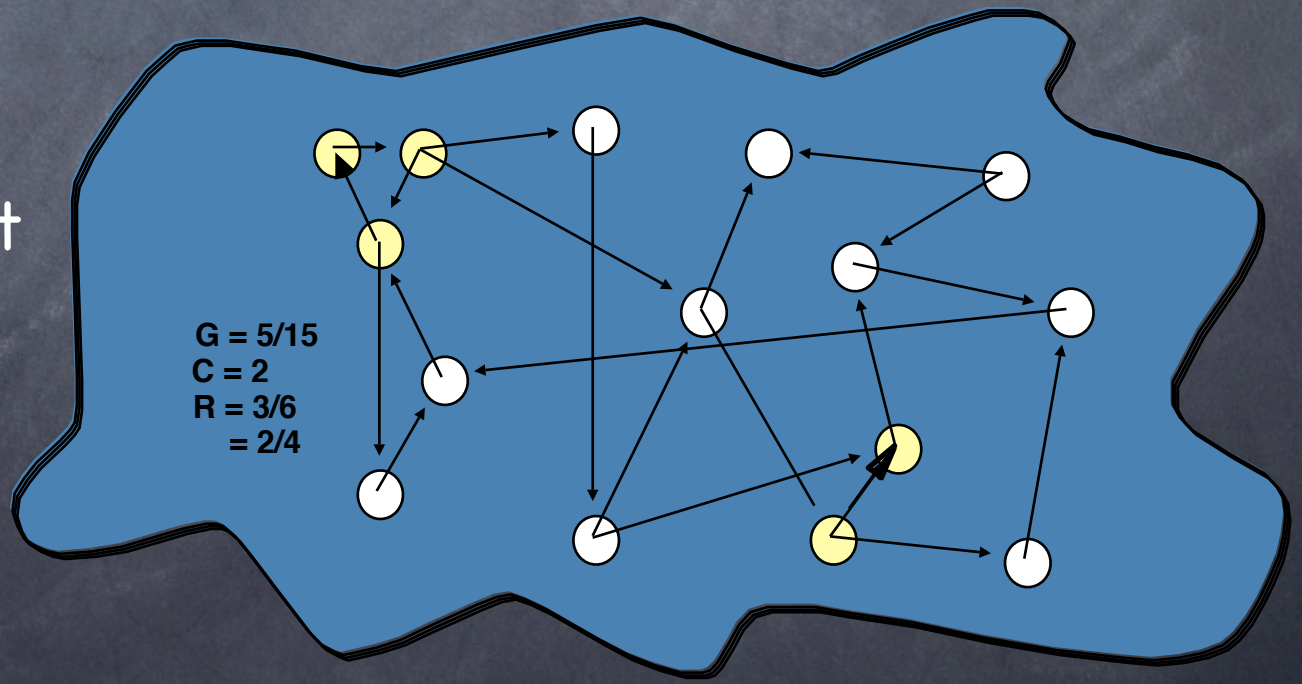






The "link-cluster" conjecture

- Connection between semantic topology (topicality or relevance) and link topology (hypertext)
  - G = Pr[rel(p)] ~ fraction of relevant pages (generality)
  - R = Pr[rel(p) | rel(q) AND link(q,p)]
- Related nodes are "clustered" if **R > G** (modularity)
  - Necessary and sufficient condition for a random crawler to find pages related to start points



ICML 1997

Link-cluster conjecture

- Stationary hit rate for a random crawler:

$$\eta(t+1) = \eta(t) \cdot R + (1 - \eta(t)) \cdot G \geq \eta(t)$$

$$\eta \xrightarrow{t \to \infty} \eta^* = \frac{G}{1 - (R - G)}$$

$$\eta^* > G \Leftrightarrow R > G$$

Conjecture

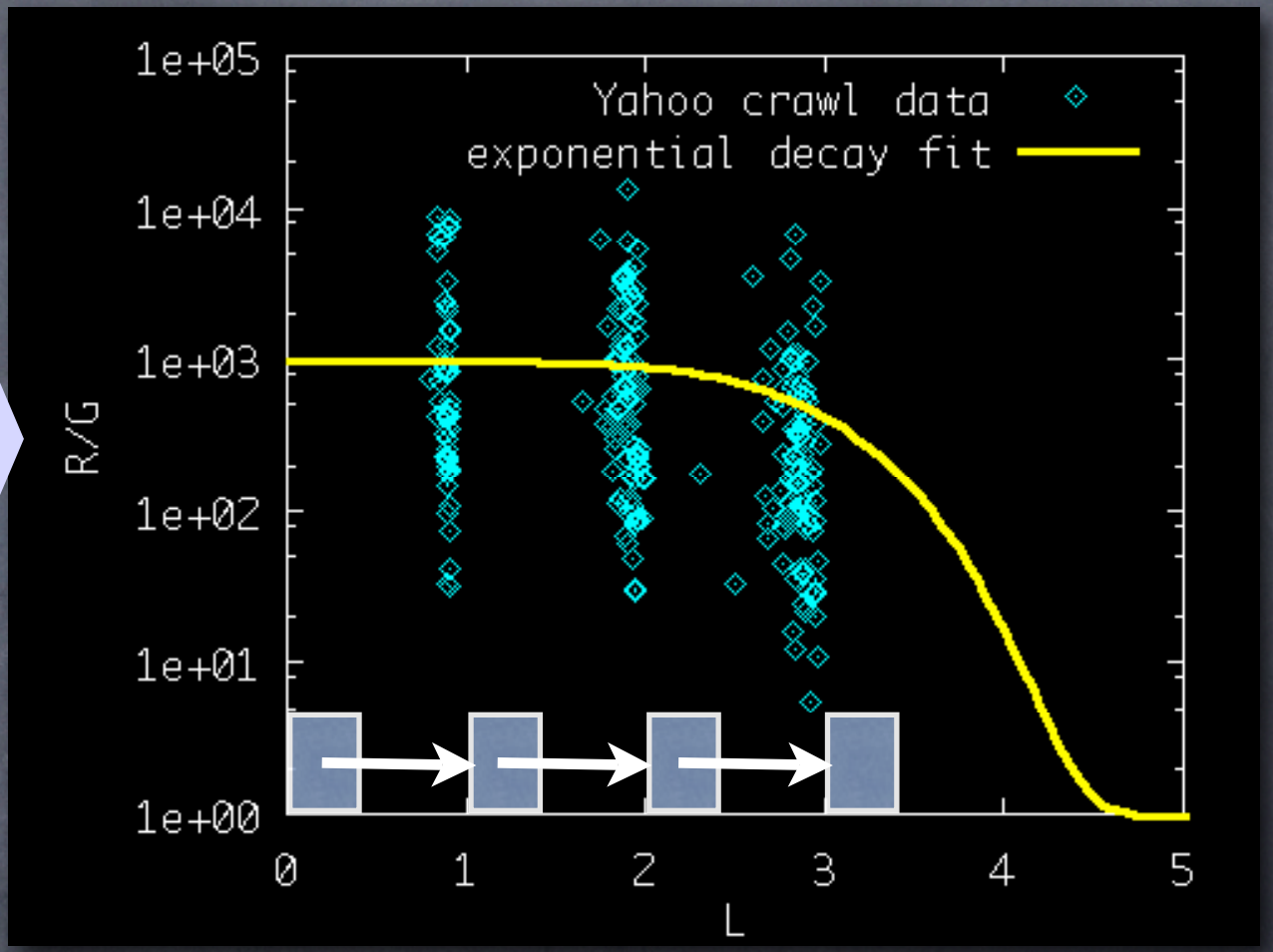
$$\frac{\eta^*}{G} - 1 = \frac{R - G}{1 - (R - G)}$$

Value added

$$\frac{R(q,\delta)}{G(q)} \equiv \frac{\Pr\left[rel(p) \mid rel(q) \wedge \|path(q,p)\| \le \delta\right]}{\Pr[rel(p)]}$$

Link-cluster conjecture

- Pages that link to each other tend to be related
- Preservation of semantics (meaning)
- A.k.a. topic drift



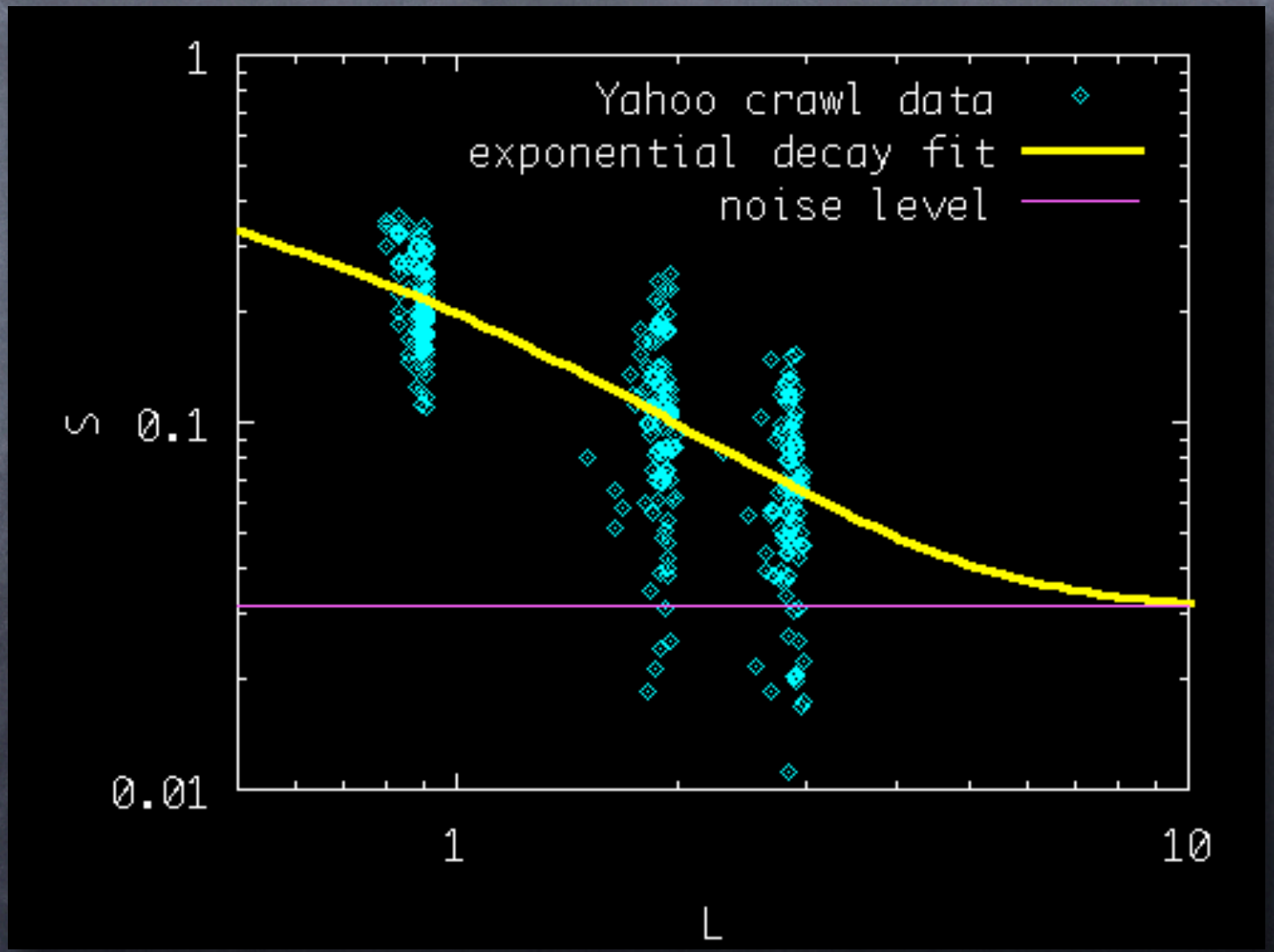
$$L(q,\delta) \equiv \frac{\sum_{\{p:\|path(q,p)\| \le \delta\}} \|path(q,p)\|}{\left|\{p : \|path(q,p)\| \le \delta\}\right|}$$

JASIST 2004

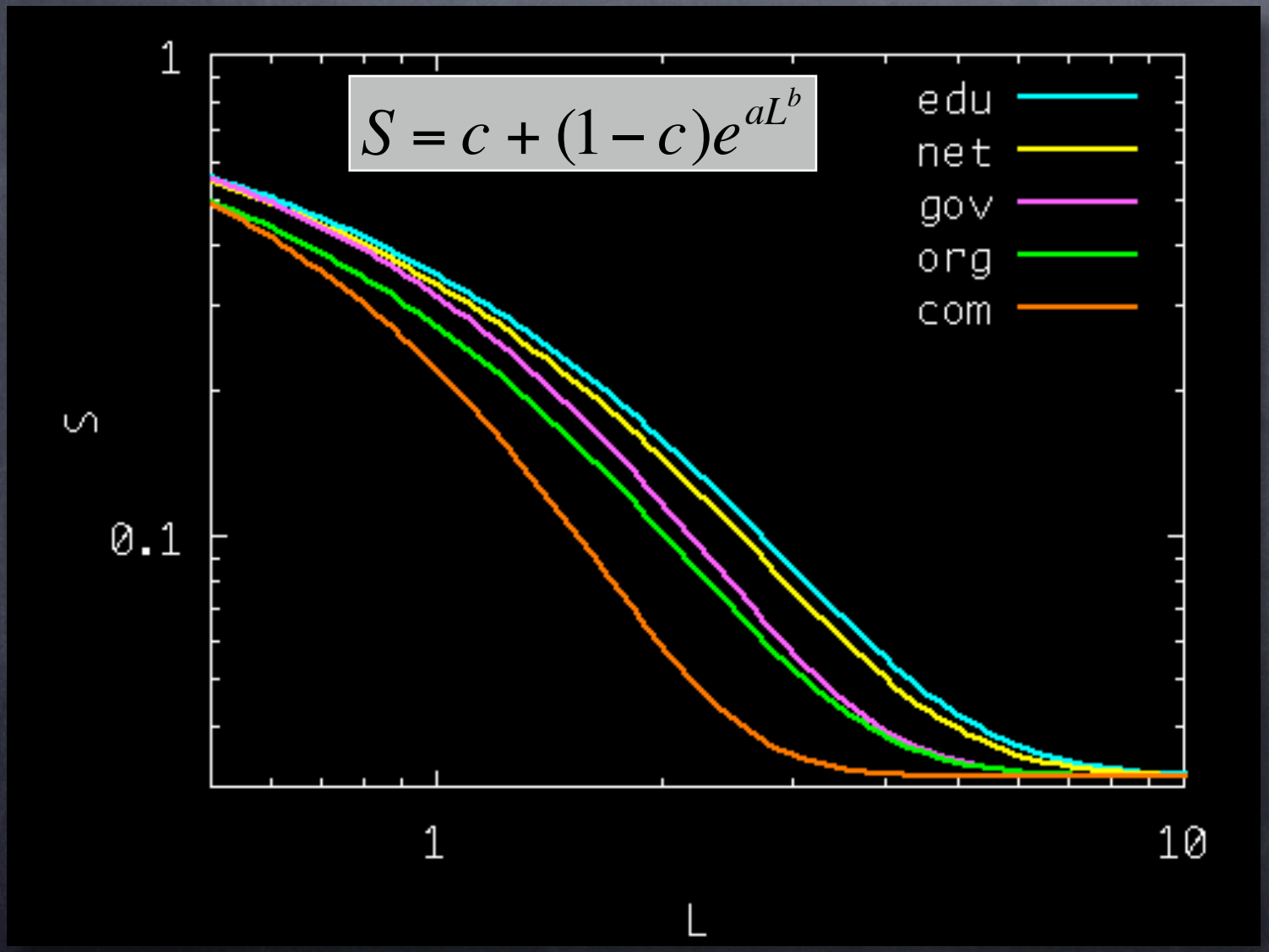
The "link-content" conjecture

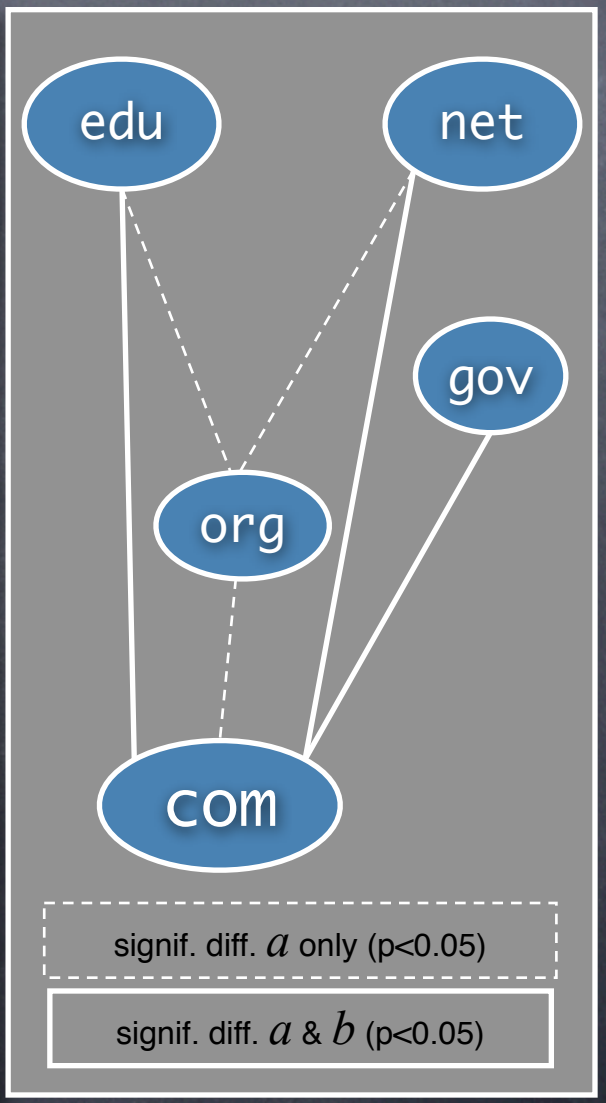
$$S(q,\delta) \equiv \frac{\sum_{\{p:\|path(q,p)\|\le\delta\}} sim(q,p)}{\left|\{p:\|path(q,p)\|\le\delta\}\right|}$$

- Correlation of lexical and linkage topology
- **L(δ)**: average link distance
- **S(δ)**: average similarity to start (topic) page from pages up to distance δ
- Correlation ρ(L,S) = -0.76



Heterogeneity of link-content correlation

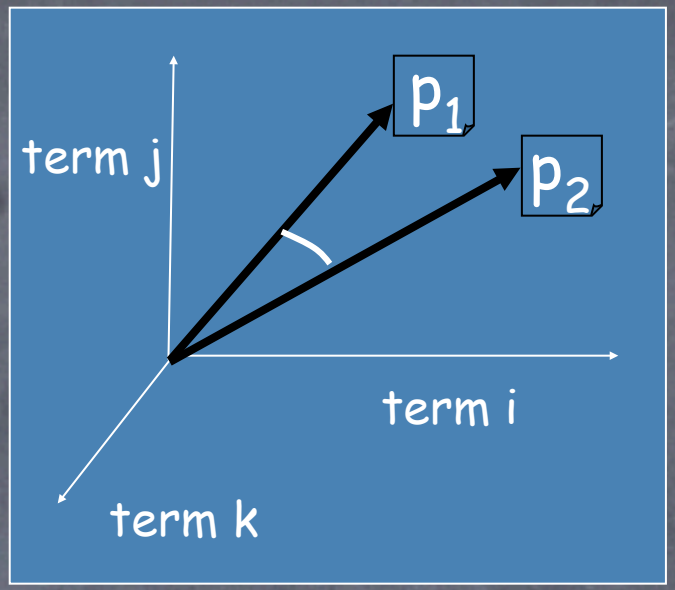




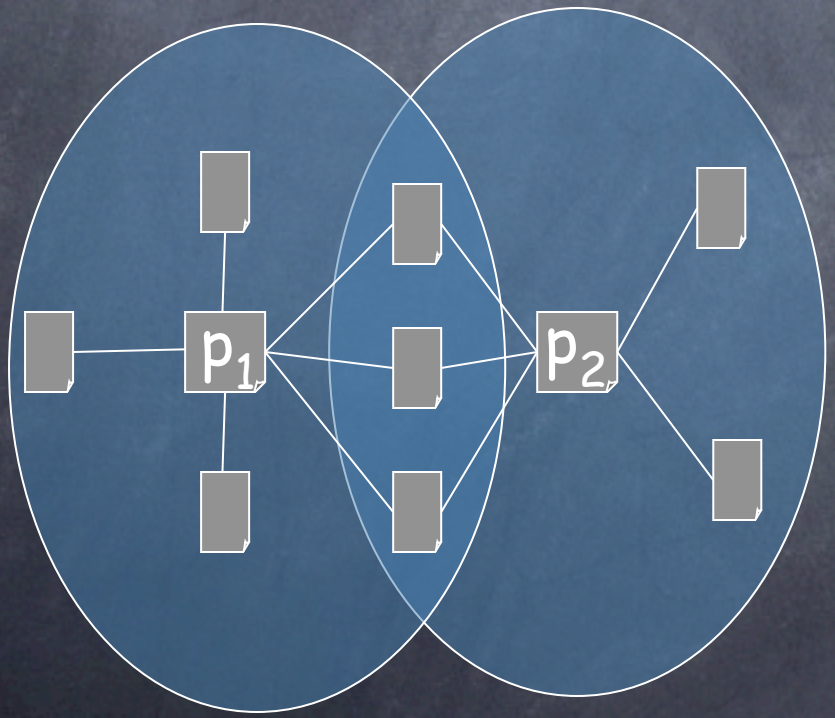
Mapping the relationship between links, content, and semantic topologies

- Given any pair of pages, need 'similarity' or 'proximity' metric for each topology:
  - Content: textual/lexical (cosine) similarity
  - Link: co-citation/bibliographic coupling
  - Semantic: relatedness inferred from manual classification
- Data: Open Directory Project (dmoz.org)
  - ~ 1 M pages after cleanup
  - ~ $1.3*10^{12}$ page pairs!

$$\sigma_c\left(\overrightarrow{p_1}, \overrightarrow{p_2}\right) = \frac{\overrightarrow{p_1} \cdot \overrightarrow{p_2}}{\left\|\overrightarrow{p_1}\right\| \cdot \left\|\overrightarrow{p_2}\right\|}$$



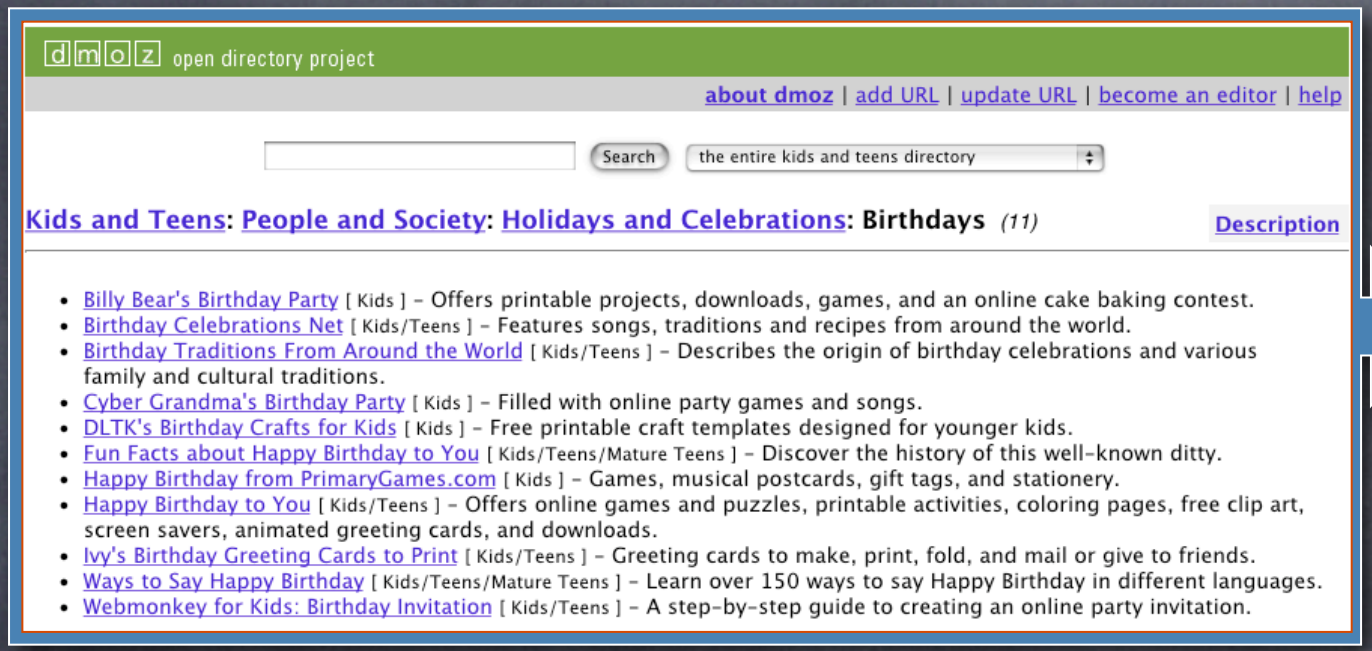
Content similarity

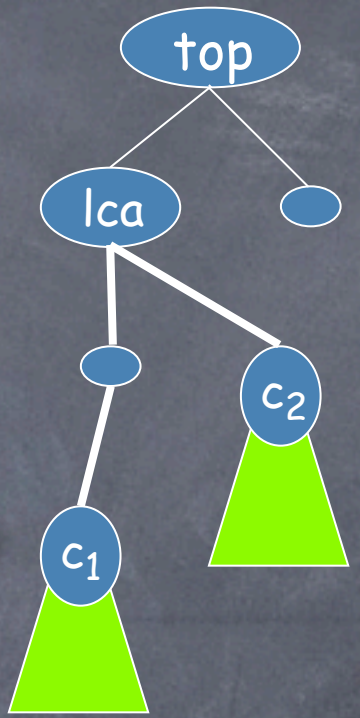


Link similarity

$$\sigma_l(p_1, p_2) = \frac{\left|U_{p_1} \cap U_{p_2}\right|}{\left|U_{p_1} \cup U_{p_2}\right|}$$

Semantic similarity



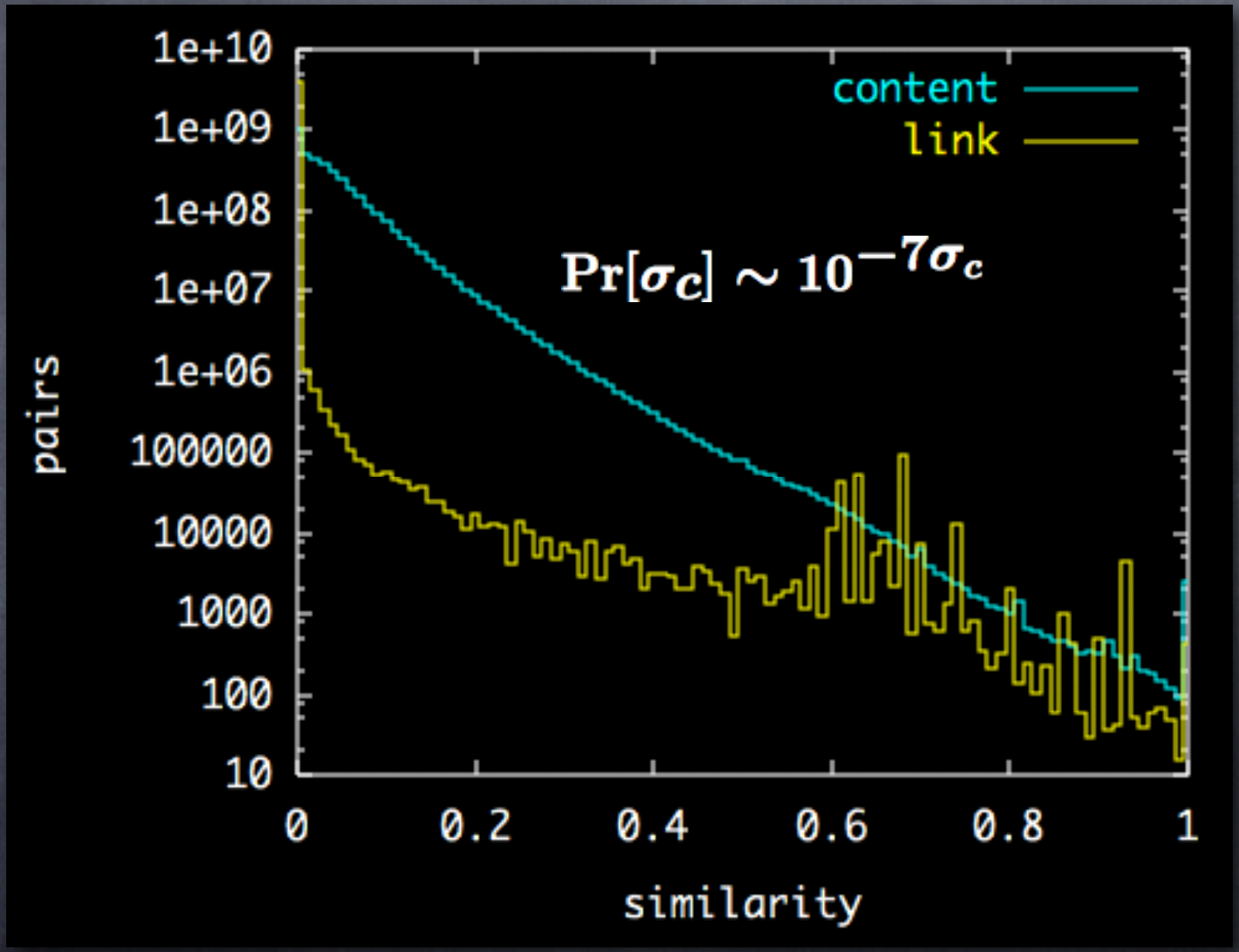


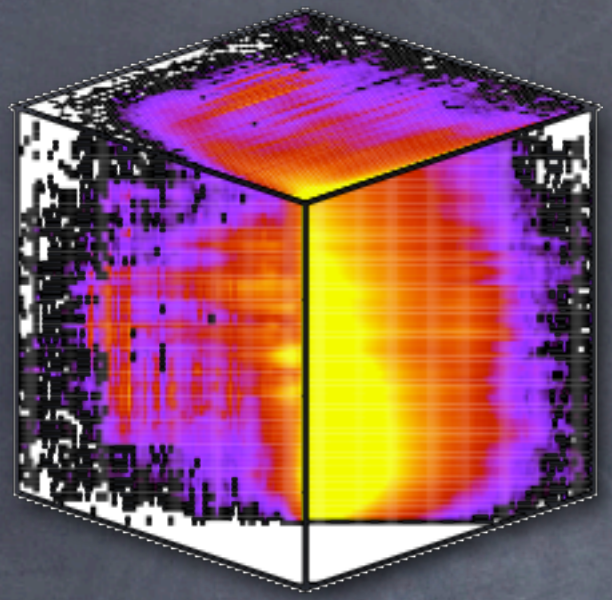
- Information-theoretic measure based on classification tree (Lin 1998)

$$\sigma_s(c_1,c_2) = \frac{2\log \Pr[\text{lca}(c_1,c_2)]}{\log \Pr[c_1] + \log \Pr[c_2]}$$

- Classic path distance in special case of balanced tree

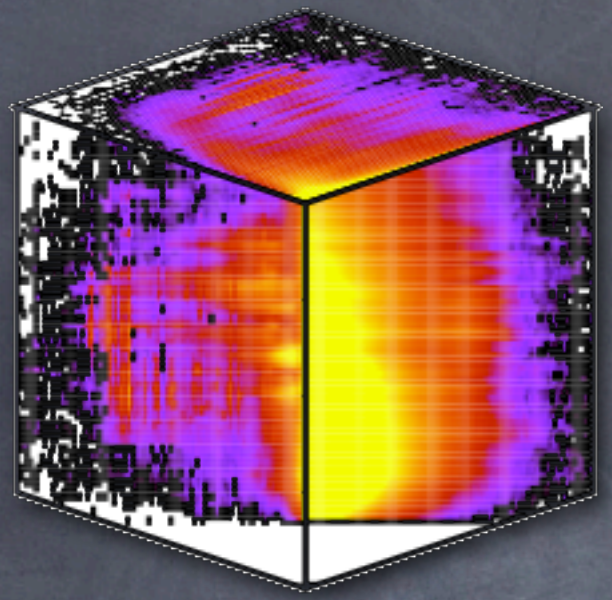
Individual metric distributions





$$\text{Precision} = \frac{|\ \text{Retrieved} \ \&\ \text{Relevant}\ |}{|\ \text{Retrieved}\ |}$$

$$\text{Recall} = \frac{|\ \text{Retrieved} \ \&\ \text{Relevant}\ |}{|\ \text{Relevant}\ |}$$



$$\text{Precision} = \frac{|\ \text{Retrieved} \ \&\ \text{Relevant}\ |}{|\ \text{Retrieved}\ |}$$

$$\text{Recall} = \frac{|\ \text{Retrieved} \ \&\ \text{Relevant}\ |}{|\ \text{Relevant}\ |}$$

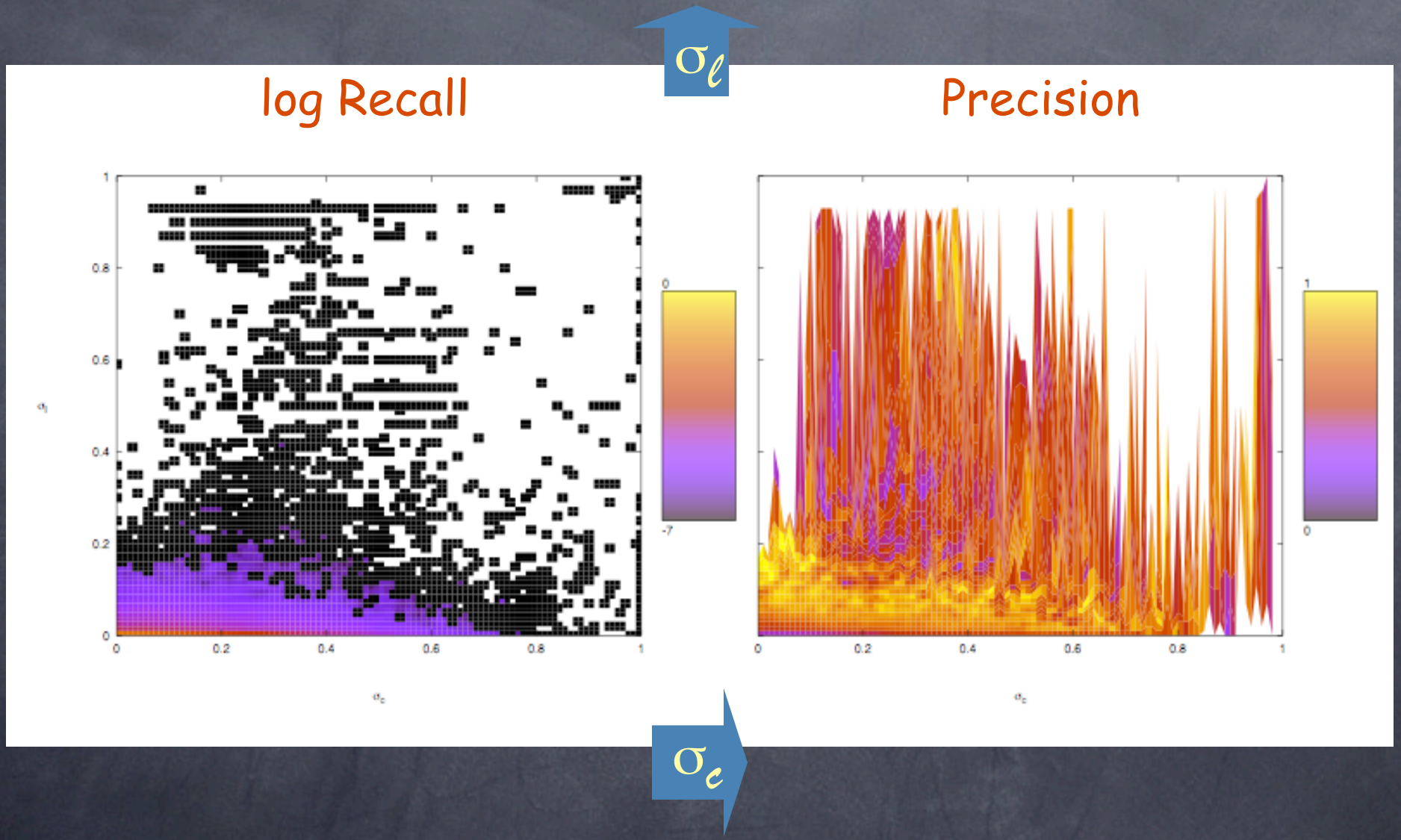
$$P(s_c, s_l) = \frac{\sum_{\{p,q:\sigma_c = s_c, \sigma_l = s_l\}} \sigma_s(p,q)}{\left|\{p,q : \sigma_c = s_c, \sigma_l = s_l\}\right|}$$

Averaging semantic similarity

$$R(s_c, s_l) = \frac{\sum_{\{p,q:\sigma_c = s_c, \sigma_l = s_l\}} \sigma_s(p,q)}{\sum_{\{p,q\}} \sigma_s(p,q)}$$

Summing semantic similarity

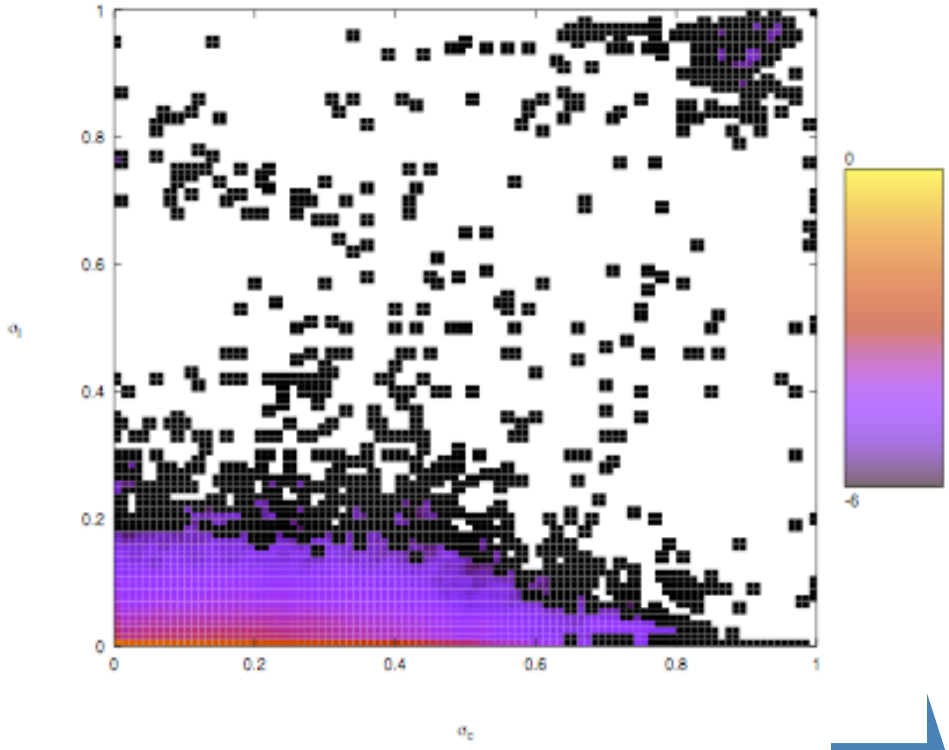
Science


σ_ℓ
log Recall
Precision
1
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1
0
-7
1
0
σ_c
σ_c
σ_c

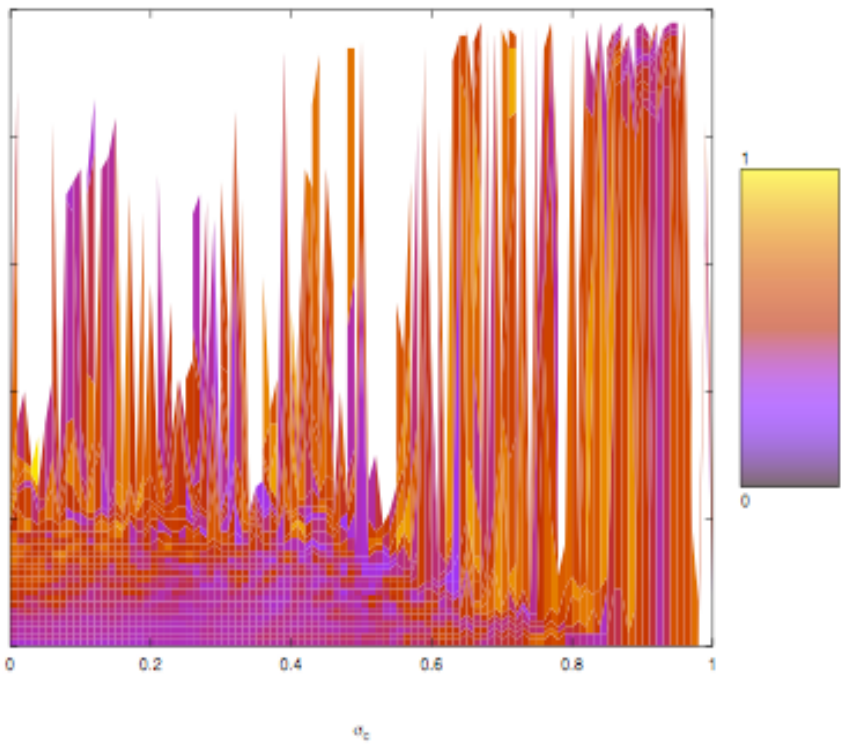
Adult

σ_ℓ

log Recall

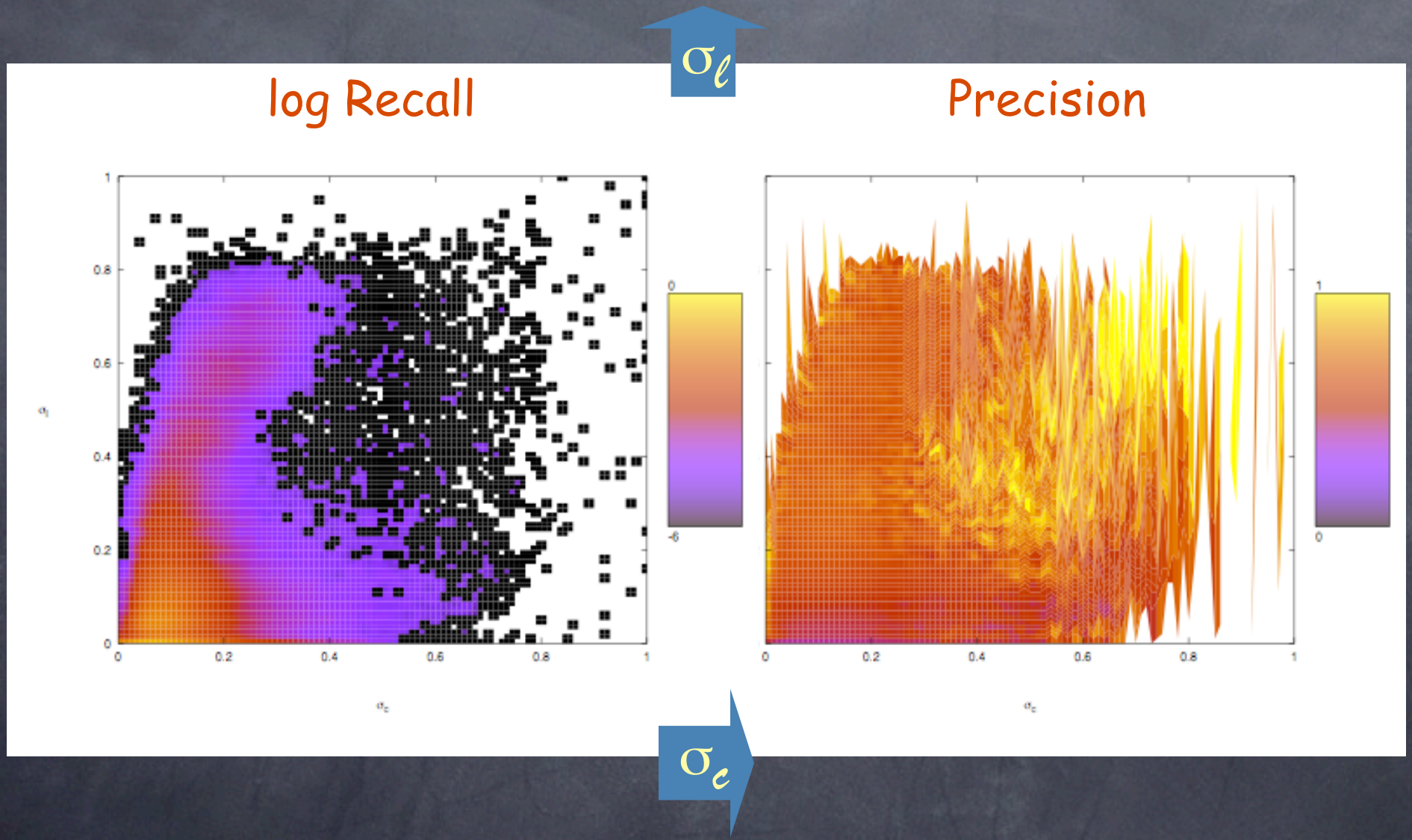


Precision

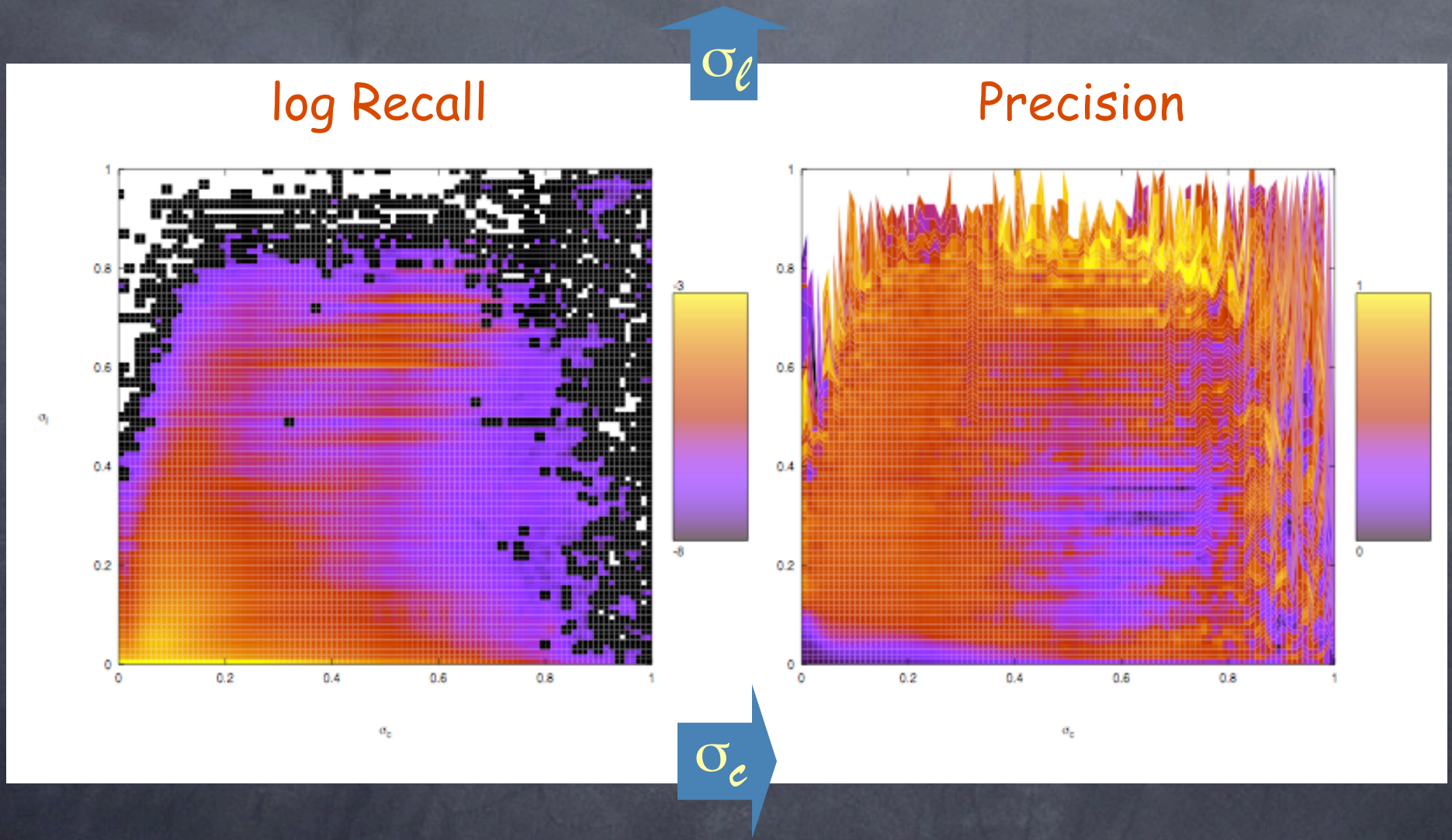


σ_c

News


σ_ℓ
log Recall
Precision
1
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1
σ_l
σ_c
0
-6
1
0
σ_c
σ_c

All pairs


log Recall
Precision
σ_ℓ
σ_c
σ_l
1
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1
-3
-8
0

Outline

- Topical locality: Content, link, and semantic topologies
- Implications for growth models and navigation
- Applications
  - Topical Web crawlers
  - Distributed collaborative peer search

Link probability vs lexical distance

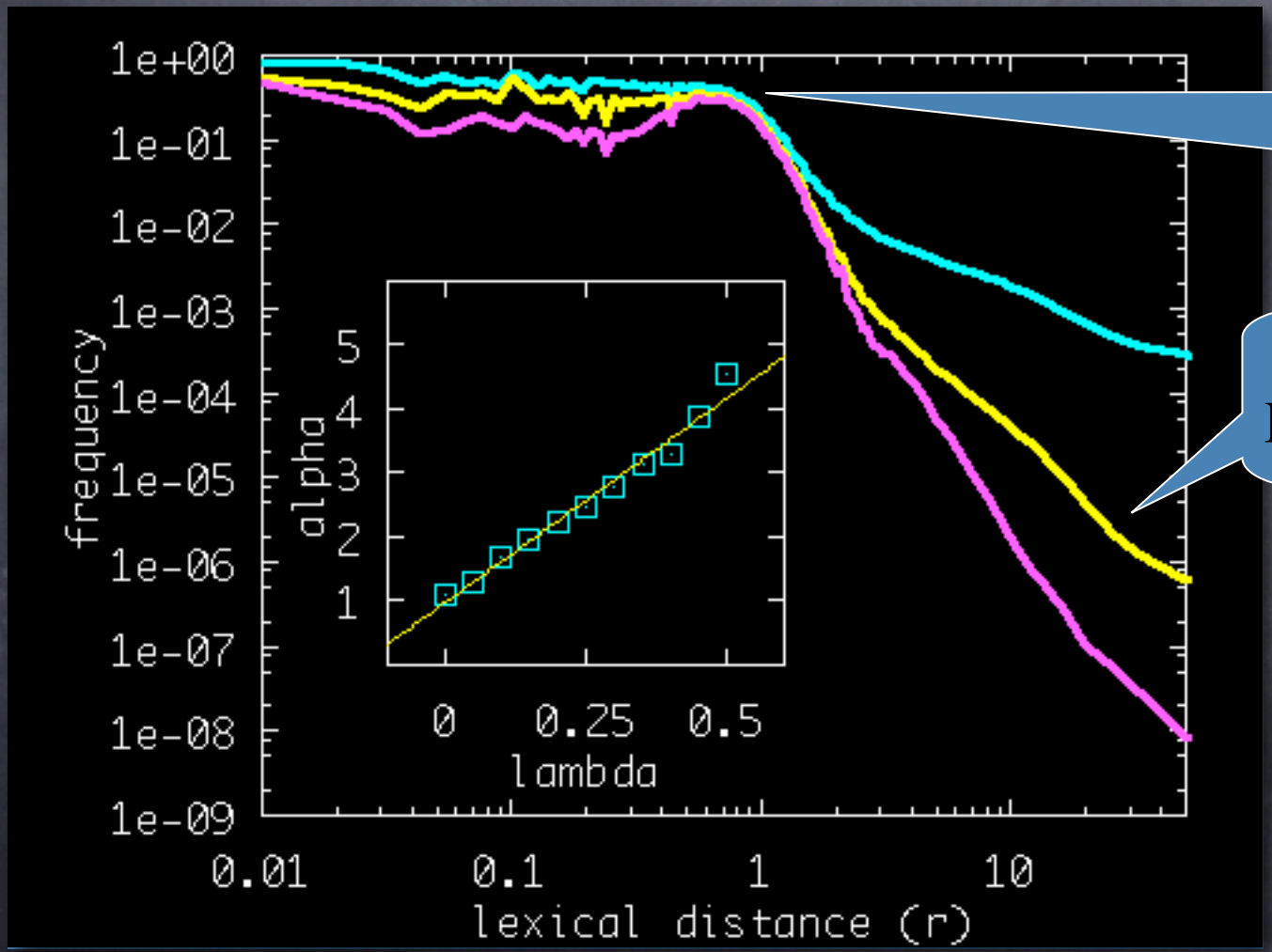
$$r = 1/\sigma_c - 1$$

$$\Pr(\lambda \mid \rho) = \frac{\left|(p,q) : r = \rho \wedge \sigma_l > \lambda\right|}{\left|(p,q) : r = \rho\right|}$$

Link probability vs lexical distance

$$r = 1/\sigma_c - 1$$

$$\Pr(\lambda \mid \rho) = \frac{\left|(p,q) : r = \rho \wedge \sigma_l > \lambda\right|}{\left|(p,q) : r = \rho\right|}$$



Phase transition ρ^*

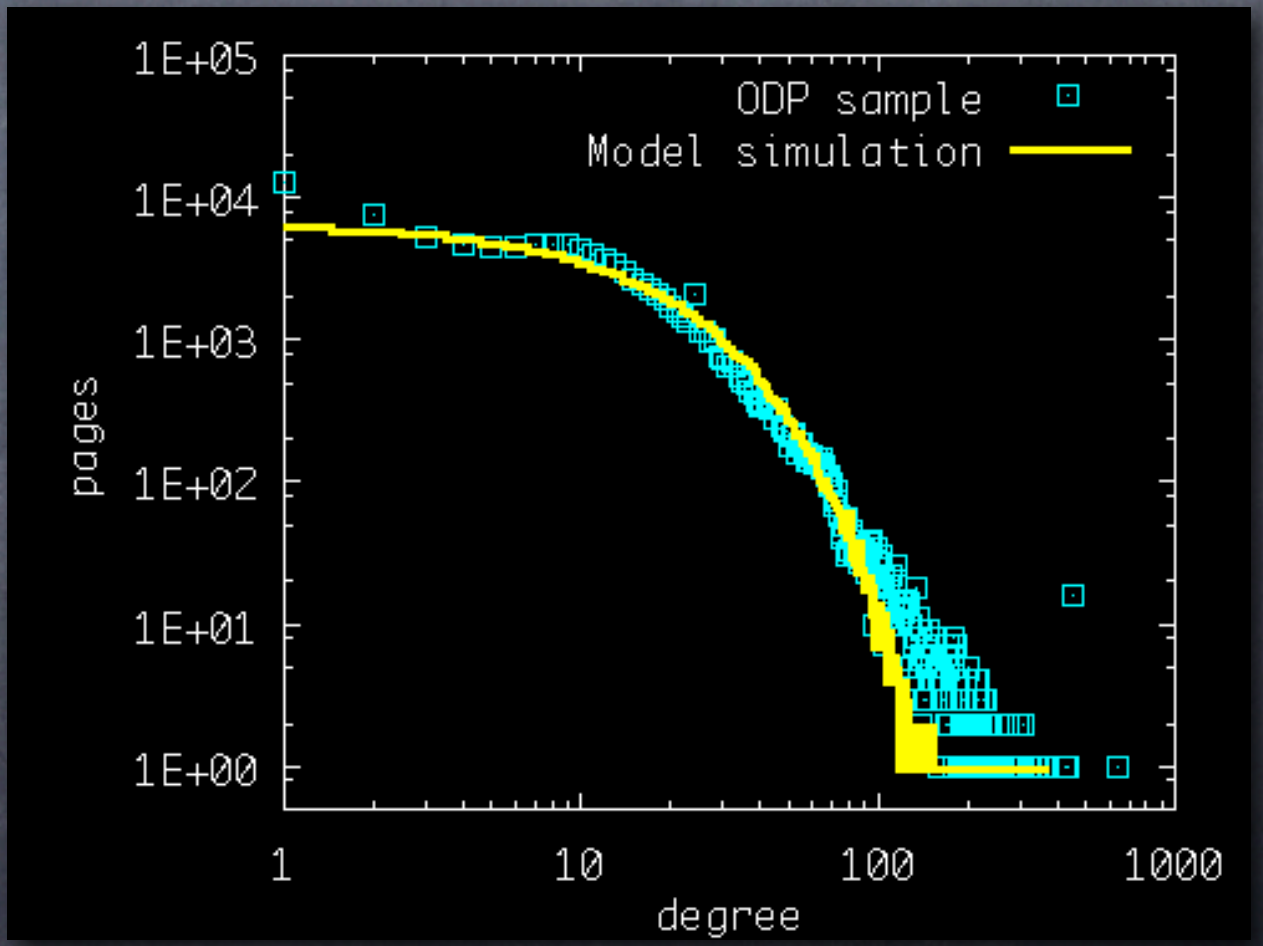
Power law tail $\Pr(\lambda \mid \rho) \sim \rho^{-\alpha(\lambda)}$

Proc. Natl. Acad. Sci. USA 99(22): 14014-14019, 2002

Local content-based growth model

$$\Pr(p_t \rightarrow p_{i<t}) = \begin{cases} \dfrac{k(i)}{mt} & \text{if } r(p_i, p_t) < \rho^* \\ c[r(p_i, p_t)]^{-\alpha} & \text{otherwise} \end{cases}$$

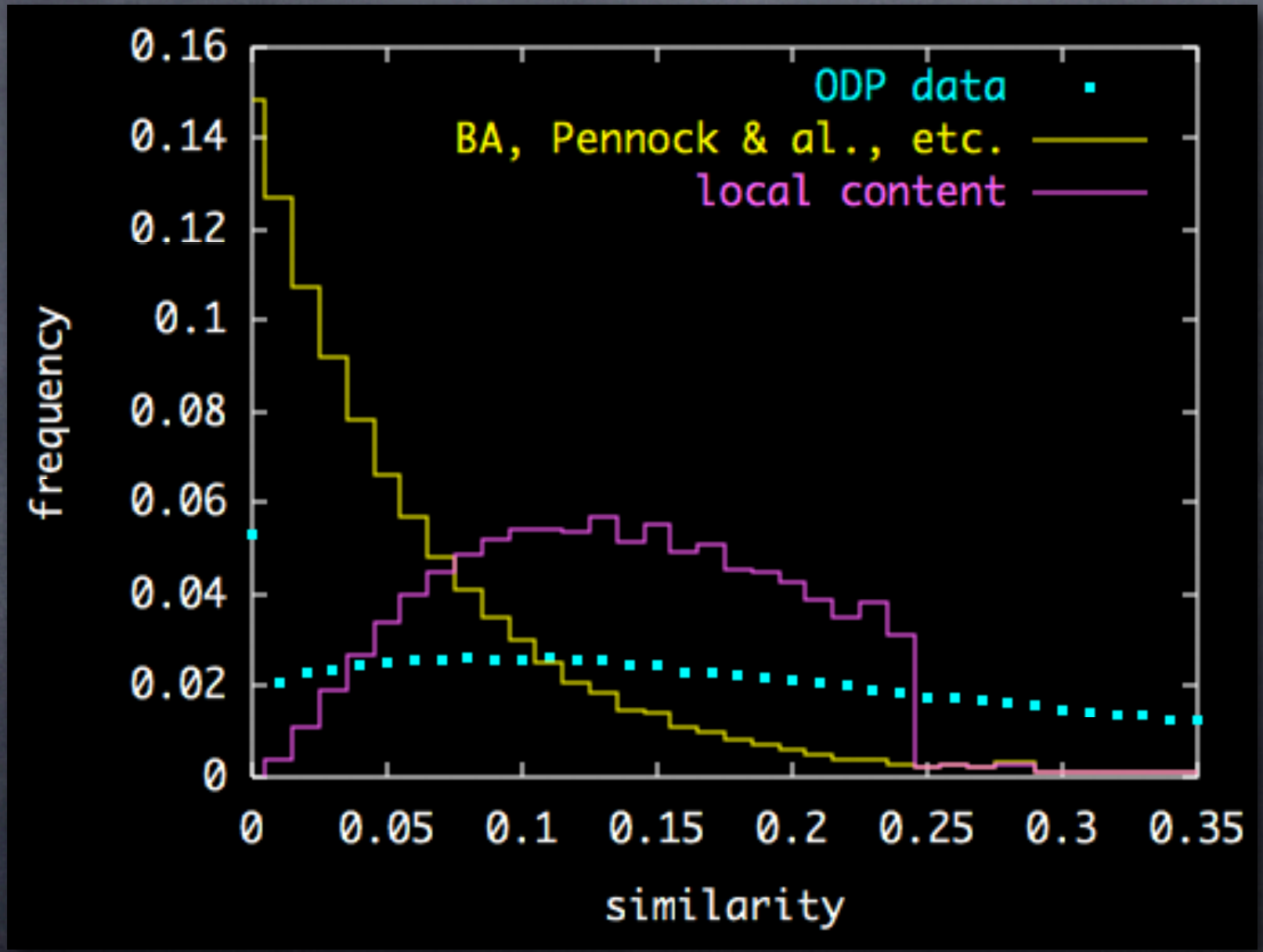
- Similar to preferential attachment (BA)
- Use degree info (popularity/ importance) only for nearby (similar/ related) pages



So, many models can predict degree distributions...

- Which is "right" ?
- Need an independent observation (other than degree) to validate models
- Distribution of content similarity across linked pairs

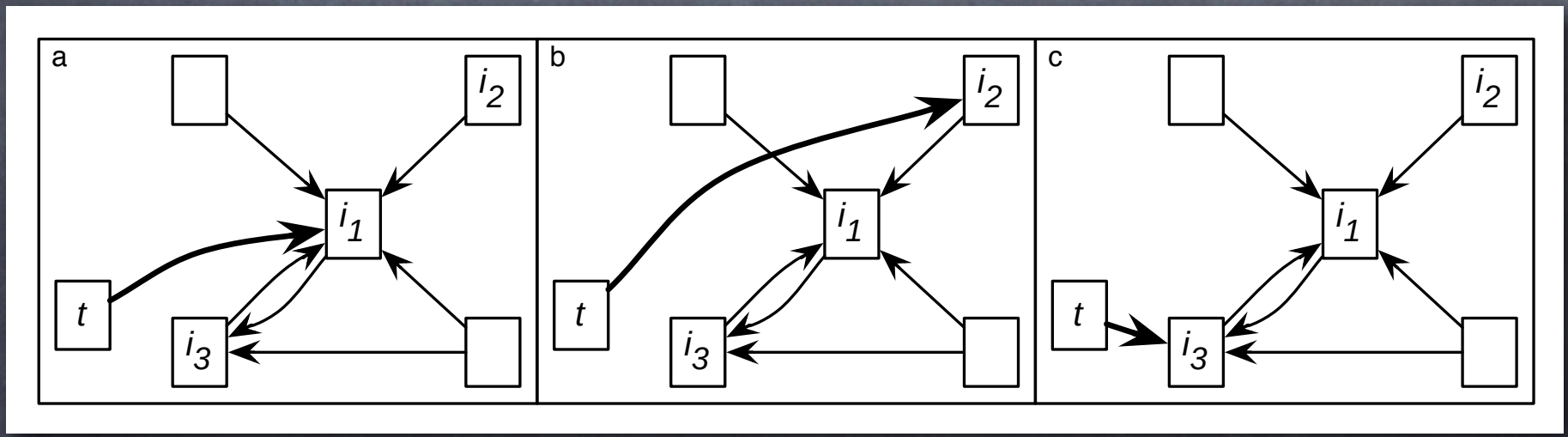
None of these models is right!



The mixture model

$$\Pr(i) \propto \psi \cdot \frac{1}{t} + (1-\psi) \cdot \frac{k(i)}{mt}$$

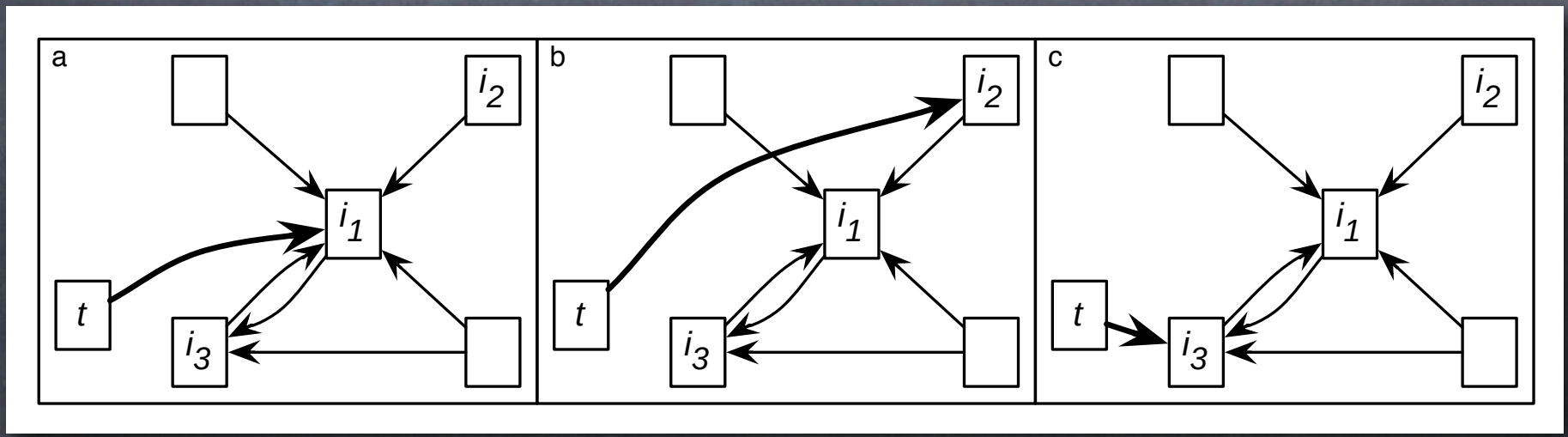
degree-uniform mixture



The mixture model

$$\Pr(i) \propto \psi \cdot \frac{1}{t} + (1-\psi) \cdot \frac{k(i)}{mt}$$

degree-uniform mixture



Bias choice by content similarity instead of uniform distribution

Degree-similarity mixture model

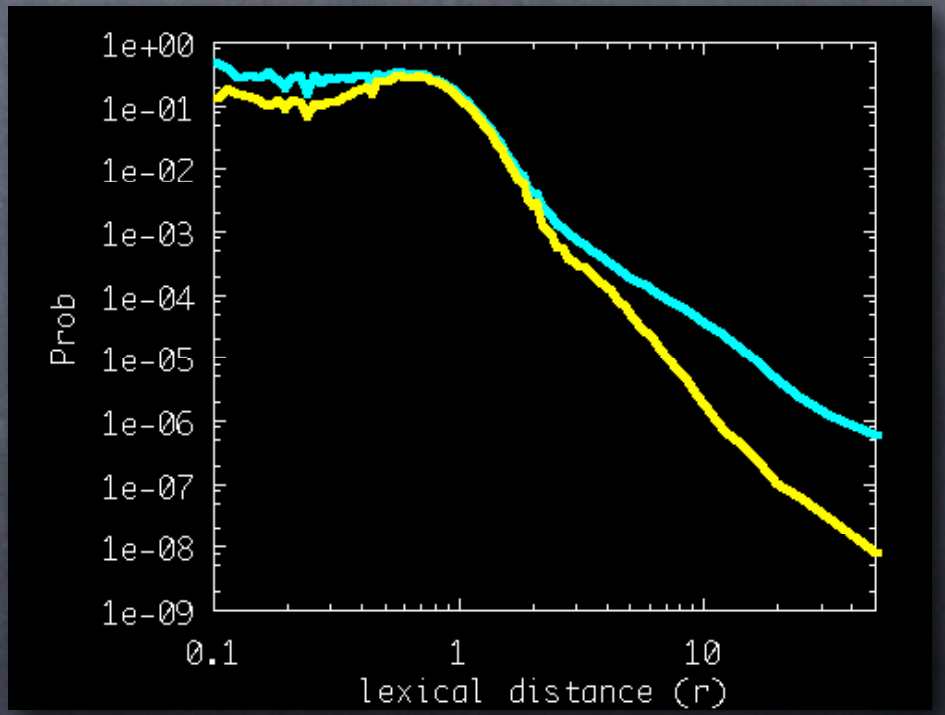
$$\Pr(i) \propto \psi \cdot \hat{\Pr}(i) + (1 - \psi) \cdot \frac{k(i)}{mt}$$

Degree-similarity mixture model

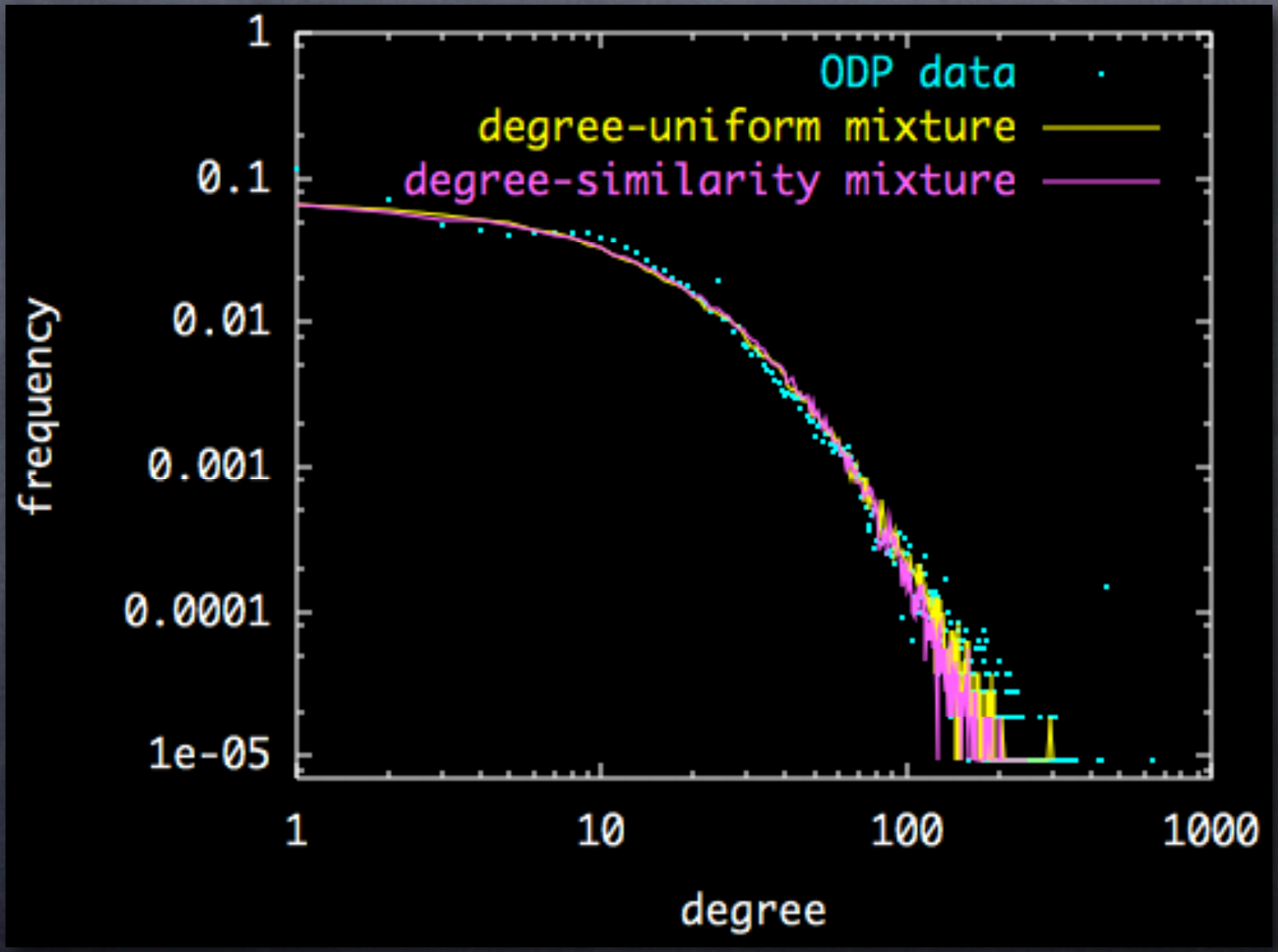
$$\Pr(i) \propto \psi \cdot \hat{\Pr}(i) + (1 - \psi) \cdot \frac{k(i)}{mt}$$

$$\hat{\Pr}(i) \propto [r(i,t)]^{-\alpha}$$

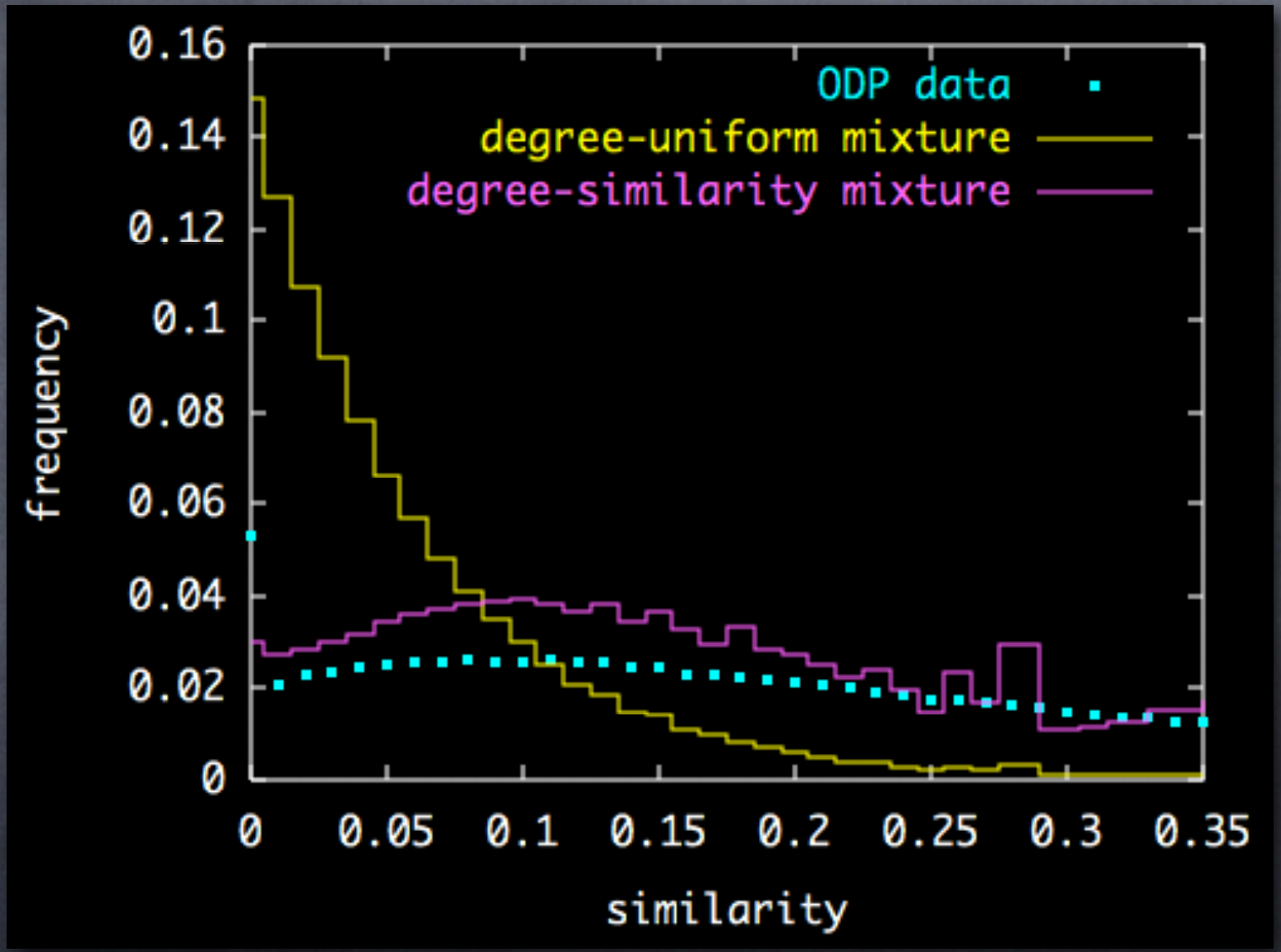
$$\psi = 0.2, \alpha = 1.7$$



Both mixture models get the degree distribution right...



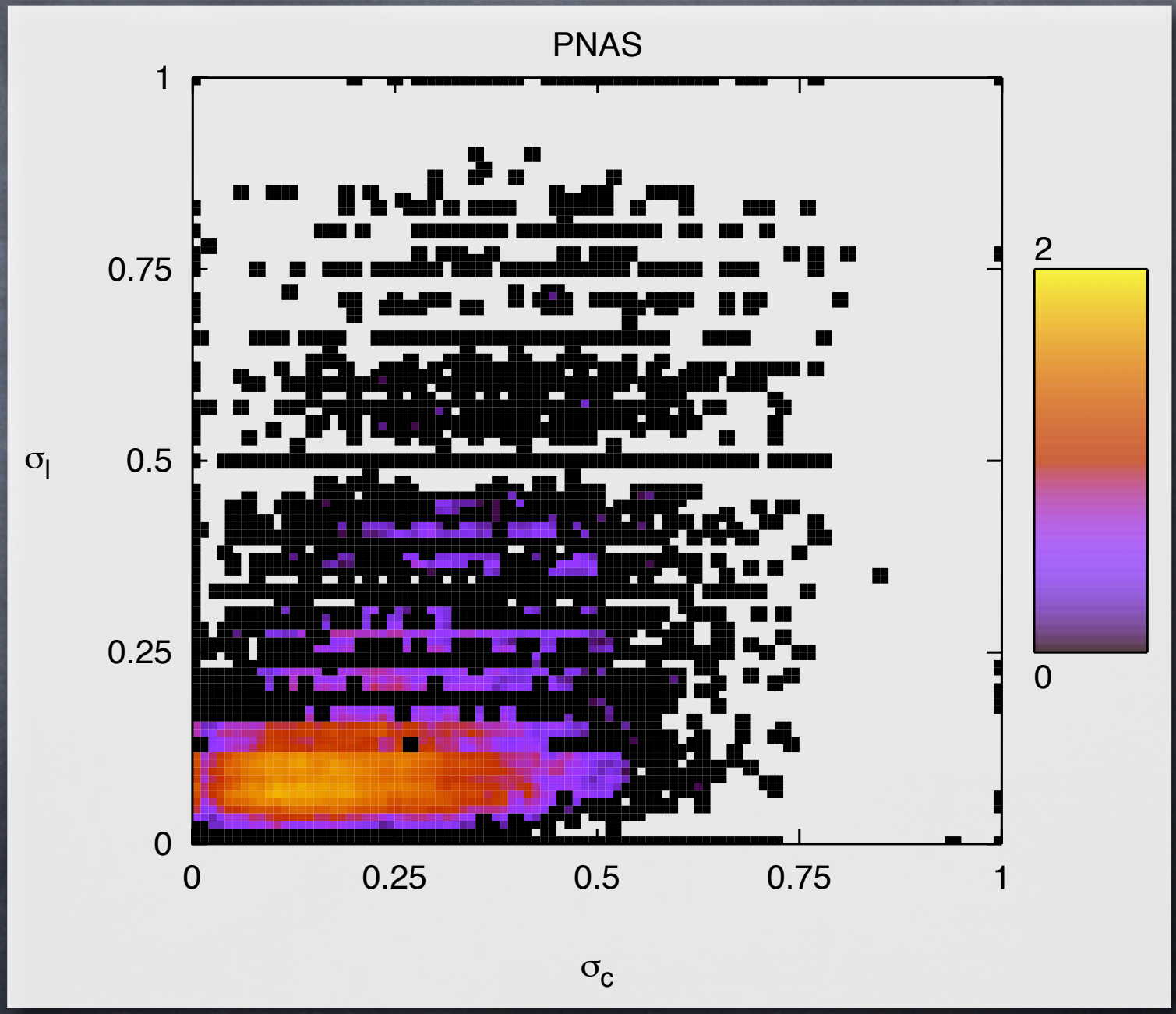
…but the degree-similarity mixture model predicts the similarity distribution better



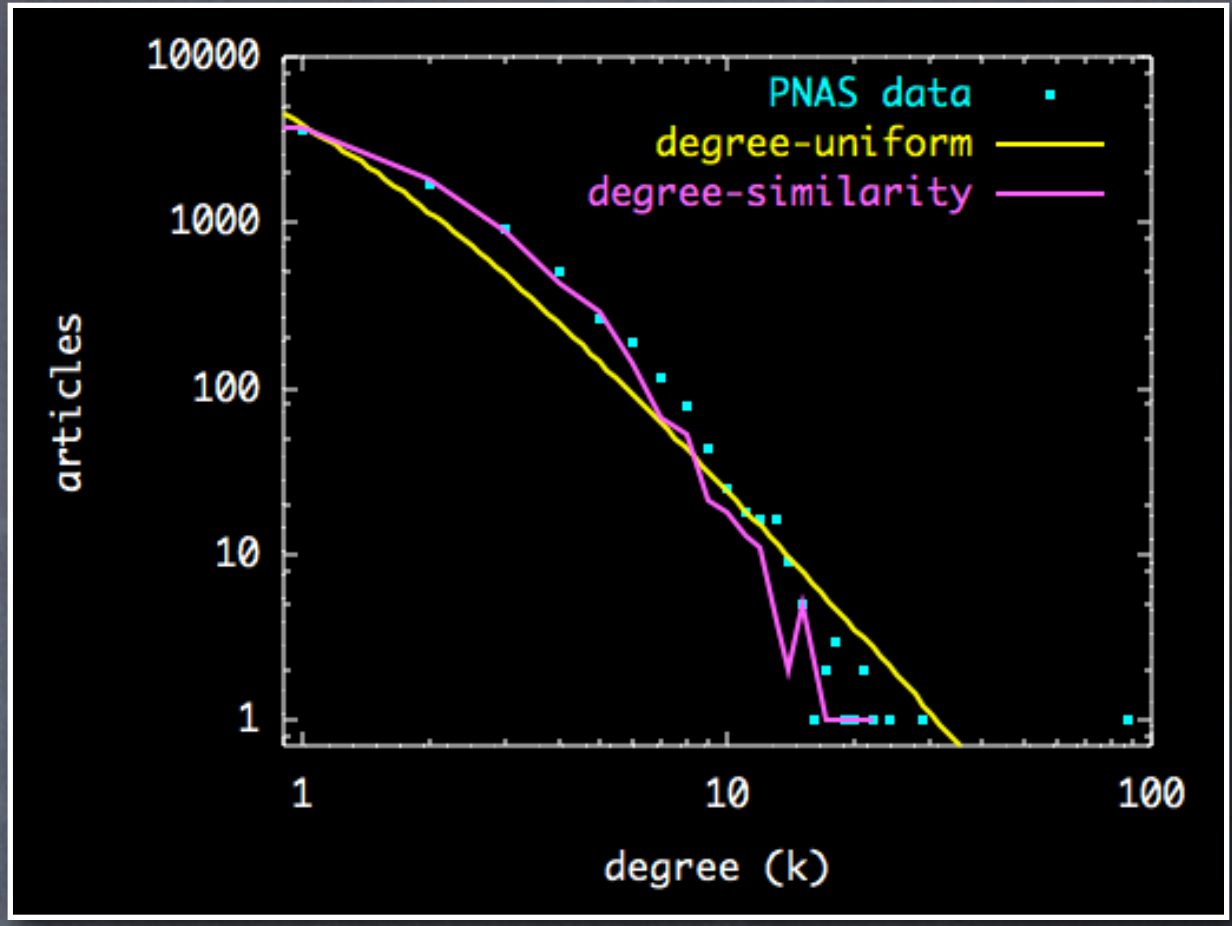
Proc. Natl. Acad. Sci. USA 101: 5261-5265, 2004

Citation networks

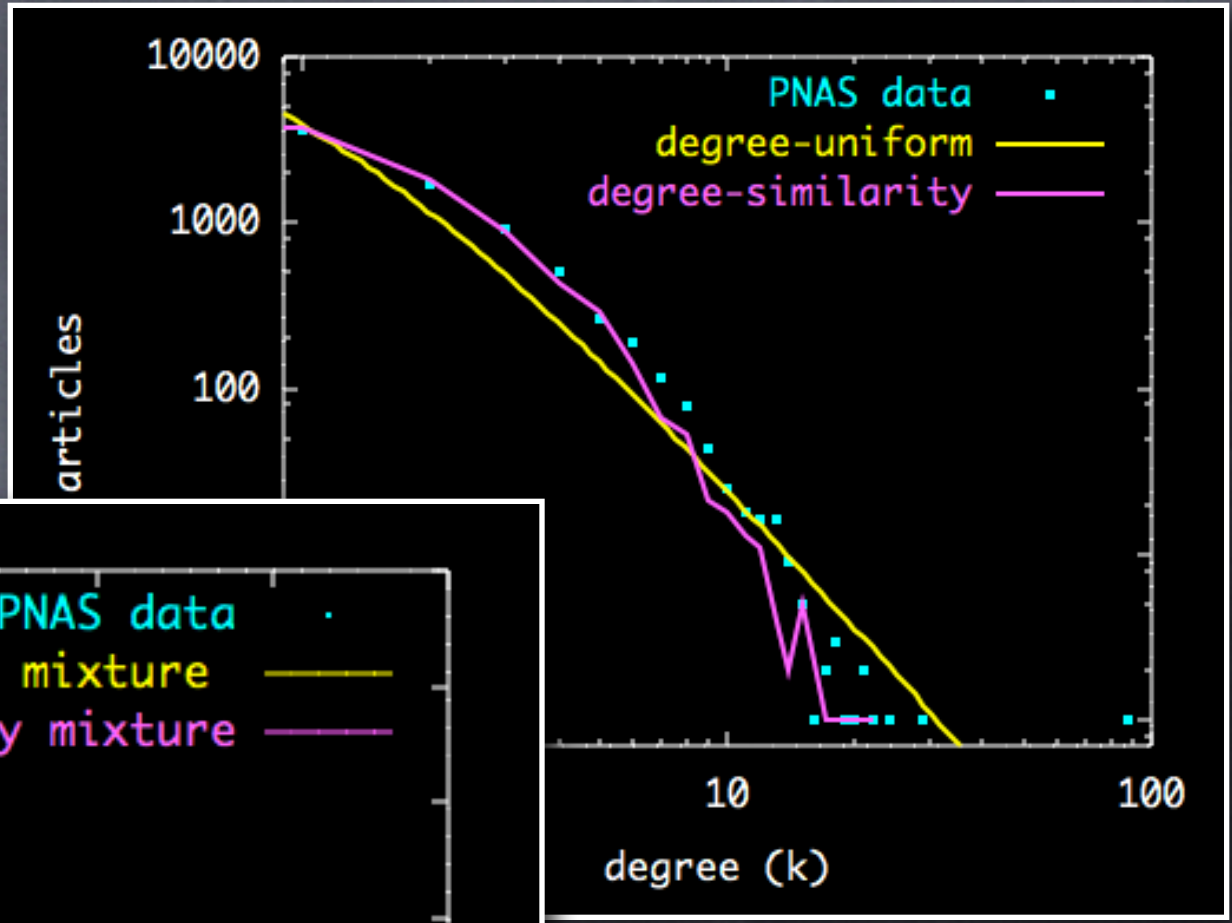
15,785 articles published in PNAS between 1997 and 2002

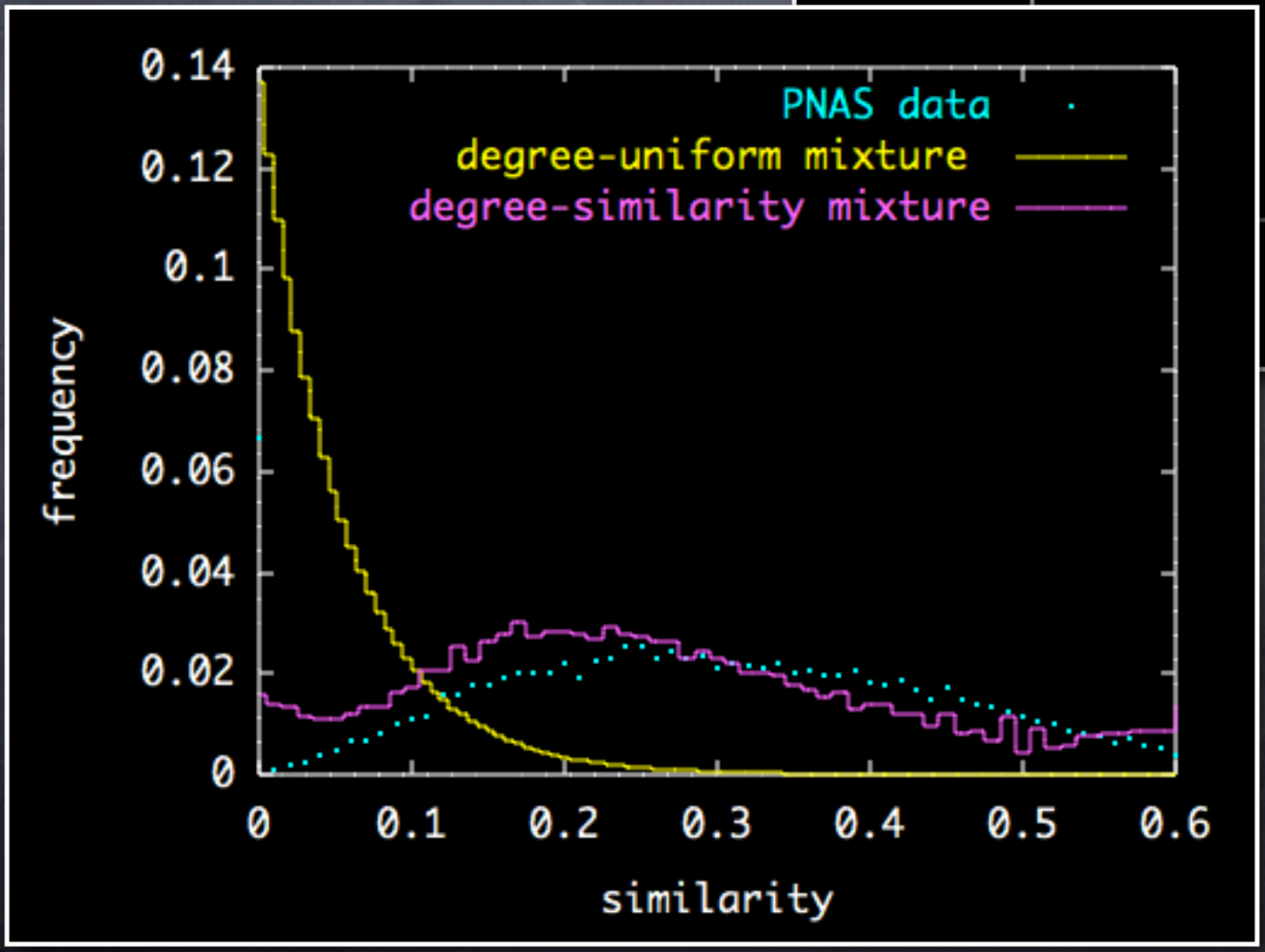


Citation networks



Citation networks





Open Questions

- Understand distribution of content similarity across all pairs of pages
- Growth model to explain co-evolution of both link topology and content similarity
- The role of search engines

Efficient crawling algorithms?

Theory: since the Web is a small world network, or has a scale free degree distribution, short paths exist between any two pages:

- ~ log N (Barabasi & Albert 1999)
- ~ log N / log log N (Bollobas 2001)

Efficient crawling algorithms?

Theory: since the Web is a small world network, or has a scale free degree distribution, short paths exist between any two pages:

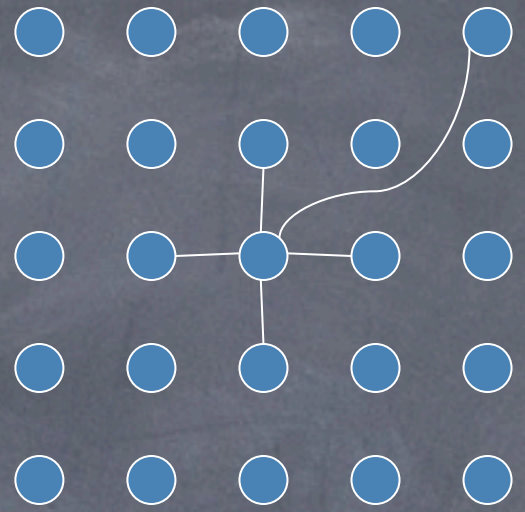
~ log N (Barabasi & Albert 1999)

~ log N / log log N (Bollobas 2001)

Practice: can't find them!

- Greedy algorithms based on location in geographical small world networks: ~ poly(N) (Kleinberg 2000)
- Greedy algorithms based on degree in power law networks: ~ N (Adamic, Huberman & al. 2001)

Exception # 1



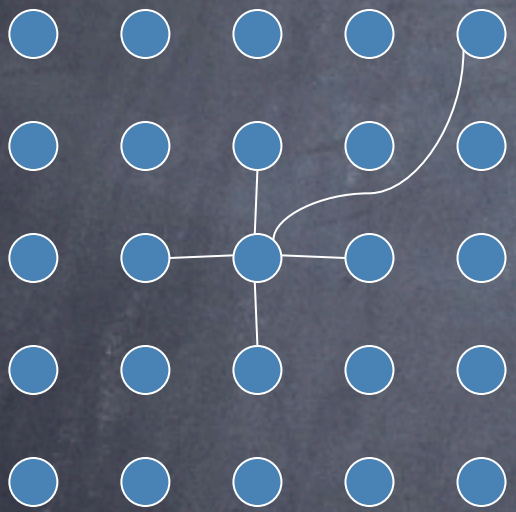
- Geographical networks (Kleinberg 2000)
  - Local links to all lattice neighbors
  - Long-range link probability distribution: power law $Pr \sim r^{-\alpha}$
    - r: lattice (Manhattan) distance
    - α: constant clustering exponent

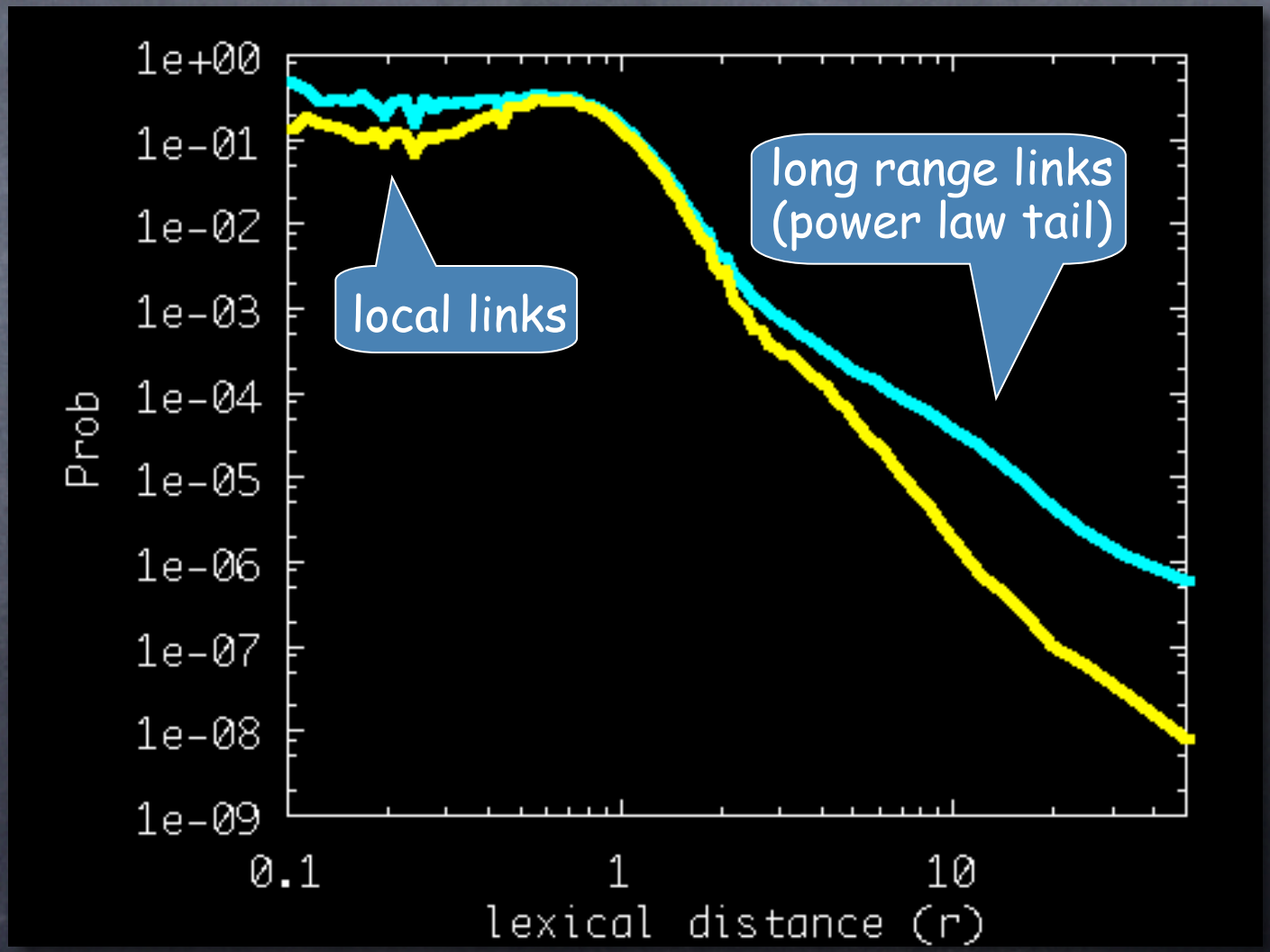
$$t \sim \log^2 N \Leftrightarrow \alpha = D$$

Is the Web a geographical network?

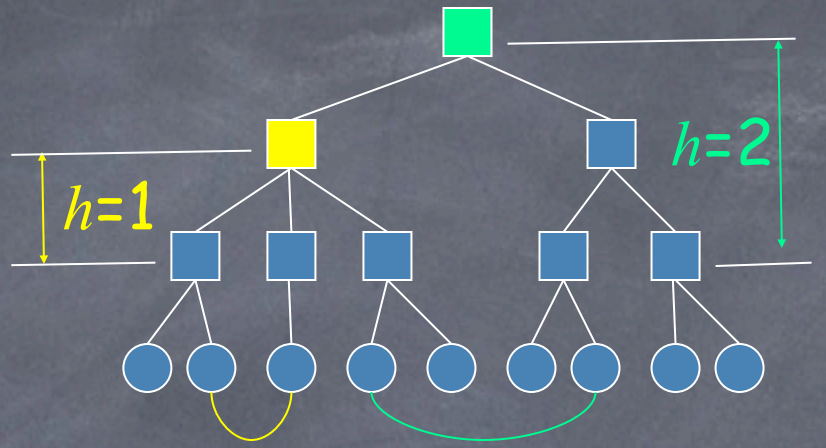
Replace lattice distance by lexical distance

$r = (1 / \sigma_c) - 1$





Exception # 2



- Hierarchical networks (Kleinberg 2002, Watts & al. 2002)
  - Nodes are classified at the leaves of tree
  - Link probability distribution: exponential tail

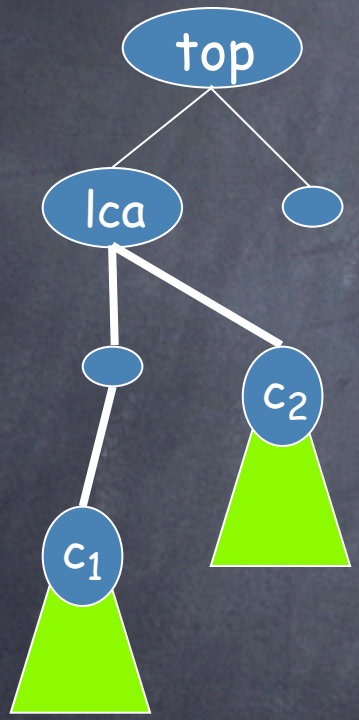
    $Pr \sim e^{-h}$
    - h: tree distance (height of lowest common ancestor)

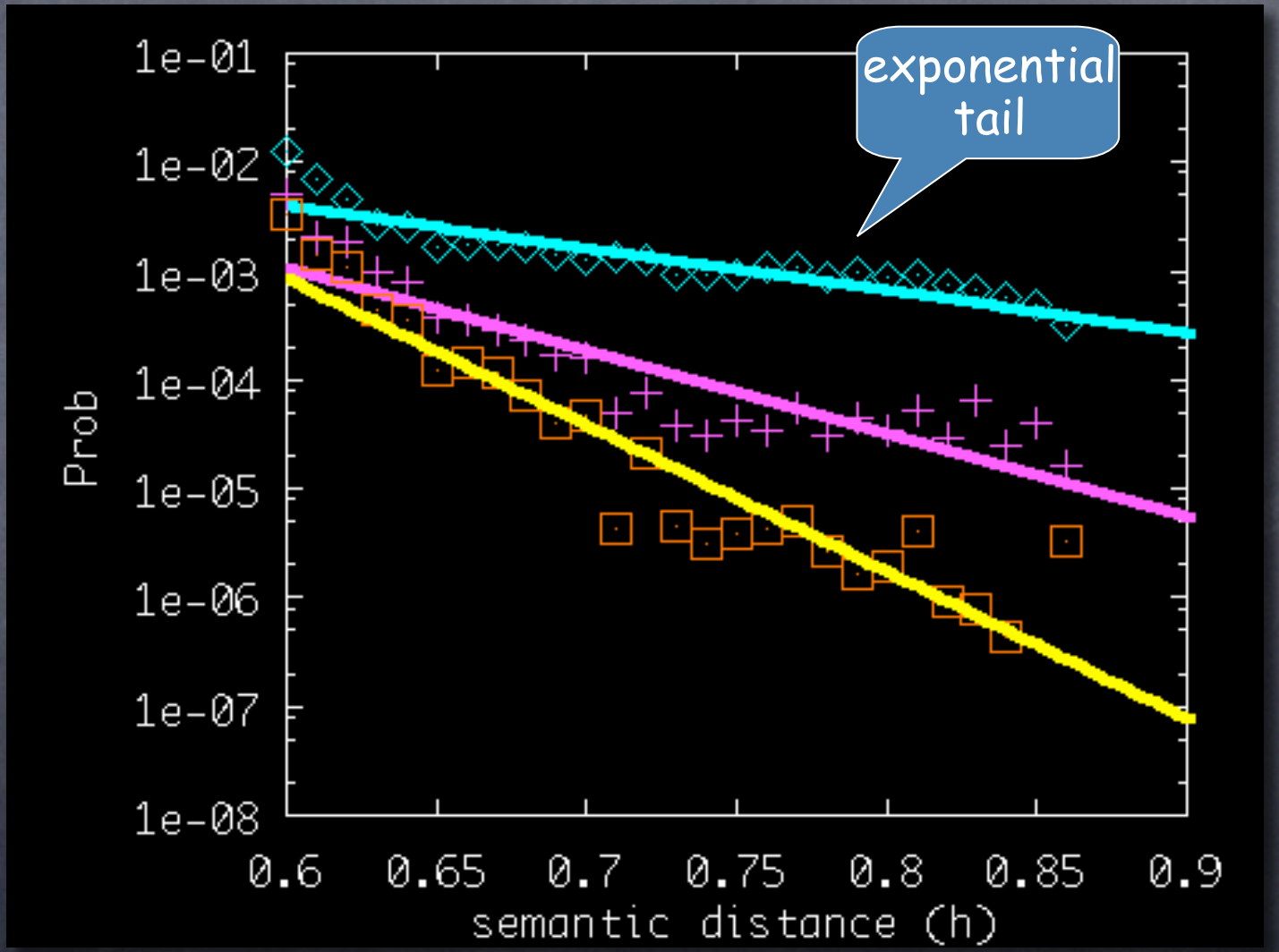
$$t \sim \log^{\varepsilon} N, \varepsilon \geq 1$$

Is the Web a hierarchical network?

Replace tree distance by semantic distance

$h = 1 - \sigma_s$





Take home message:
the Web is a “friendly” place!

Outline

- Topical locality: Content, link, and semantic topologies
- Implications for growth models and navigation
- Applications
  - Topical Web crawlers
  - Distributed collaborative peer search

Crawler applications

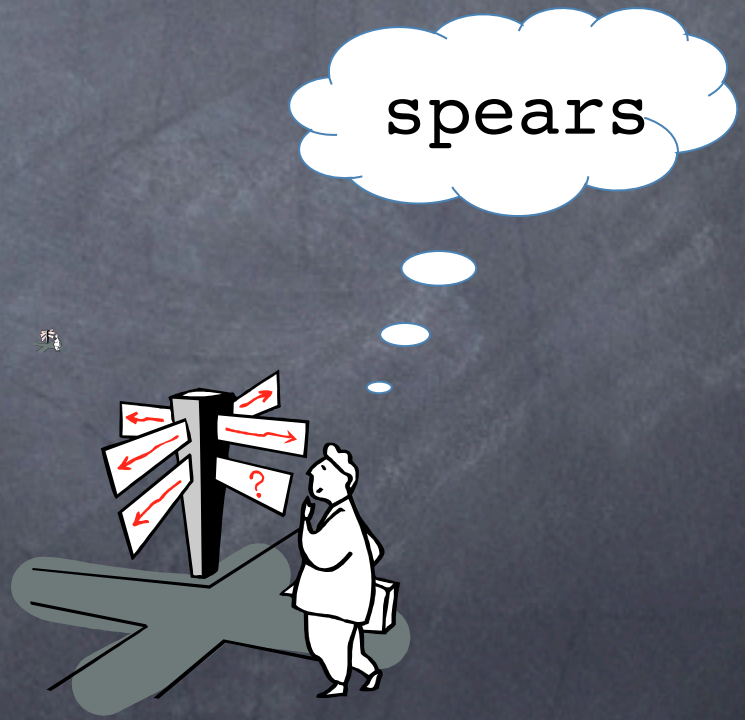
- Universal Crawlers
  - Search engines!
- Topical crawlers
  - Live search
    (e.g., myspiders.informatics.indiana.edu)
  - Topical search engines & portals
  - Business intelligence (find competitors/partners)
  - Distributed, collaborative search

Topical crawlers

Miselanous Physics Websites
[sic]

- Coaxial Cable Attenuation & Power Handling Calculator
- Britney Spears guide to Semiconductor Physics: semiconductor physics, Edge Emitting Lasers and VCSELs
- Particle Data Book everything you ever wanted to know about particles, and then some.
- X-Ray Data Booklet everything you ever wanted to know about x-rays, and then some.

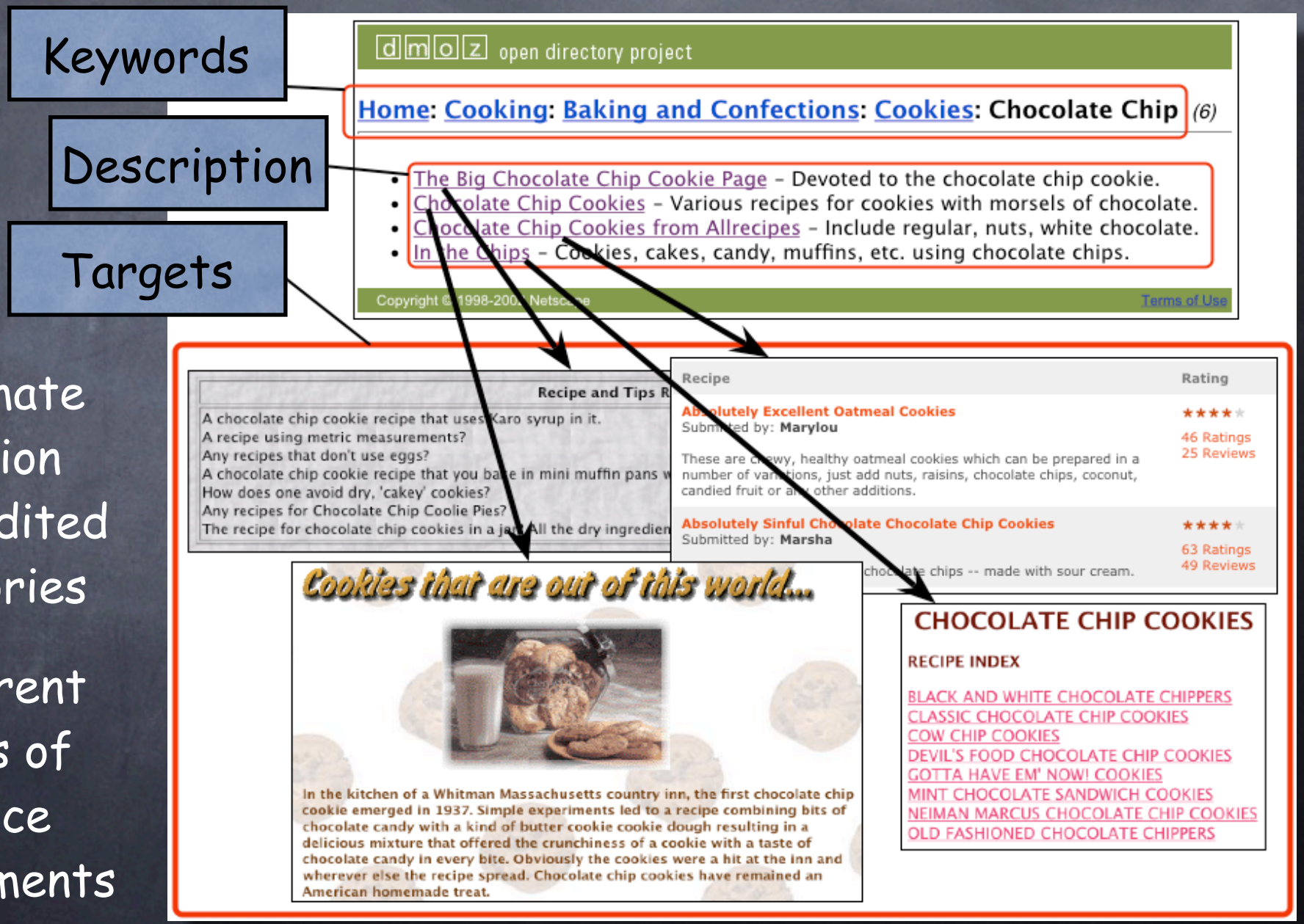
'A Brief History of Anglo-Saxon England' -
'Anglo-Saxon Military Organisation' - Article.
'Anglo-Saxon Social Organisation' - Article.
'Arms and Armour - Part 3 Axes' - Article.
'Arms and Armour - Part 7 Helmets' - Article.
'Arms and Armour - Part 6 Mail Armour' - Article.
'Arms and Armour - Part 4 Missile Weapons' - Article.
'Arms and Armour - Part 2 Scramseaxes' - Article.
'Arms and Armour - Part 8 Shields' - Article.
'Arms and Armour - Part 1 Spears' - Article.
'Arms and Armour - Part 5 Swords' - Article.
'A Nice Little Earner' - The slave trade in Anglo-Saxon England.
'A Spring Warmer' - An alternative recipe for jugged hare!
'The Battle of Hastings' - Article.
'Bone and Antler Working' - Article.
'Braid Weaving' - Article.
'Bronzeworking' - Article.
'Charcoal Burning' - The results of an experiment in charcoal burning.
'Church Organisation' - The organisation of the church in Anglo-Saxon England.



Evaluating topical crawlers

- Goal: build "better" crawlers to support applications
- Build an unbiased evaluation framework
  - Define common tasks of measurable difficulty
  - Identify topics, relevant targets
  - Identify appropriate performance measures
    - Effectiveness: quality of crawler pages, order, etc.
    - Efficiency: separate CPU & memory of crawler algorithms from bandwidth & common utilities

Evaluating topical crawlers: Topics

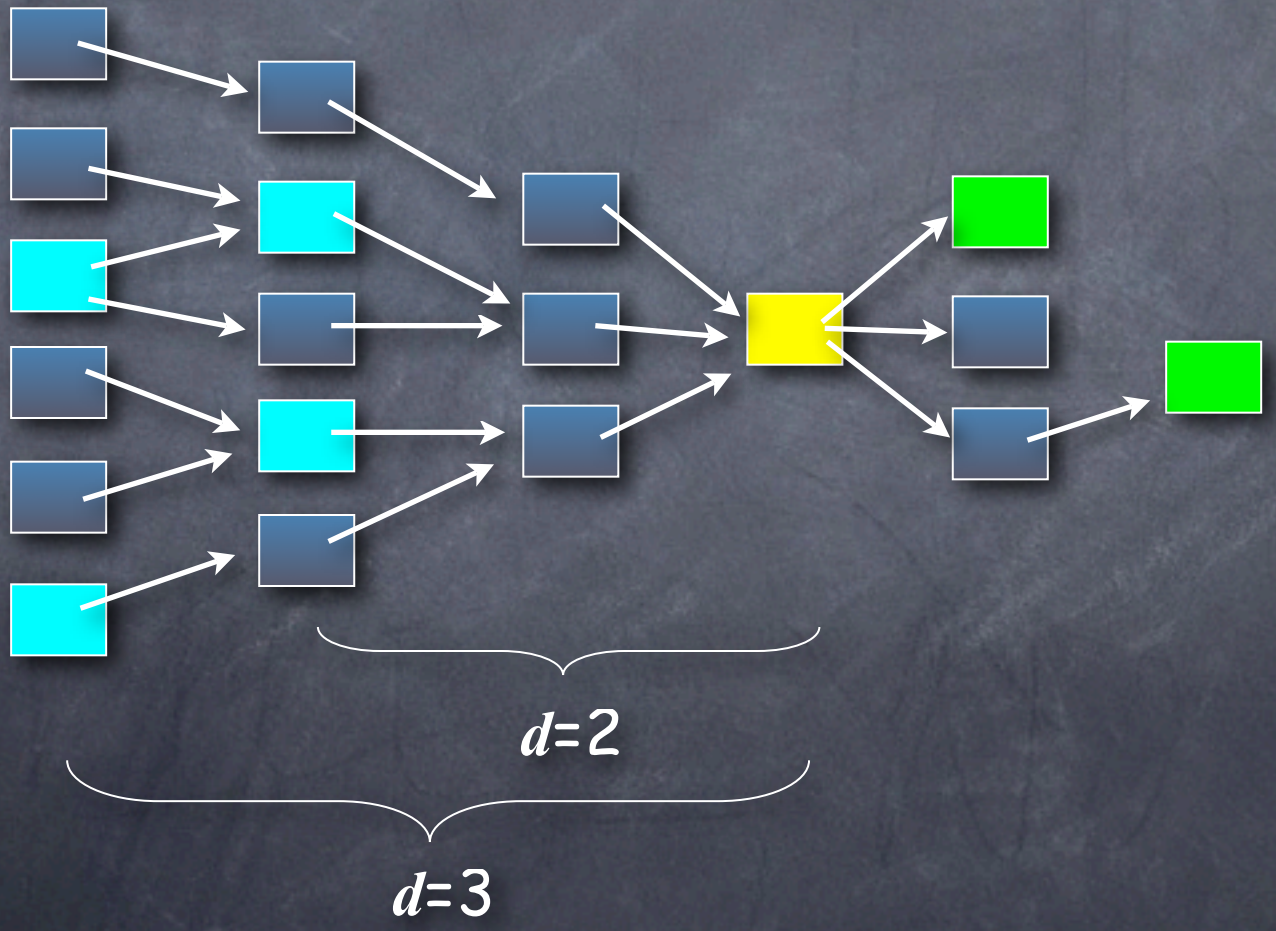


- Automate evaluation using edited directories
- Different sources of relevance assessments

Evaluating topical crawlers: Tasks

Start from seeds, find targets

and/or pages similar to target descriptions



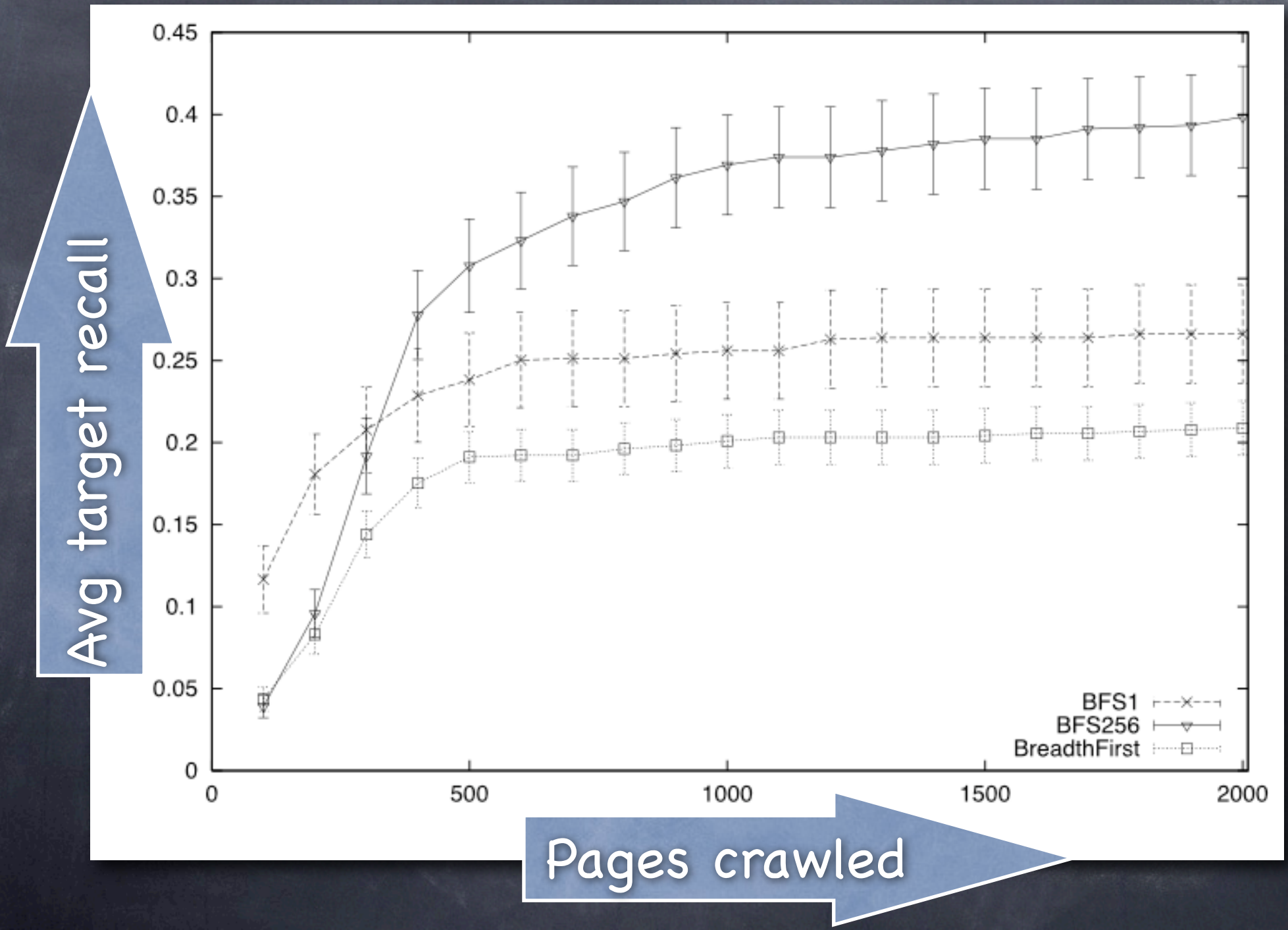
Examples of crawling algorithms

- Breadth-First
  - Visit links in order encountered
- Best-First
  - Priority queue sorted by similarity
  - Variants:
    - explore top N at a time
    - tag tree context
    - hub scores
- SharkSearch
  - Priority queue sorted by combination of similarity, anchor text, similarity of parent, etc.
- InfoSpiders

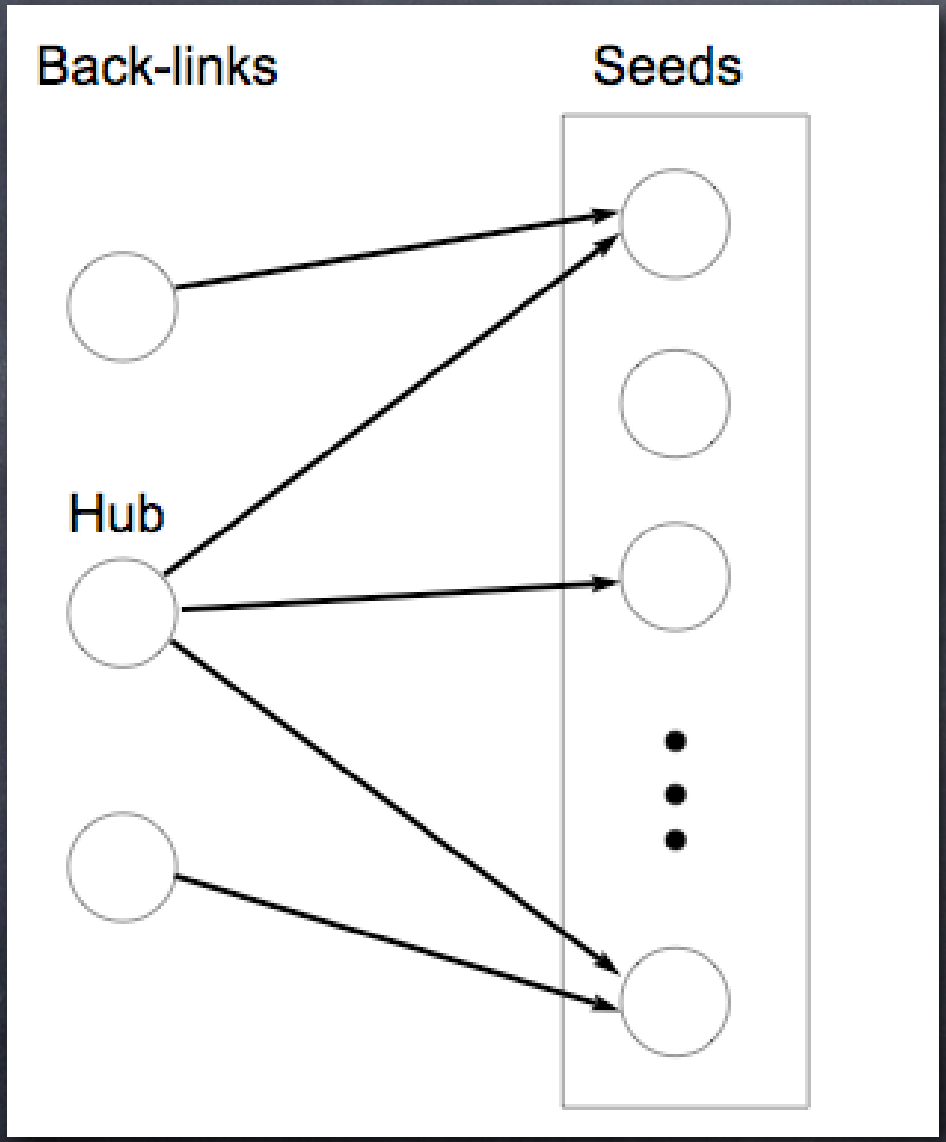
Examples of crawling algorithms

- Breadth-First
  - Visit links in order encountered
- Best-First
  - Priority queue sorted by similarity
  - Variants:
    - explore top N at a time
    - tag tree context
    - hub scores
- SharkSearch
  - Priority queue sorted by combination of similarity, anchor text, similarity of parent, etc.
- InfoSpiders

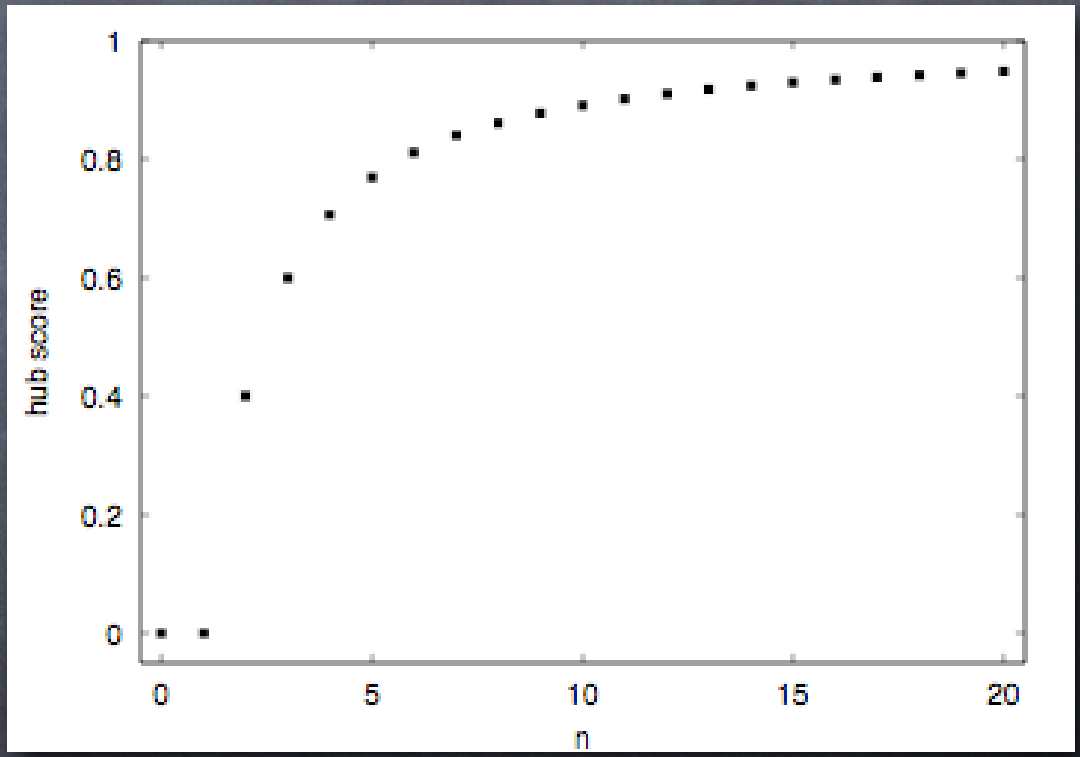
Exploration vs. Exploitation



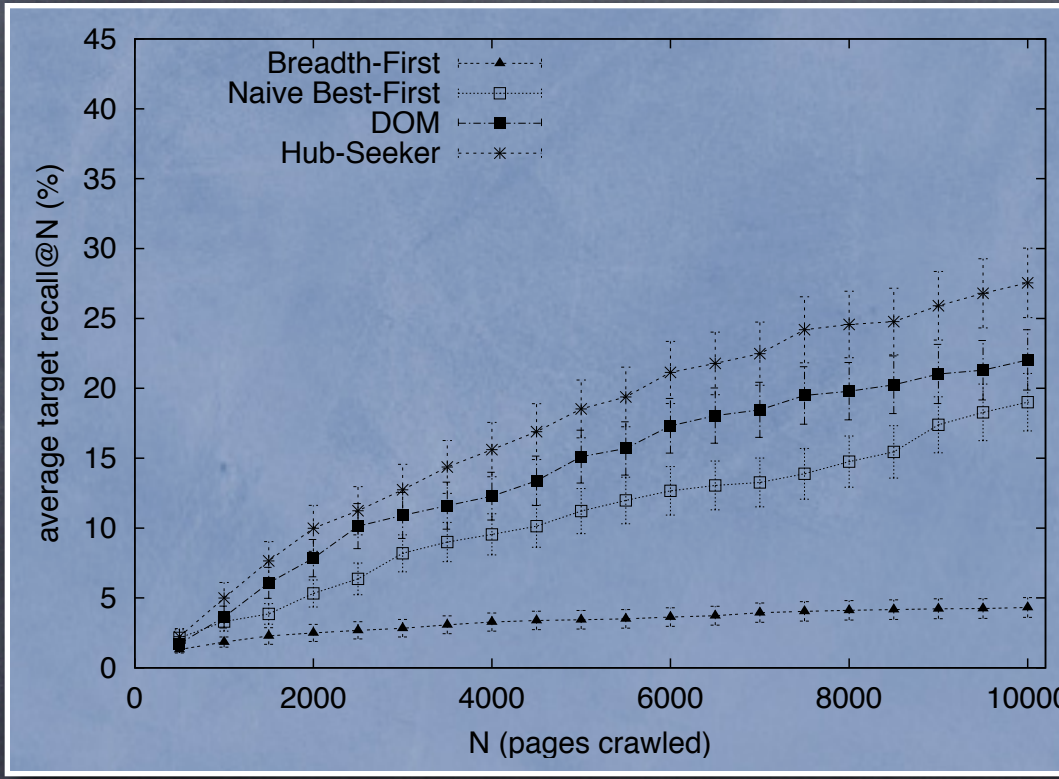
Co-citation: hub scores

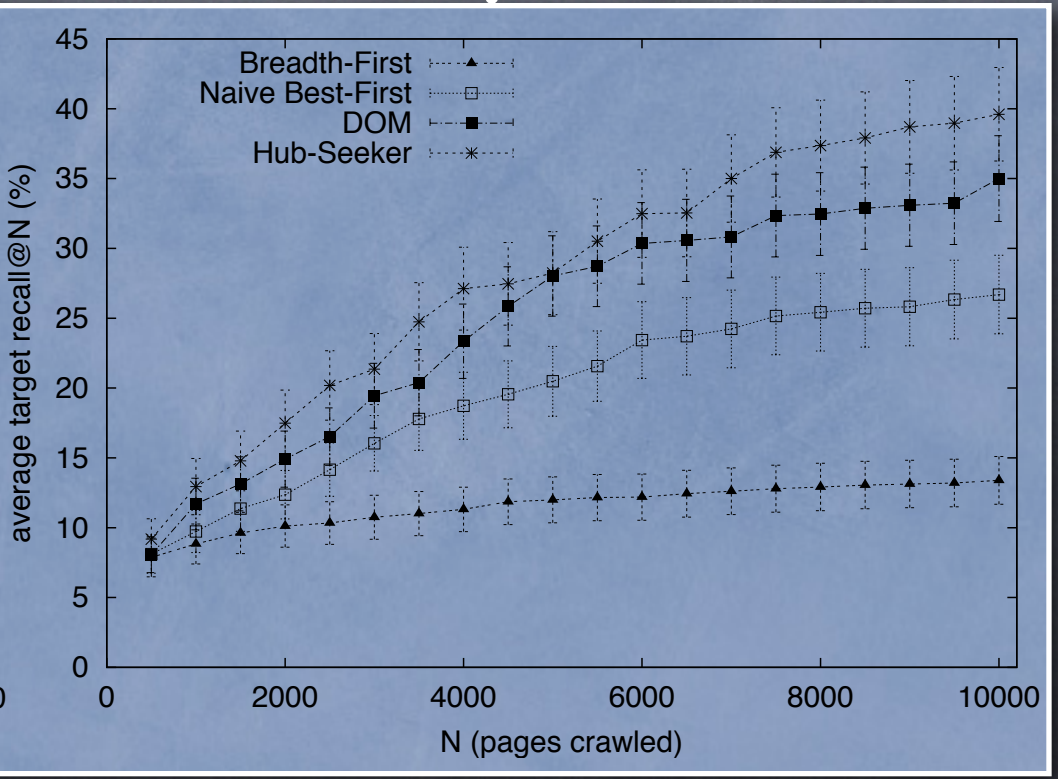


Link score$_{hub}$ = linear combination between link and hub score



Recall (159 ODP topics)





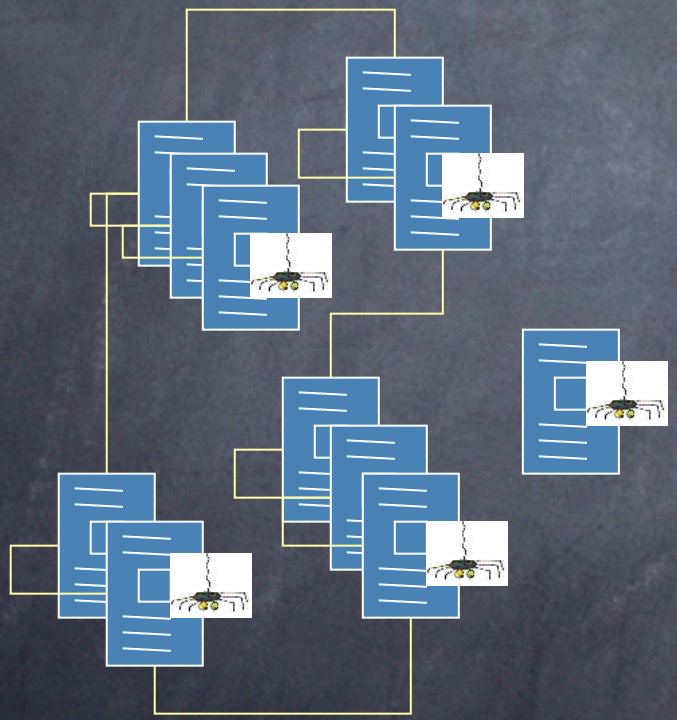
Split ODP URLs between seeds and targets

Add 10 best hubs to seeds for 94 topics

ECDL 2003

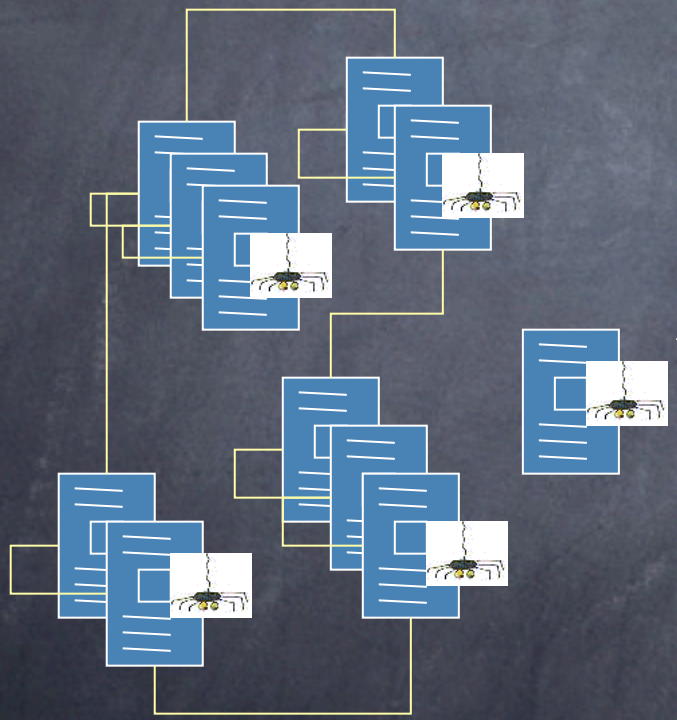
InfoSpiders

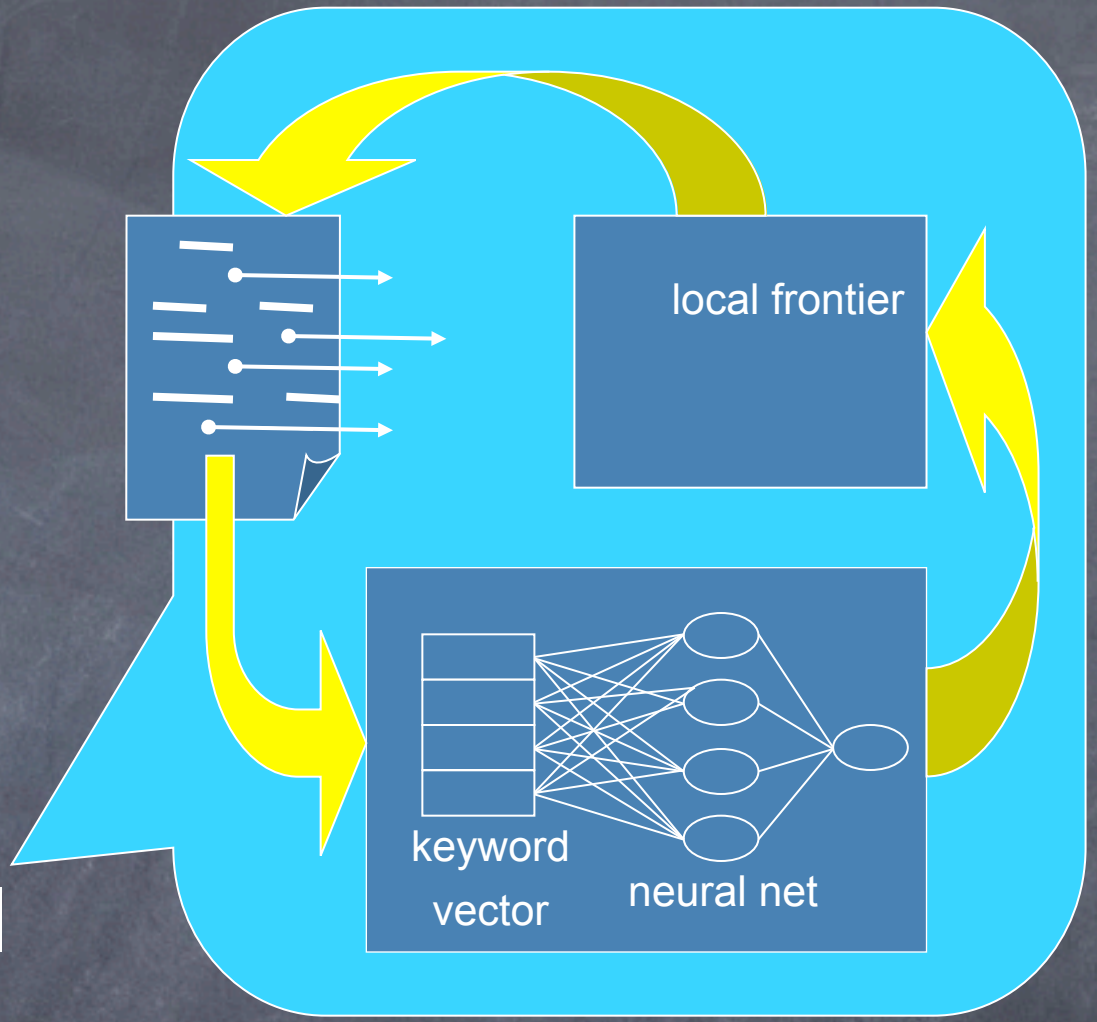
adaptive distributed algorithm using an evolving population of learning agents



InfoSpiders

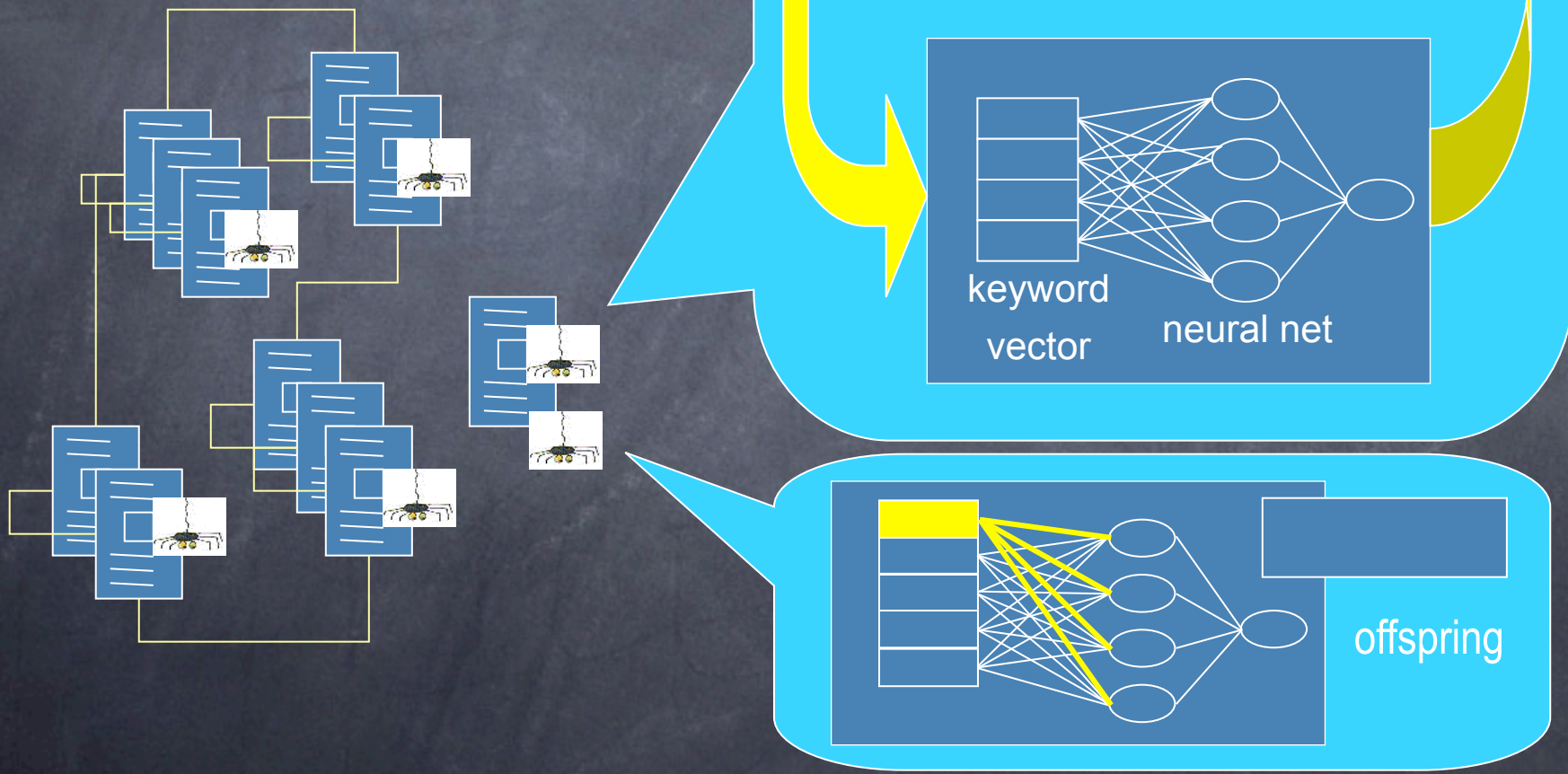
adaptive distributed algorithm using an evolving population of learning agents





InfoSpiders

adaptive distributed algorithm using an evolving population of learning agents



InfoSpiders

Author.....FILIPPO MENCZER

Advisor.....RIK BELEW

University of California, San Diego

Easy query:
The population rapidly focuses
on the relevant areas of the
information space

InfoSpiders

Author.....FILIPPO MENCZER

Advisor.....RIK BELEW

University of California, San Diego

Ambiguous query:
A subpopulation eventually
locates the relevant pages

Mozilla Firefox

http://myspiders.informatics.indiana.edu/myspiders2.html

Query: search censorship in france

Start Stop

Max. Pages: 100

MySpiders

Crawler Name

InfoSpiders

Pages Crawled: 100

Population: 0

Source	URL	Score	Rece...
Spider2	http://www.multilingual-search.com/new-tool-shows-cens...	0.43	?
Seed	http://www.technologynewsdaily.com/node/2283	0.34	1
Seed	http://www.rpi.edu/~bulloj/search/CENSORSHIP.html	0.33	0
Seed	http://en.wikipedia.org/wiki/Censorship_in_France	0.32	0.14
Seed	http://www.laboratorytalk.com/news/iqd/iqd109.html	0.3	0.33
Seed	http://www.informatics.indiana.edu/news/news.asp?id=313	0.27	?
Seed	http://blog.searchenginewatch.com/blog/050117-090638	0.12	0.02

- Spider Hierarchy
  - Spiders
    - Spider1
    - Spider2
      - Spider13
    - Spider3
    - Spider4
    - Spider5
      - Spider11
      - Spider12
    - Spider6
    - Spider7

New tool shows censorship by search engine in China, France, Germany and the US [Multilingual Search]

New tool shows censorship by search engine in China, France, Germany and the US

Andy Atkins-Krüger Mar 19, 2006 | [en]

Pandia reports on a new censorship comparison tool developed by researchers at Indiana University. The tool compares the preeminance of words featured in the top ten results of Yahoo or Google by displaying words graphically giving weight to those terms which are more frequent - the end result is a graphic somewhat reminscent of Technorati tags.

CENSEARCHIP

Called Censearchip - the team behind the tool have clearly chosen China, France and Germany - compared with the US - because of the recent censorhip issues in China and the restrictions placed on search engines in terms of displaying nazi material - by Germany and France.

In addition to this 'political' censorhip - it would be useful to be able to compare results by country where 'busines' censorhip has an impact - for instance the filtering of results which takes place between the US and the UK.

Spider Details

- Details
  - Spider13
    - Status
    - Energy
      - 1.1394327
    - Query
      - Term1
      - Term2
      - Term3
      - Term4
        - itali
    - History

Spider Details

- Details
  - Spider11
    - Status
    - Energy
    - Query
      - Term1
      - Term2
      - Term3
      - Term4
        - engin
    - History

Evolutionary Local Selection Algorithm (ELSA)

```
Foreach agent thread:
 Pick & follow link from local frontier
 Evaluate new links, merge frontier
 Adjust link estimator
 E := E + payoff - cost
     If E < 0:
        Die
     Elsif E > Selection_Threshold:
        Clone offspring
        Split energy with offspring
        Split frontier with offspring
        Mutate offspring
```

reinforcement learning

match resource bias

selective query expansion

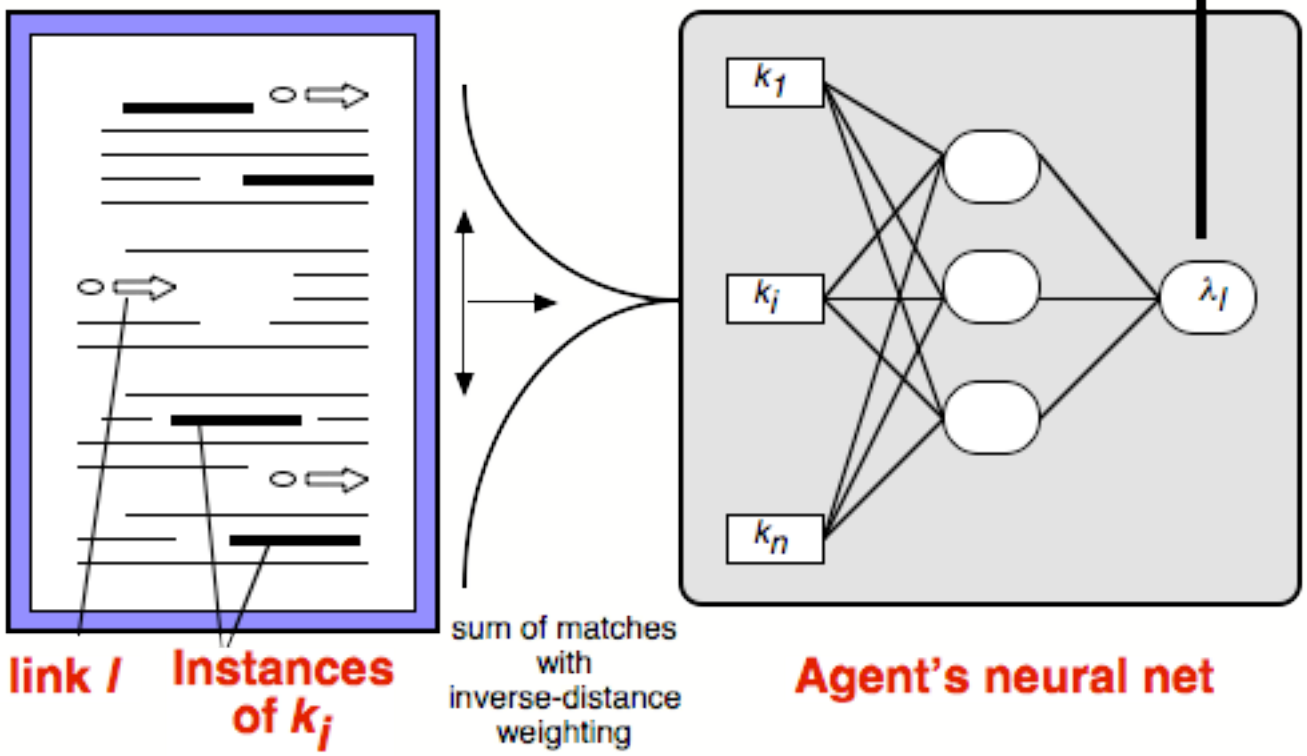
Action selection

$$\Pr[l] = \frac{e^{\beta\lambda_l}}{\sum_{l'} e^{\beta\lambda_{l'}}}$$

Stochastic selector

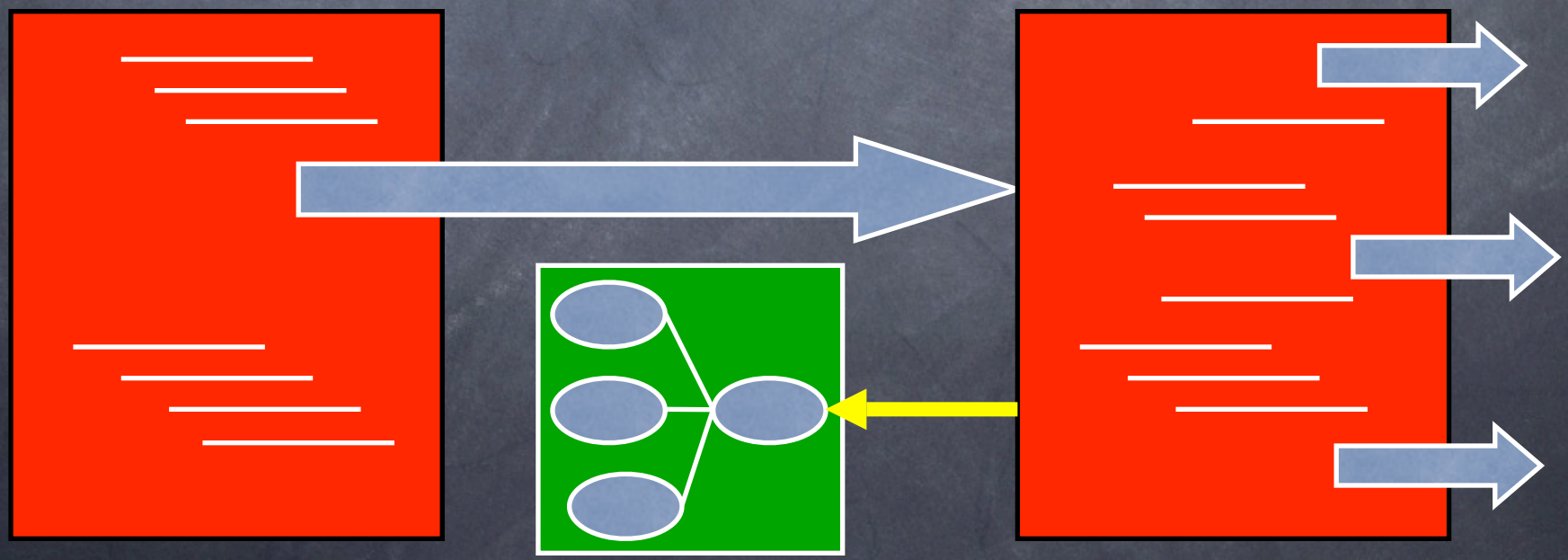
$$\lambda_l = nnet(in_1, \ldots, in_N)$$

$$in_k = \sum_{w \in D} \frac{\delta(k,w)}{dist(w,l)}$$

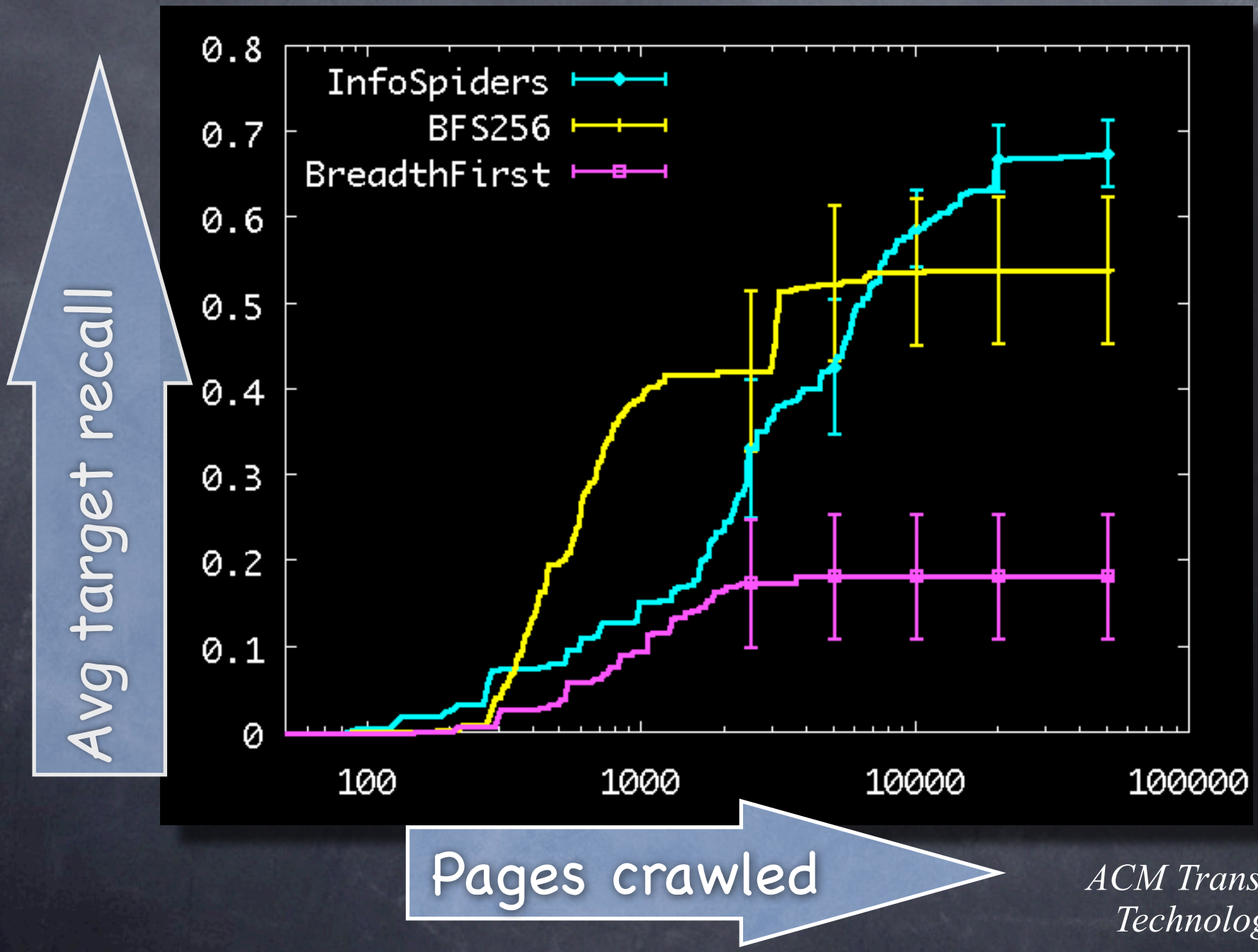


Q-learning

- Compare estimated relevance of visited document with estimated relevance of link followed from previous page
- Teaching input: $E(D) + \mu \max_{l(D)} \lambda_l$



Performance



ACM Trans. Internet Technology 2003

Google paris hilton Search Advanced Search Preferences

Web News Images Video Results 1 - 10 of about 27,100,000 for paris hilton. (0.04 seconds)

News results for paris hilton


ITN

Paris Hilton 'Loves Everything About Korea' - 21 hours ago
By Cathy Rose A. Garcia A barrage of flashing cameras greeted hotel heiress and American celebrity Paris Hilton when she attended her first press conference ...
Korea Times - 18 related articles »
Hallmark Files Motion to Dismiss Paris Hilton's - FOX News - 12 related articles »

Paris Hilton - Wikipedia, the free encyclopedia
Paris Hilton's best-known acting work is in the television series The Simple Life with her friend Nicole Richie. She has also appeared in several minor film ...
en.wikipedia.org/wiki/Paris_Hilton - 120k - Cached - Similar pages

Paris Hilton | The Official Website
ParisHilton.com Paris Hilton, Nicky Hilton Fashion, Pictures, Apparel, Jewellery, Film, and Fun.
www.parishilton.com/ - 8k - Cached - Similar pages

Paris Hilton
Paris Hilton on IMDb: Movies, TV, Celebs, and more...
www.imdb.com/name/nm0385296/ - 51k - Cached - Similar pages

1 Night in Paris (2004) (V)
The exclusive Paris Hilton sex video, where she and Rick Salomon's private ... Oh wow she's Paris Hilton so her sex tape is automatically controversial and ...
www.imdb.com/title/tt0412260/ - 34k - Cached - Similar pages

Paris Hilton Zone | Paris Hilton Pictures, Pics, Photos
4000+ new Paris Hilton pictures, Paris wallpaper, sex tape, lyrics, audio, video, daily Paris pics & news.
www.parishiltonzone.com/ - 47k - Cached - Similar pages

AskMen.com - Paris Hilton
AskMen.com feature on Paris Hilton that includes pics, pictures, biography, video, related news, vital stats, commentary, and cool facts. ...
www.askmen.com/women/models_150/171_paris_hilton.html - Similar pages

Google

Web News Images V

News results for paris

Paris Hilto
By Cathy R
American c
Korea Time
Hallmark Fi
ITN

Paris Hilton - Wikiped
Paris Hilton's best-known
friend Nicole Richie. She ha
en.wikipedia.org/wiki/Paris_

Paris Hilton | The Offi
ParisHilton.com Paris Hilto
Fun.
www.parishilton.com/ - 8k

Paris Hilton
Paris Hilton on IMDb: Mov
www.imdb.com/name/nm03

1 Night in Paris
The exclusive Paris
she's Paris Hilton s
www.imdb.com/title/

Paris Hilton Zone | Pa
4000+ new Paris Hilton pic
pics & news.
www.parishiltonzone.com/

AskMen.com - Paris H
AskMen.com feature on Pa
news, vital stats, comment
www.askmen.com/women/

tripadvisor®
get the truth. then go.™

Search results

16,863,263 Travelers Planned Trips This Week! Sign in Register! My Trips

Search Again

paris hilton GO

(e.g., Boston hotels, Las Vegas, Paris art museum)

Refine Search

- Hotel (5)
- Reviews (1,111)
- Forums (2,117)
- Travel Guide (10)
- goLists™ (6)

advertisement

You live in more than one place, so AT&T works in more places.

Name your place. Customize your design. Get your gear.

START HERE

Paris Hilton Travel Deals

Sponsored links *

Compare all deals on Paris Hilton

Paris Hilton: Special autumn and winter rates! SmartDeal
Hilton.com Our Best Rates. Guaranteed. Book now!

Paris Hilton: Great rooms from $282 SmartDeal
Expedia.com Get the lowest price on hotels with our Best Price Guarantee

Paris Hilton: Great Rooms, Great Prices SmartDeal
hotels.com Low Rates Guaranteed! Book your Trip with a Hotel Expert at 1-800-434-6835.

Location Results 1 - 5 of 5 for *paris hilton*

Paris Hilton, Paris, Ile-de-France, France
Reviews of Paris Hilton

Hilton Arc de Triomphe Paris, Paris, Ile-de-France, France
Reviews of Hilton Arc de Triomphe Paris

Hilton Paris Charles de Gaulle Airport, Roissy, Ile-de-France, France
Reviews of Hilton Paris Charles de Gaulle Airport

Hilton Paris Orly Airport, Paris, Ile-de-France, France
Reviews of Hilton Paris Orly Airport

Hilton Paris La Defense, Paris, Ile-de-France, France
Reviews of Hilton Paris La Defense

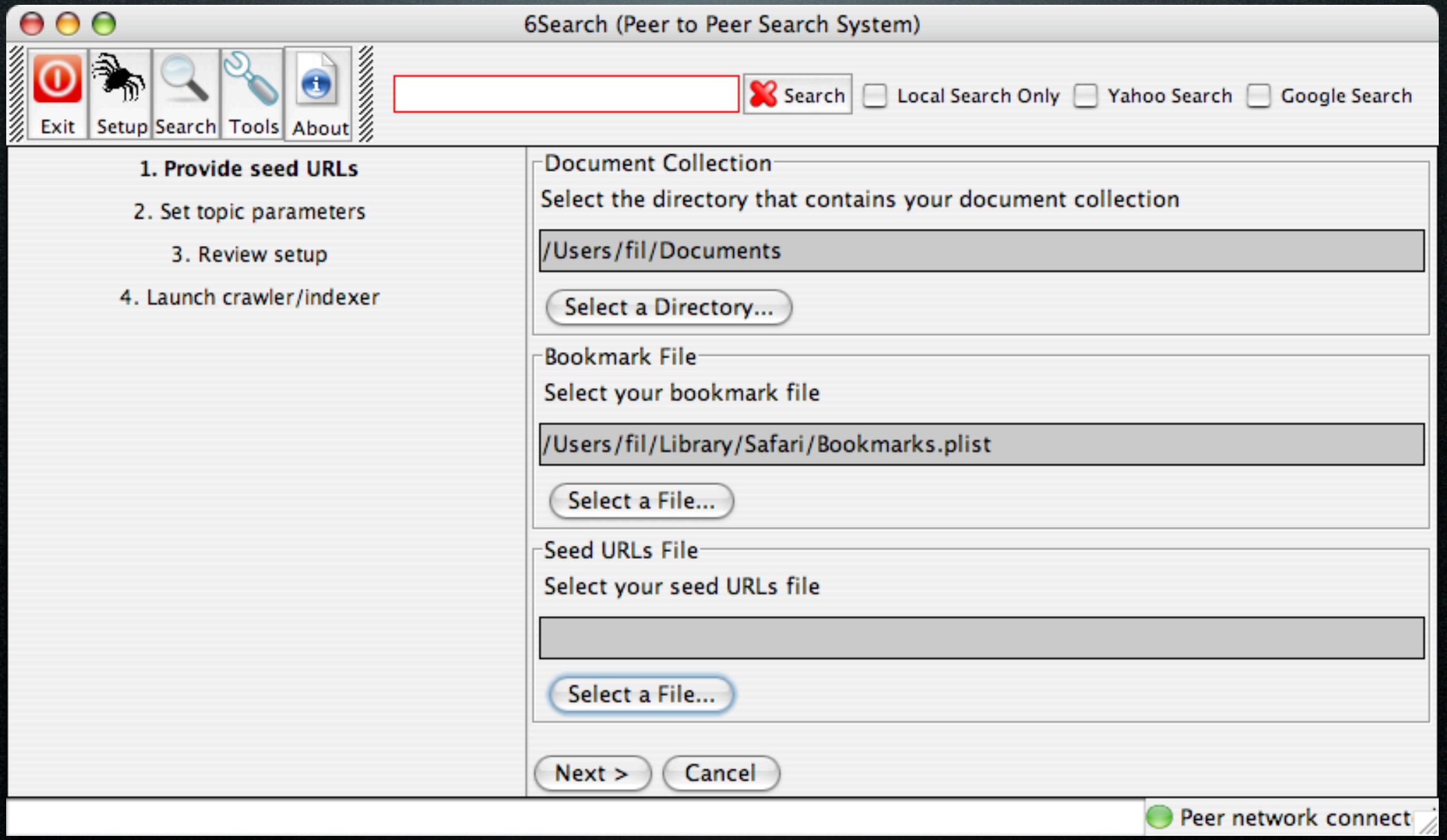
What Our Users Are Saying Results 1 - 10 of 3,244 for *paris hilton*

Review: Paris Hilton, Paris, Île-de-France, France
1 2 3 4 5 4.0 Paris Hilton
July 01, 2006 *A TripAdvisor Member*, Southern, CA
So who says **Paris** isn't a great ... ... we drank more wine in **Paris** than we have in the last 2 years in the US. Take a very short walk to the ...

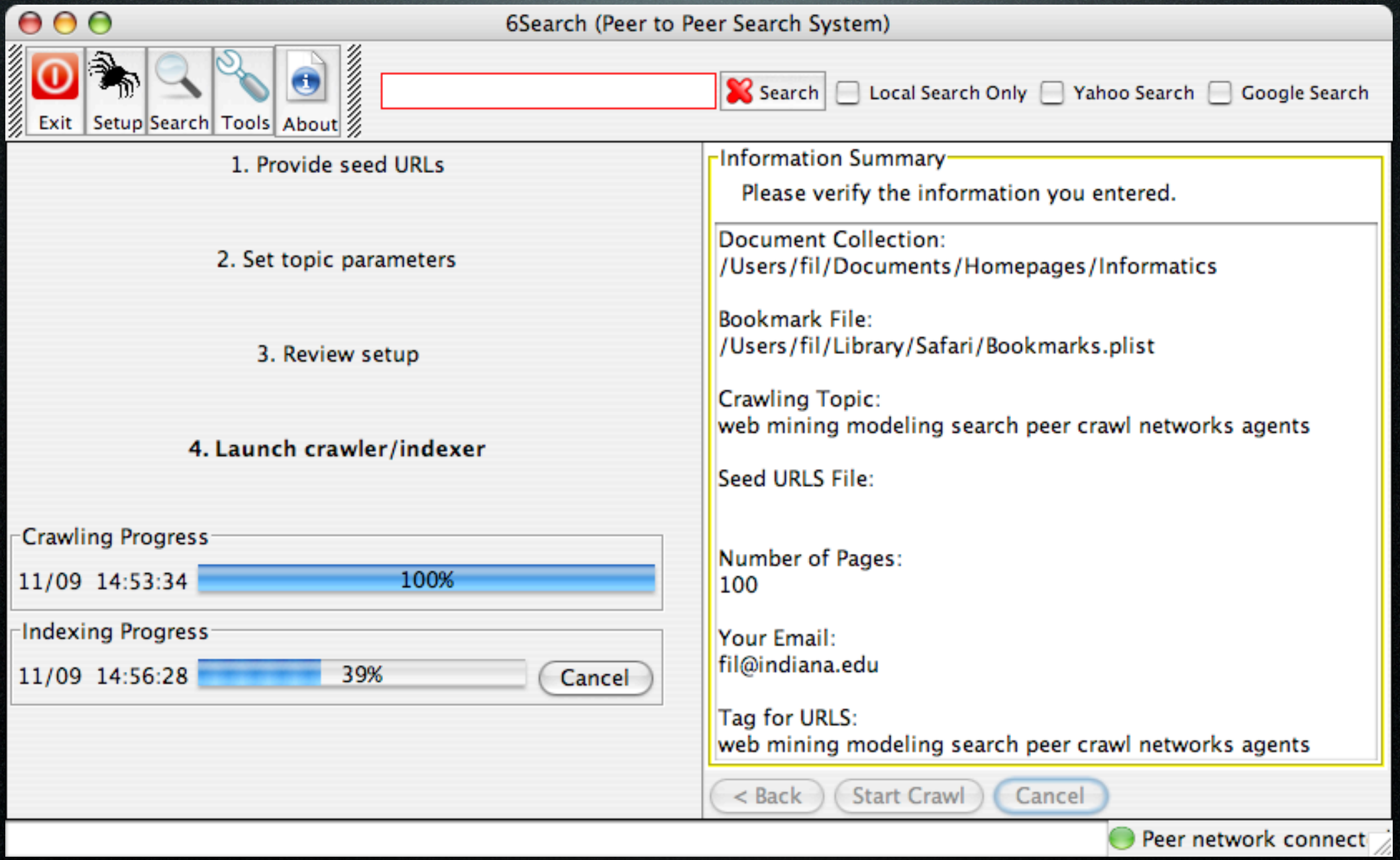
Sponsored links *

Paris Hilton: Book now and save big
The Simple Way to Cheap Travel. Just click - you're there.
CheapTickets.com

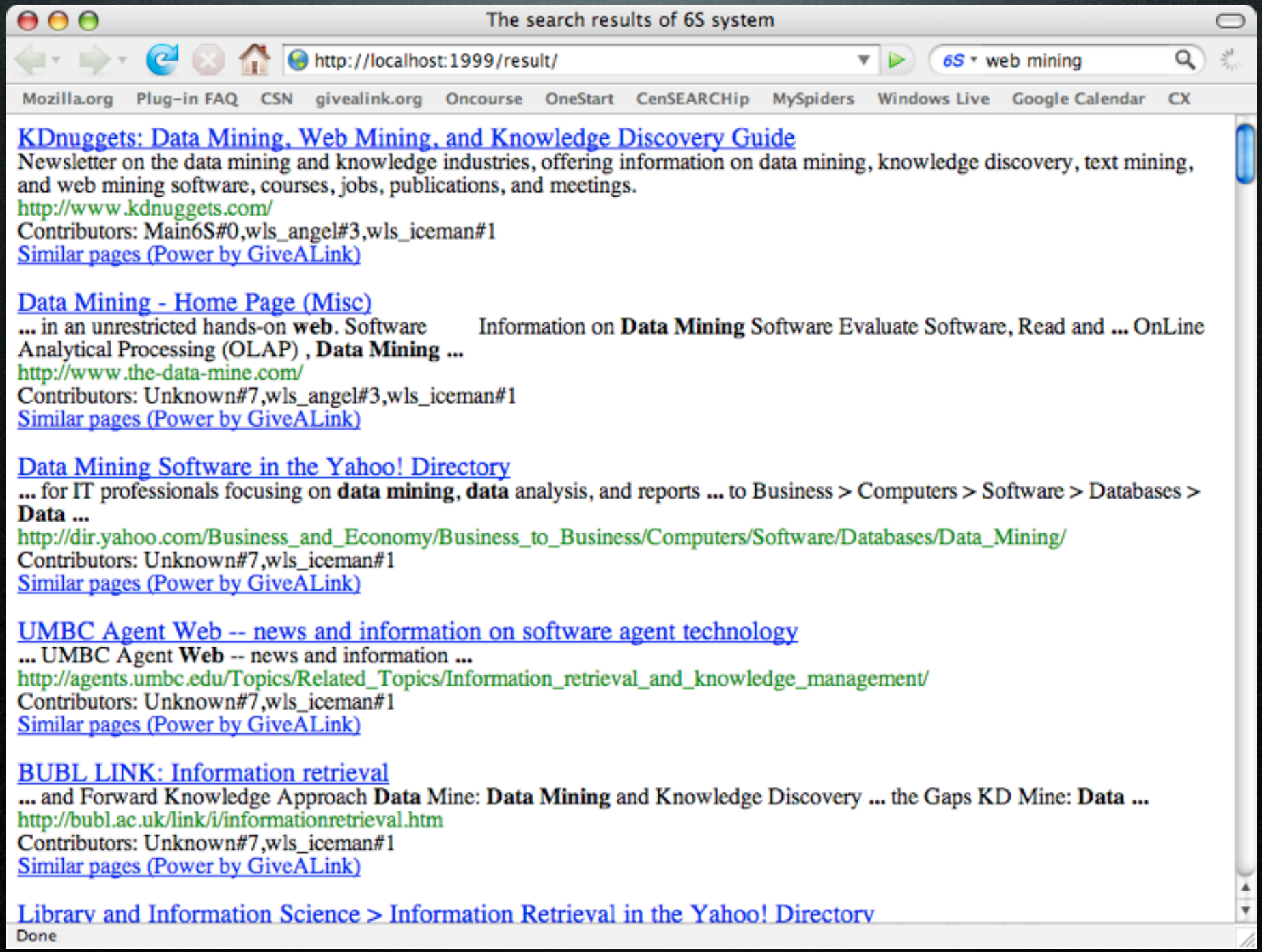
http://sixearch.org



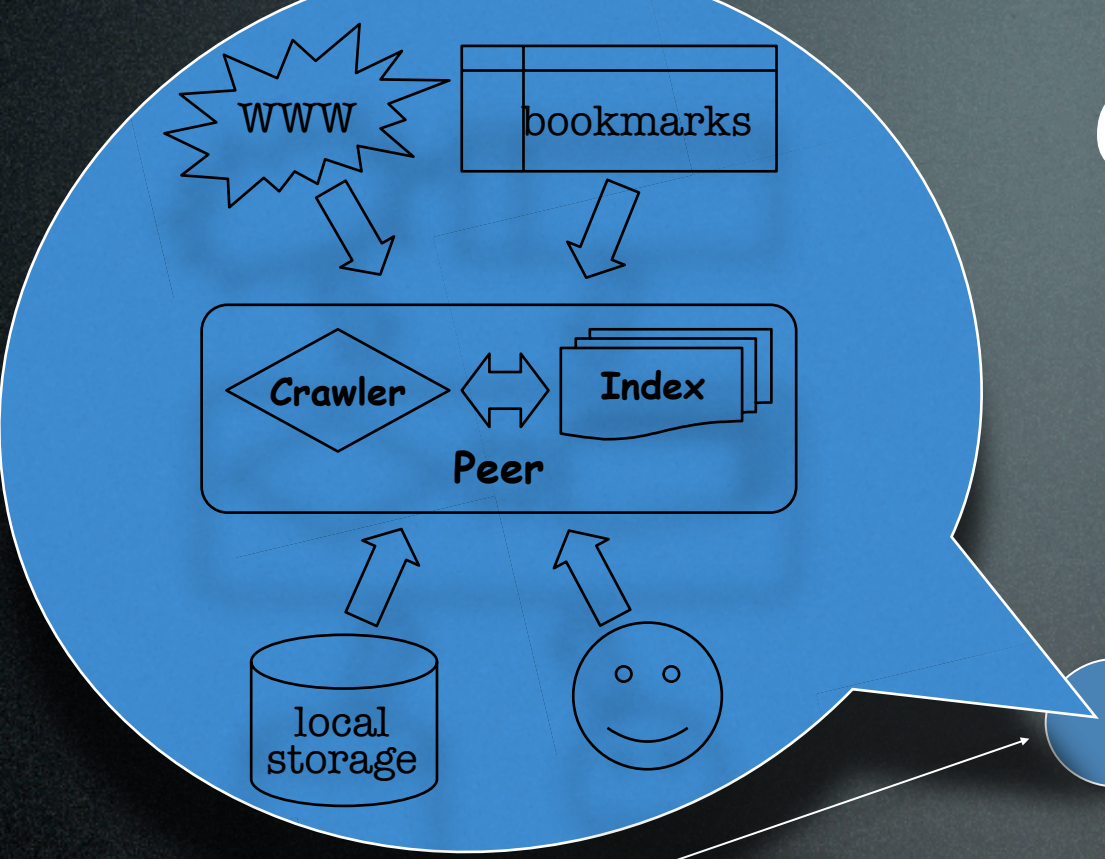
http://sixearch.org

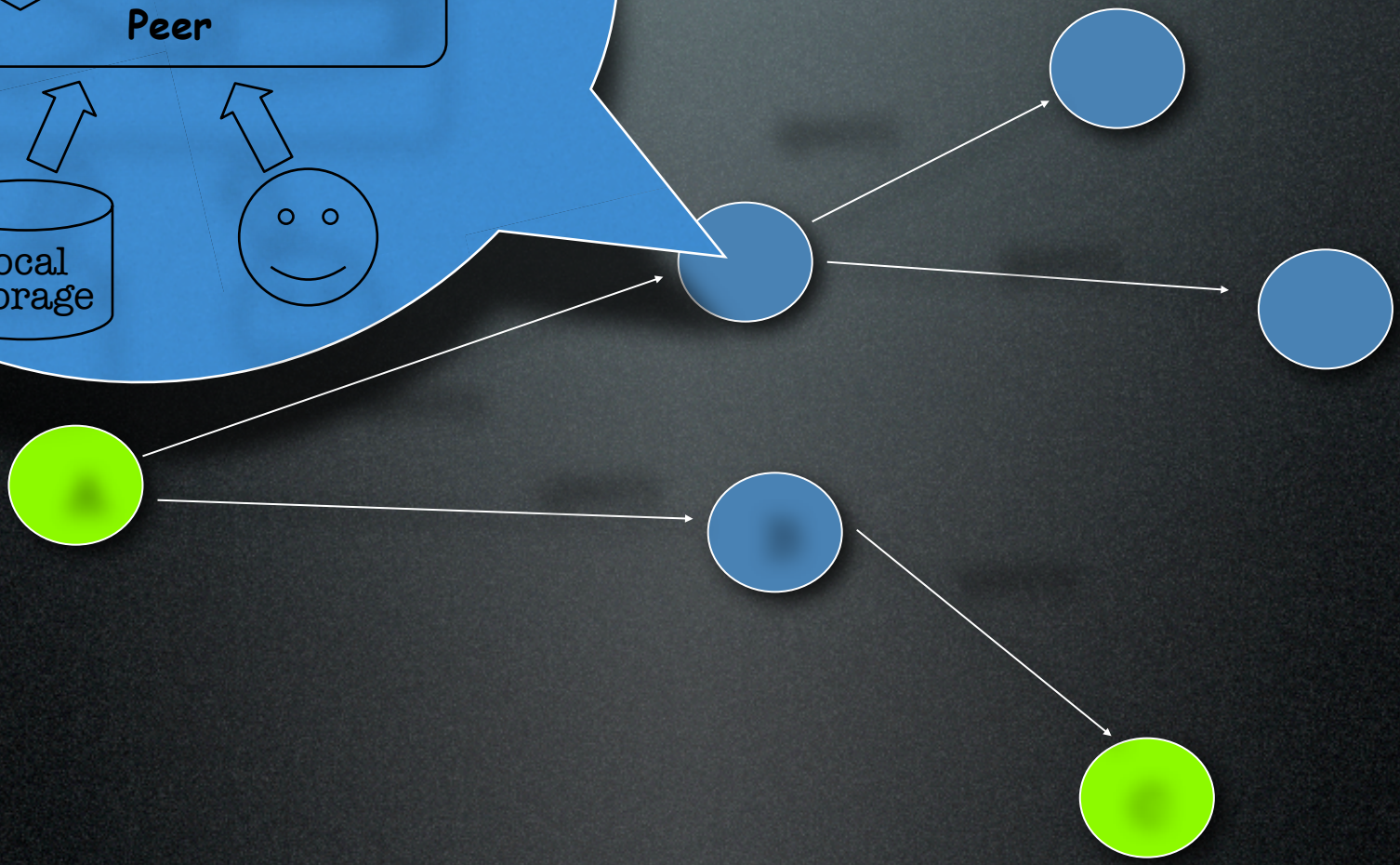


http://sixearch.org



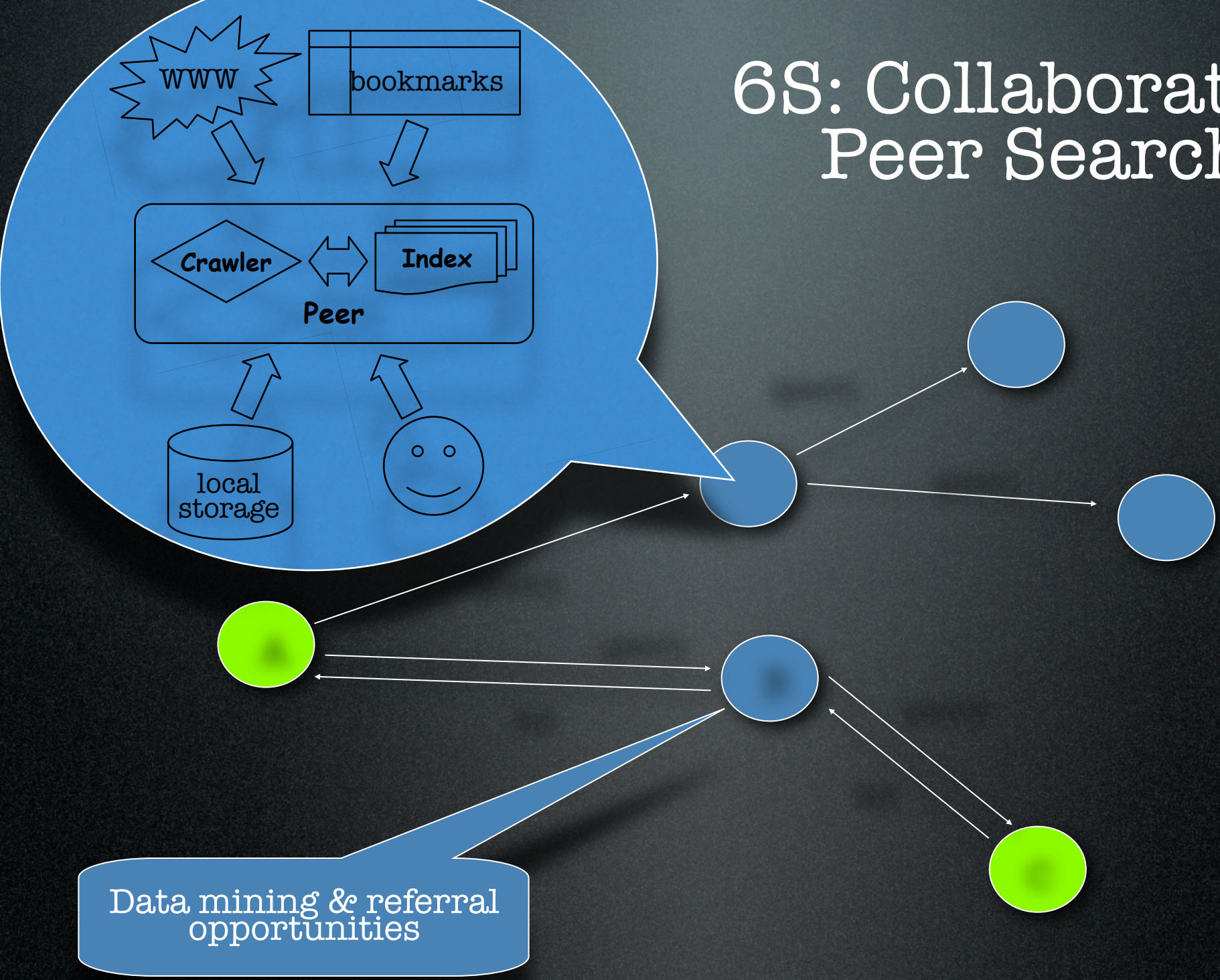
6S: Collaborative Peer Search





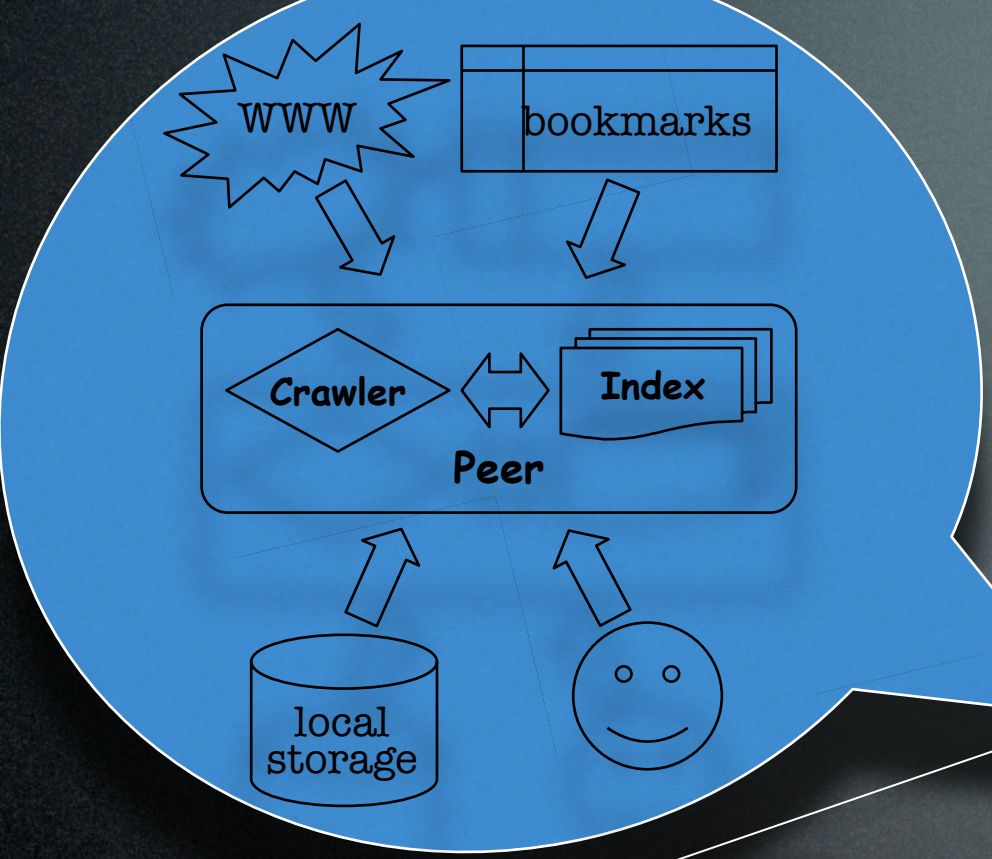
WWW2004
WWW2005
WTAS2005
P2PIR2006

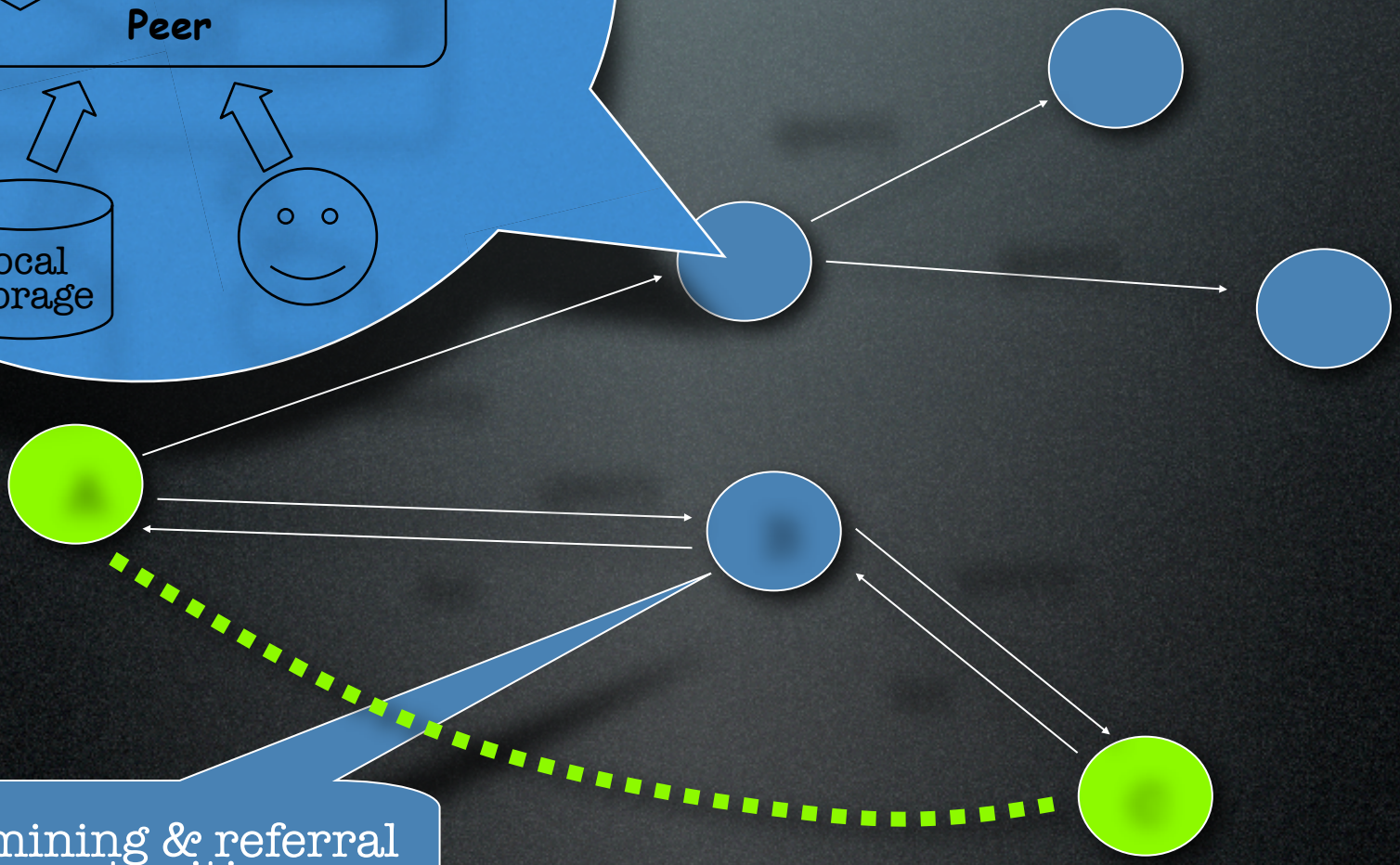
6S: Collaborative Peer Search



WWW2004
WWW2005
WTAS2005
P2PIR2006

6S: Collaborative Peer Search

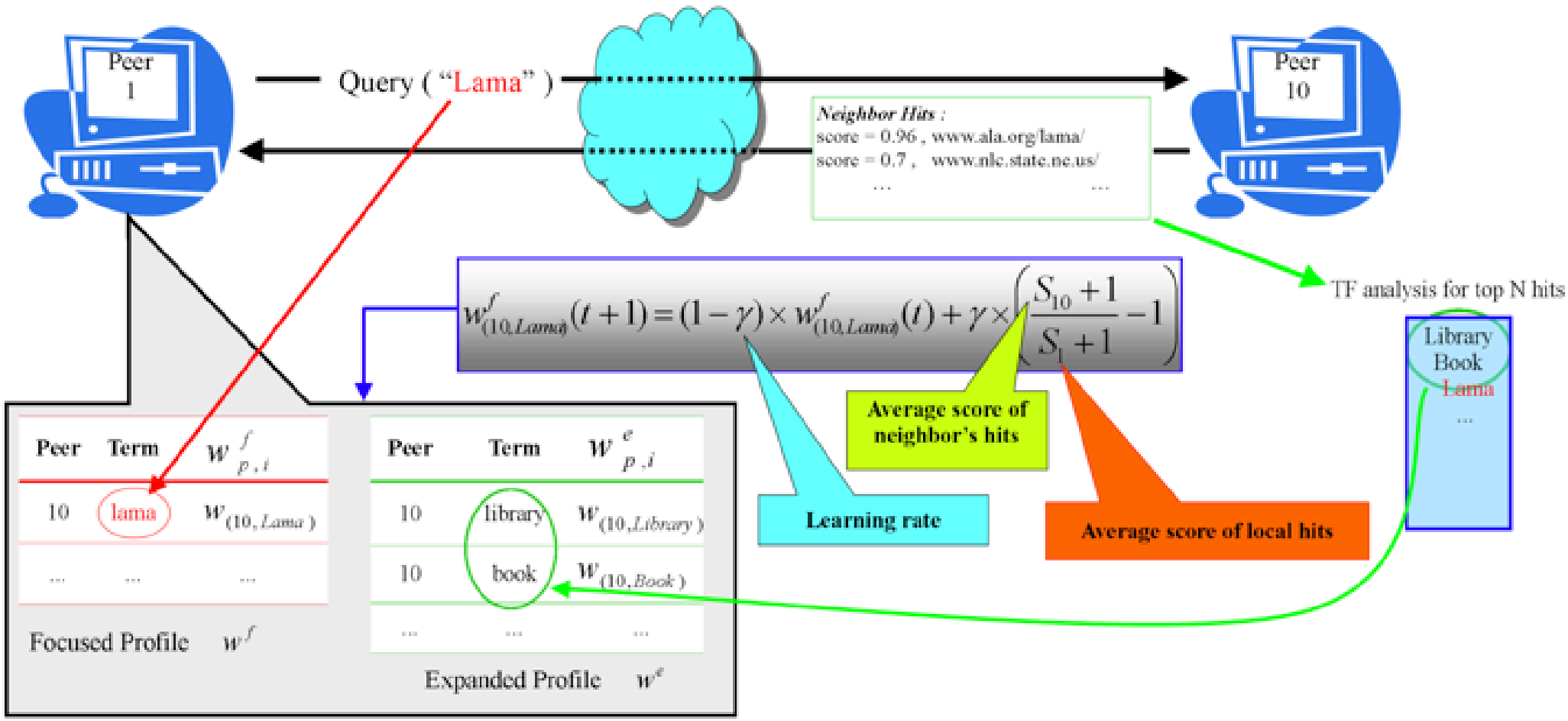




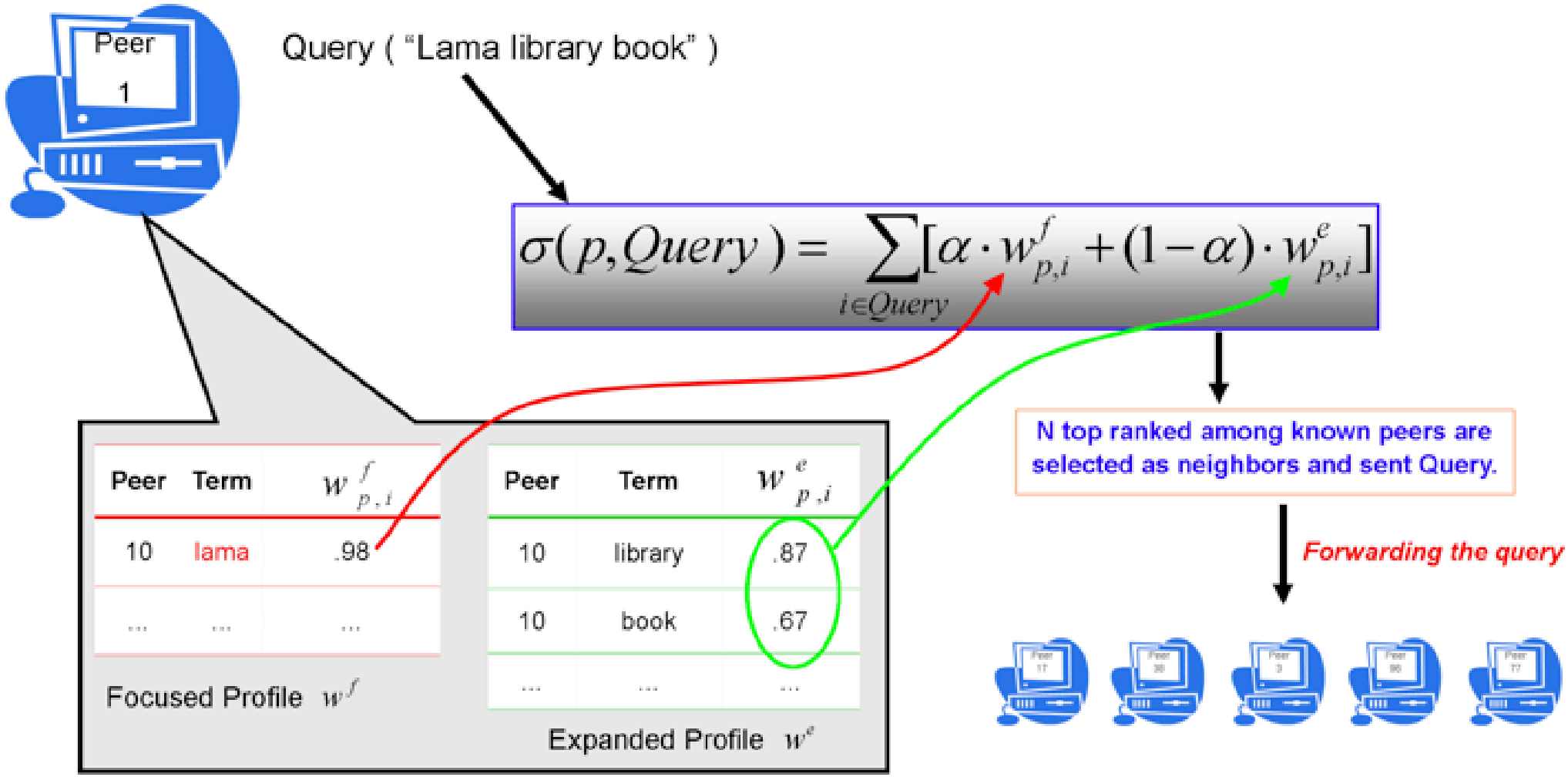
WWW2004
WWW2005
WTAS2005
P2PIR2006

Data mining & referral opportunities

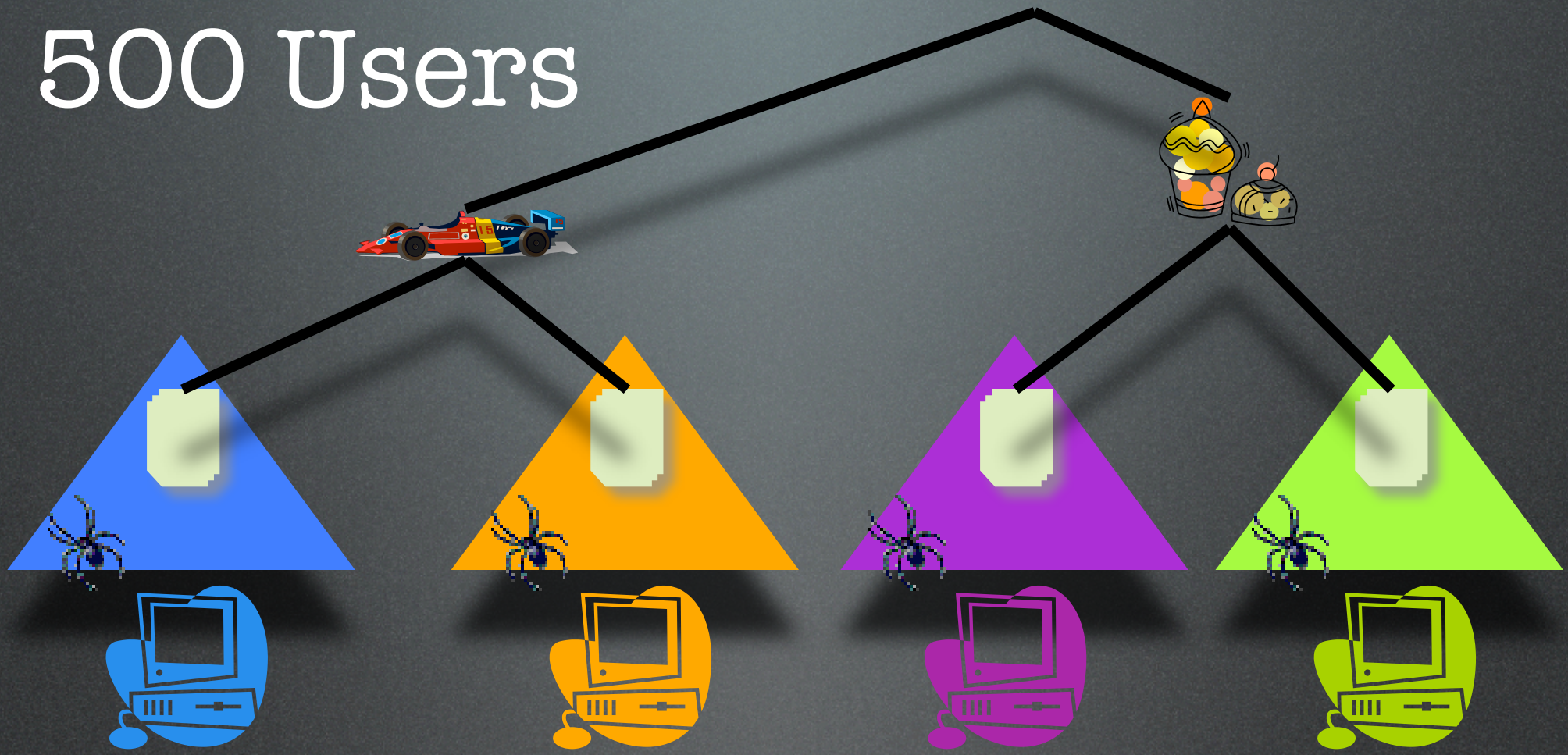
Reinforcement Learning



Query Routing

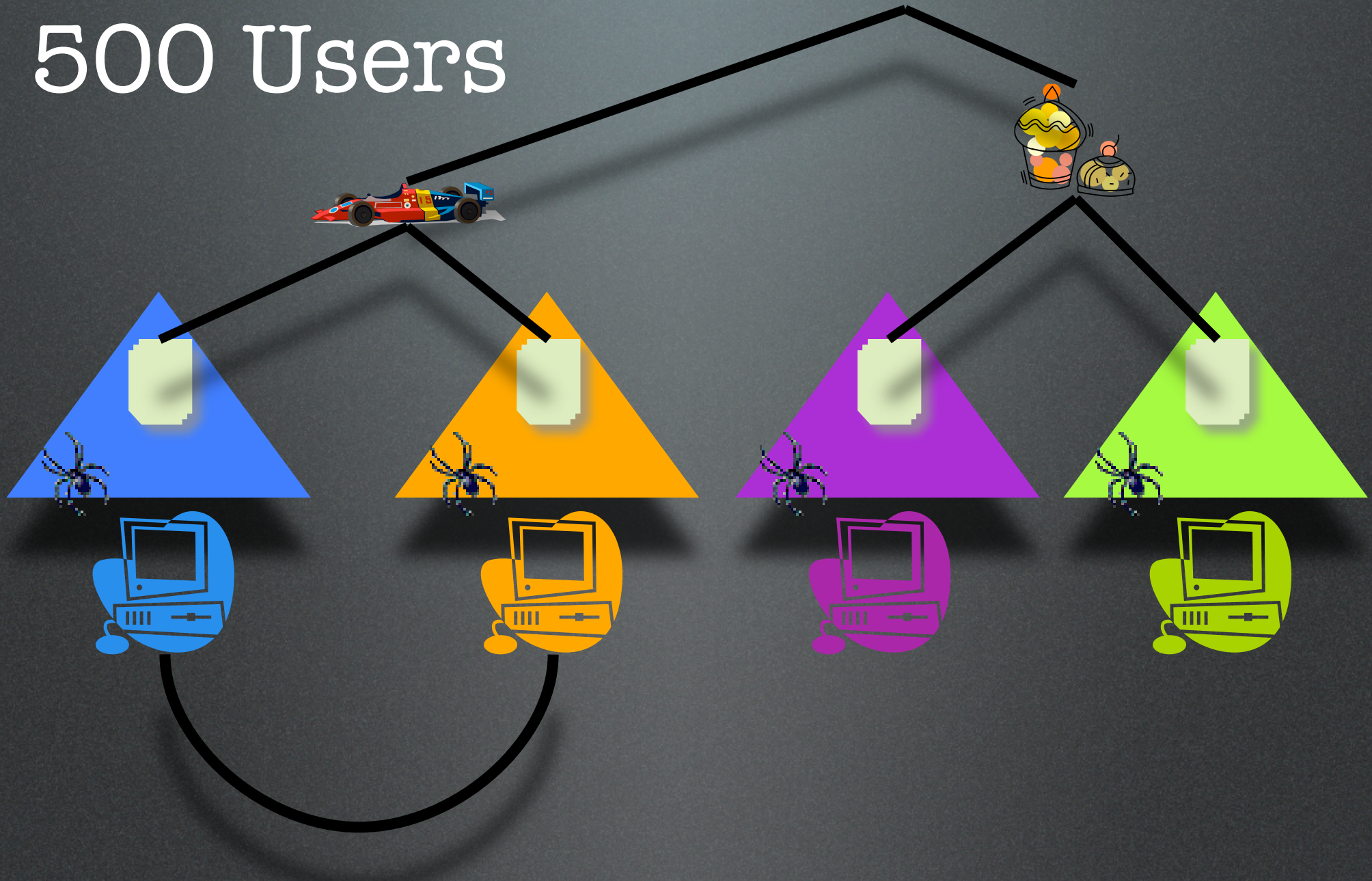


Simulating 500 Users



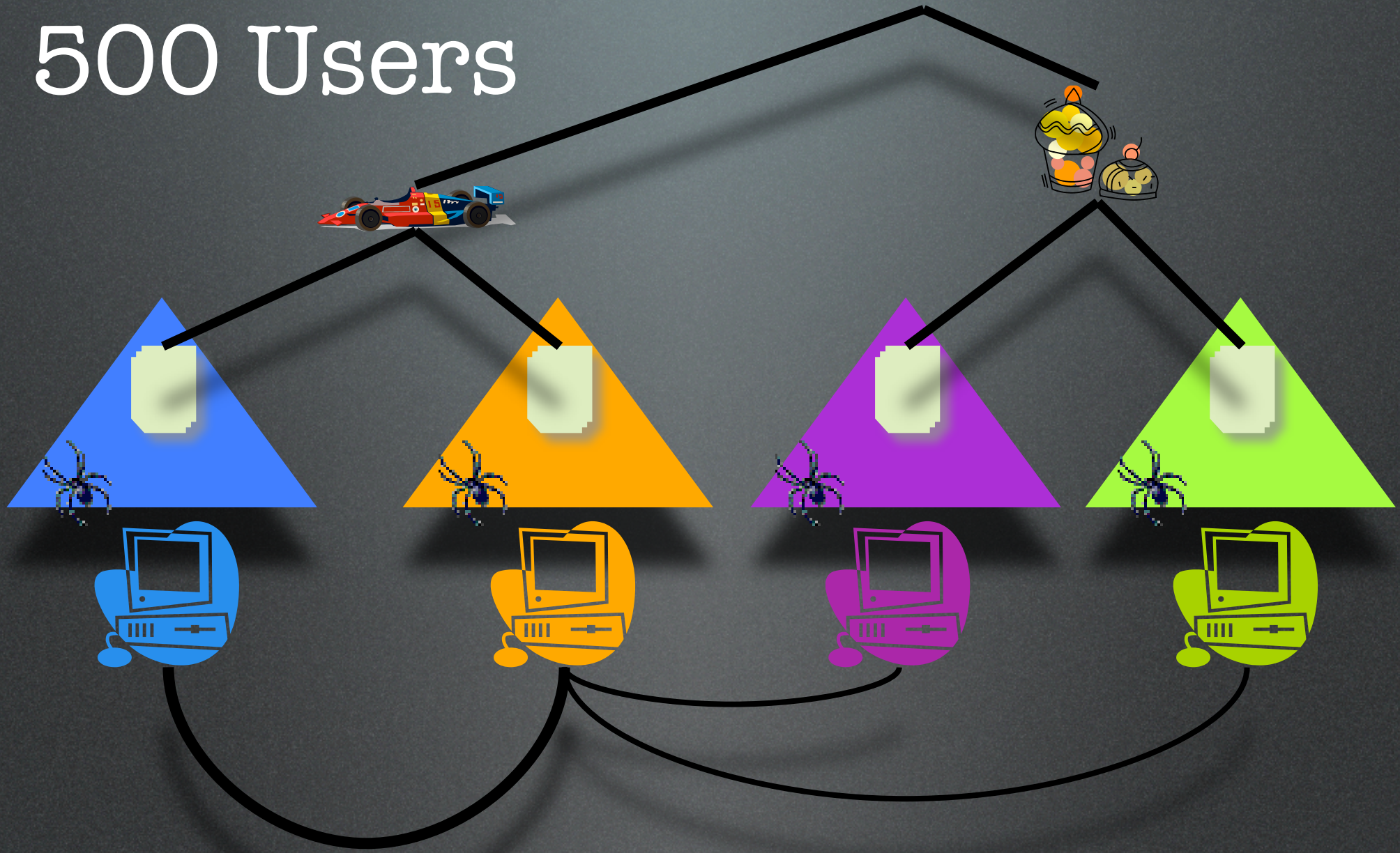
Maguitman, Menczer et al.: Algorithmic computation and approximation of semantic similarity. WWW2005, WWWJ 2006

Simulating 500 Users

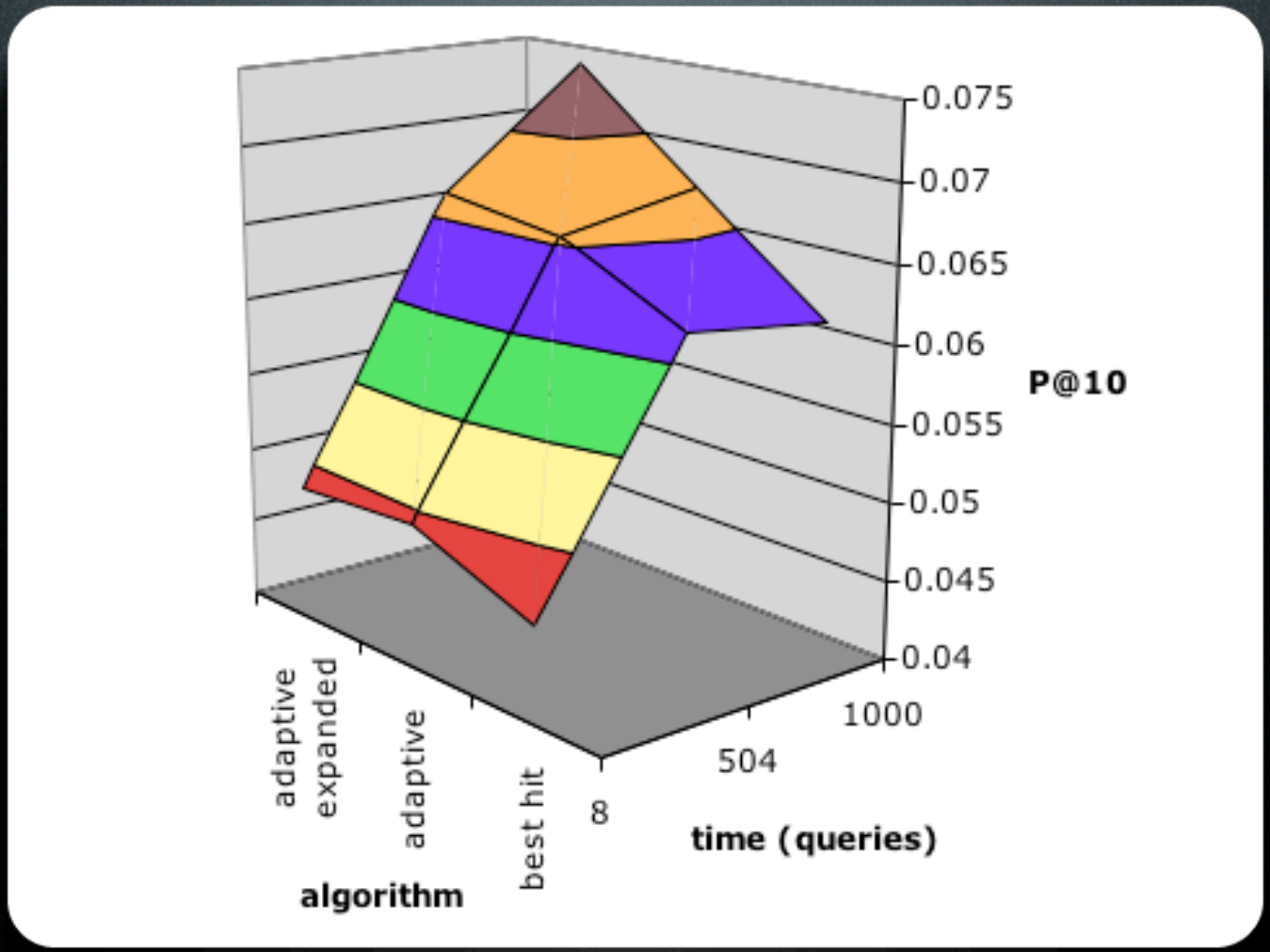


Maguitman, Menczer et al.: Algorithmic computation and approximation of semantic similarity. WWW2005, WWWJ 2006

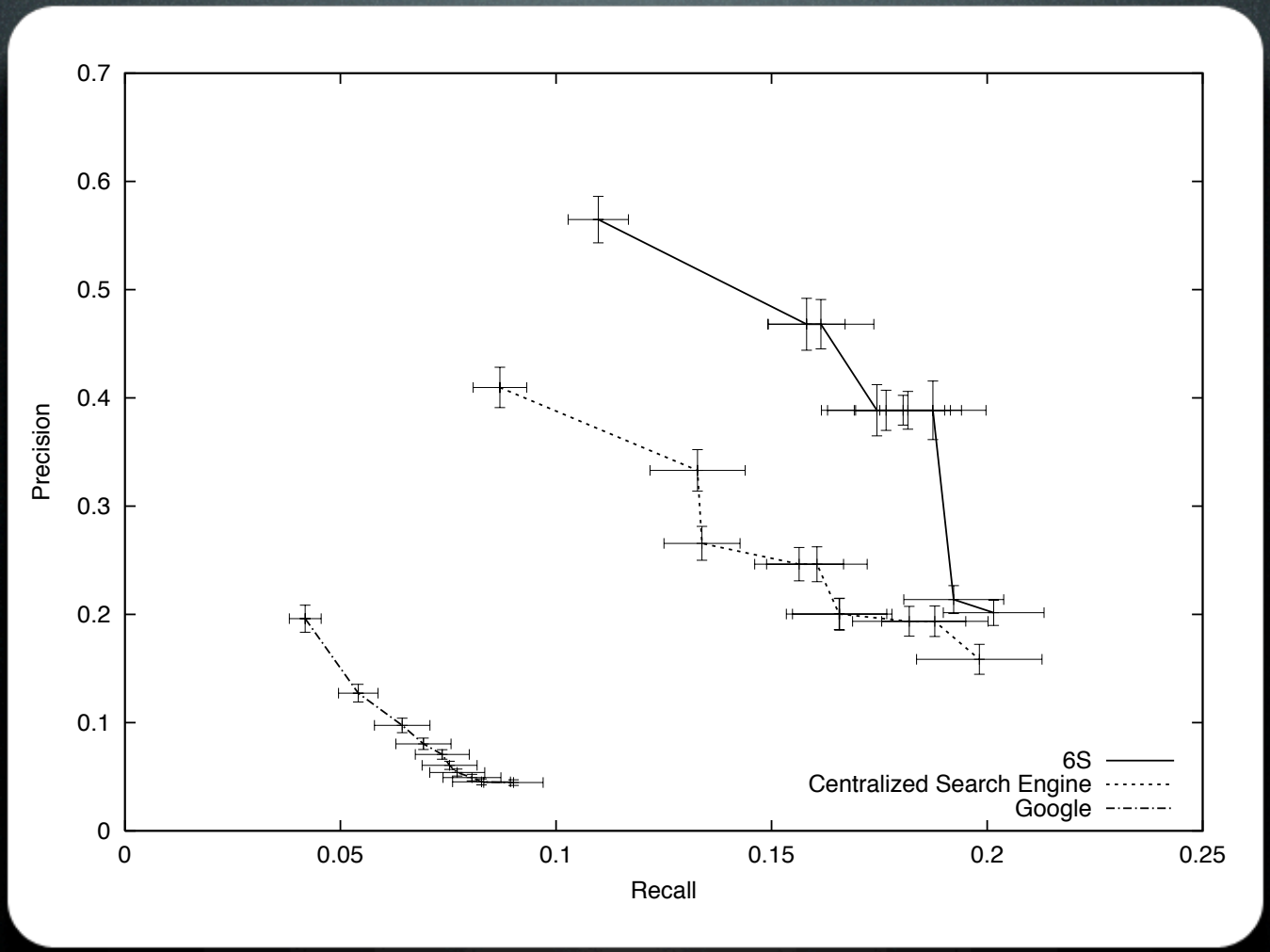
Simulating 500 Users



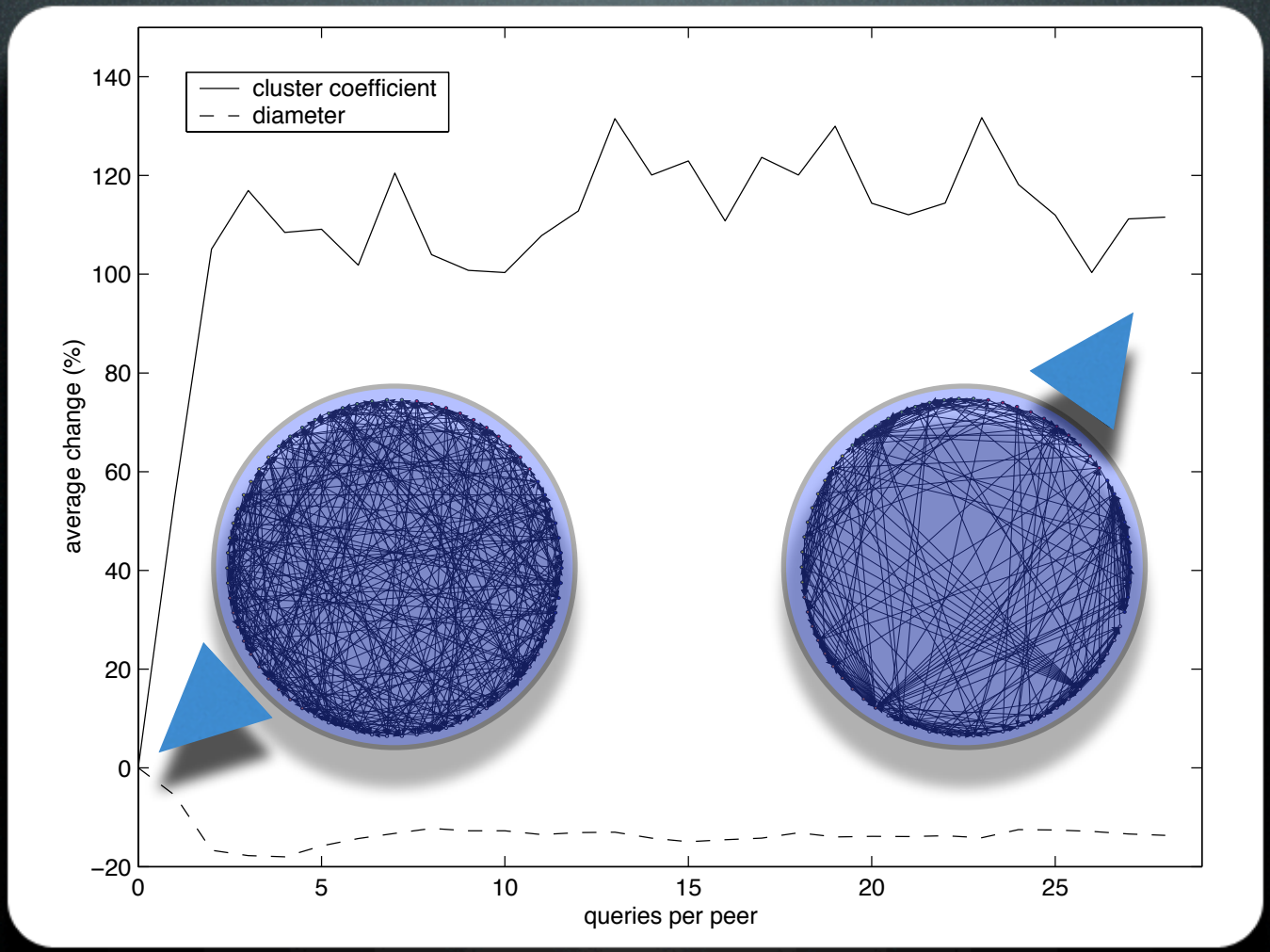
Maguitman, Menczer et al.: Algorithmic computation and approximation of semantic similarity. WWW2005, WWWJ 2006


0.075
0.07
0.065
0.06
0.055
0.05
0.045
0.04
P@10
adaptive expanded
adaptive
best hit
algorithm
8
504
1000
time (queries)

P@10

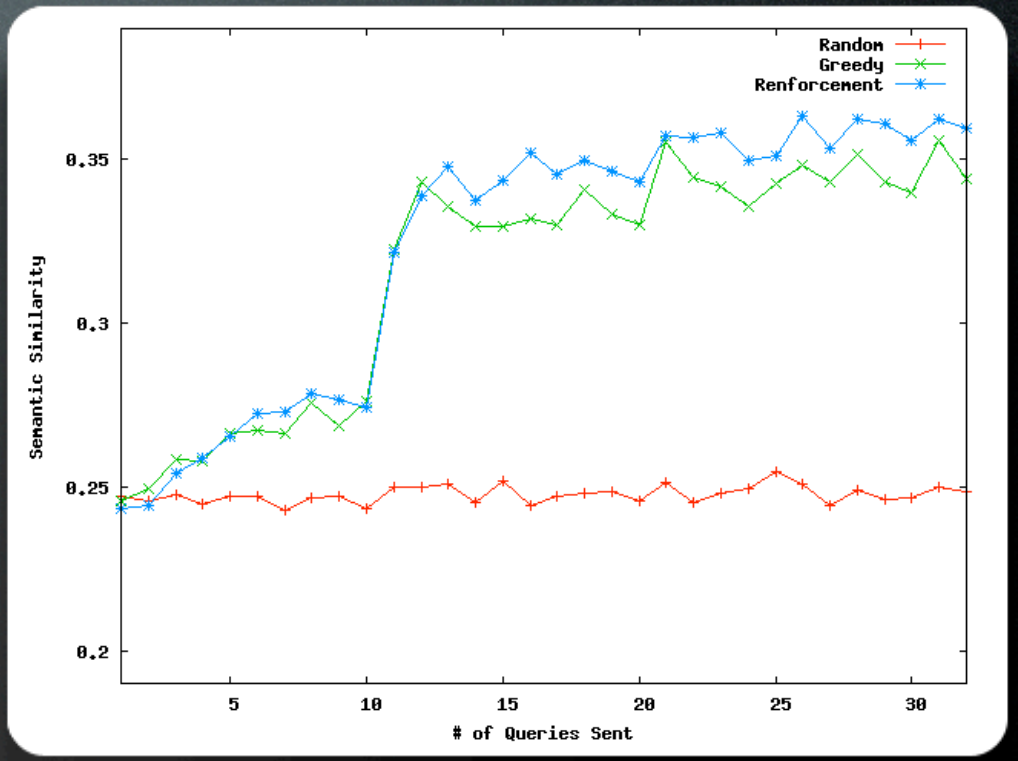


Distributed vs Centralized

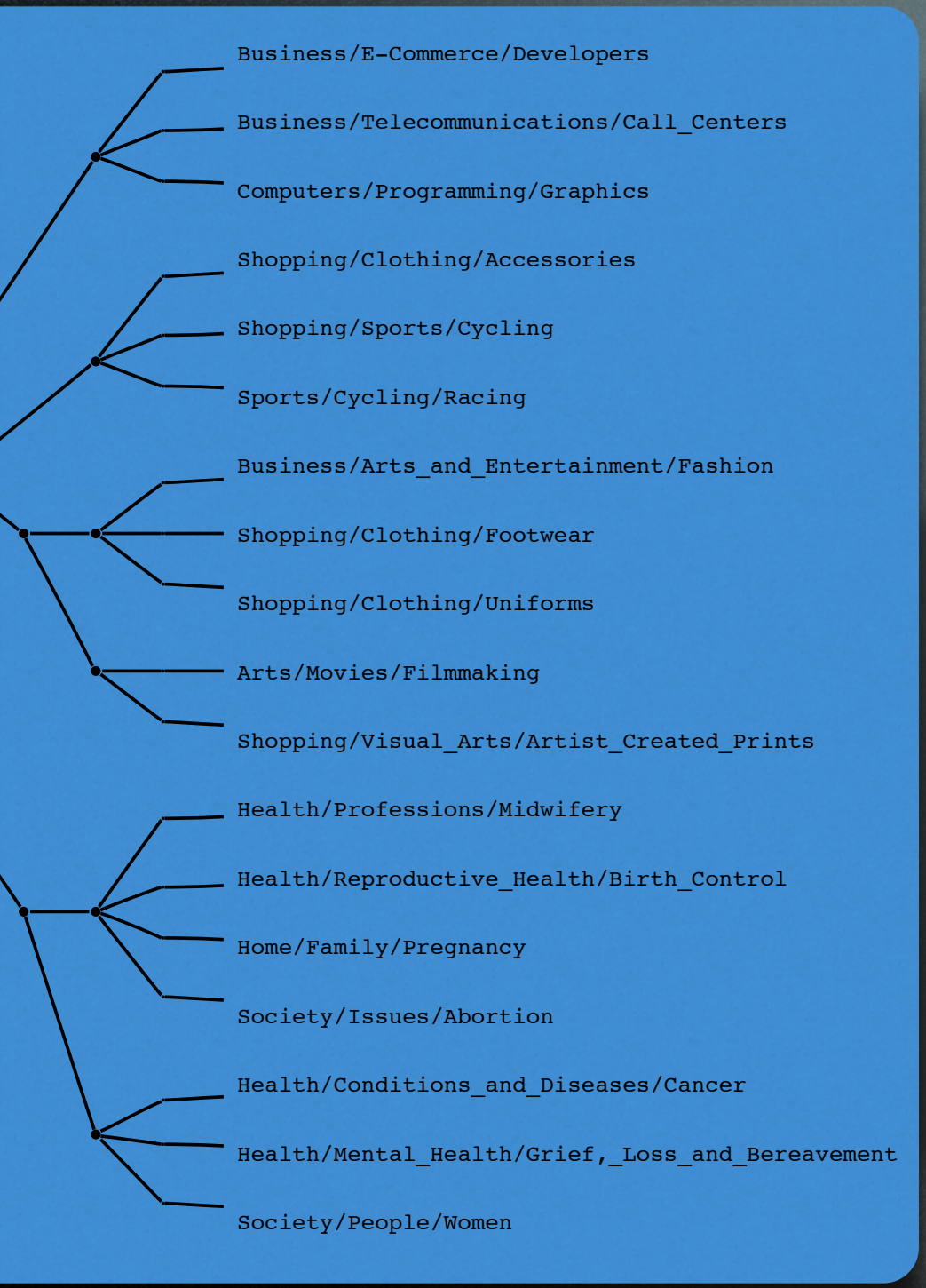

cluster coefficient
diameter
average change (%)
queries per peer
140
120
100
80
60
40
20
0
−20
0
5
10
15
20
25

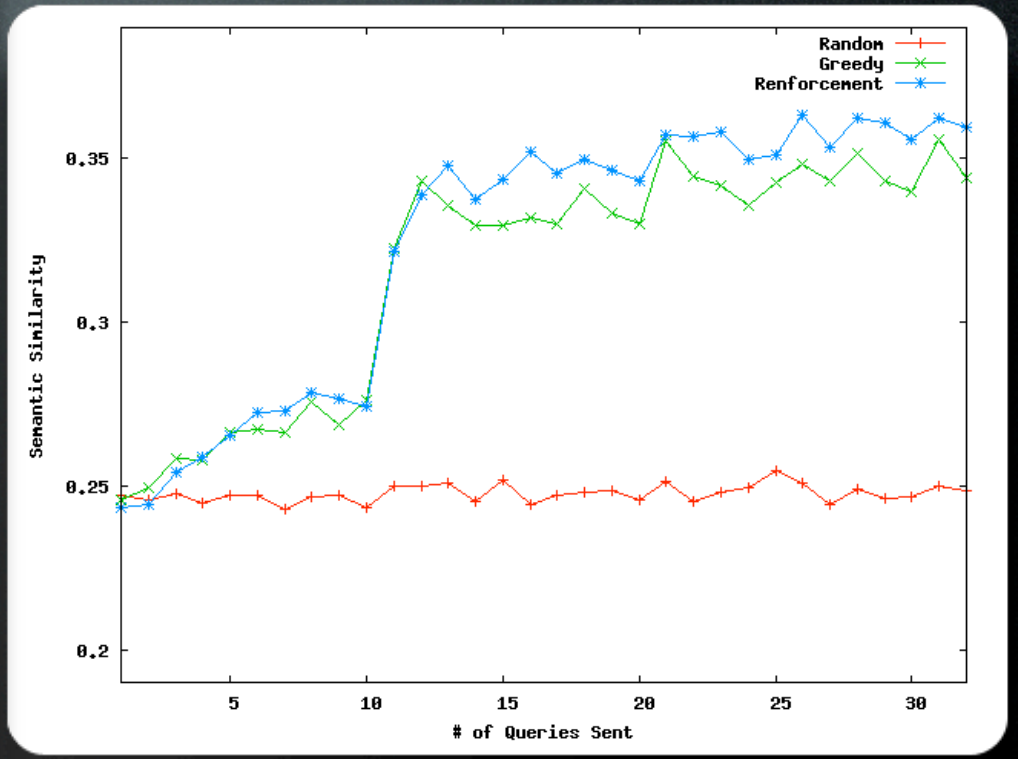
Small-world

Semantic Similarity



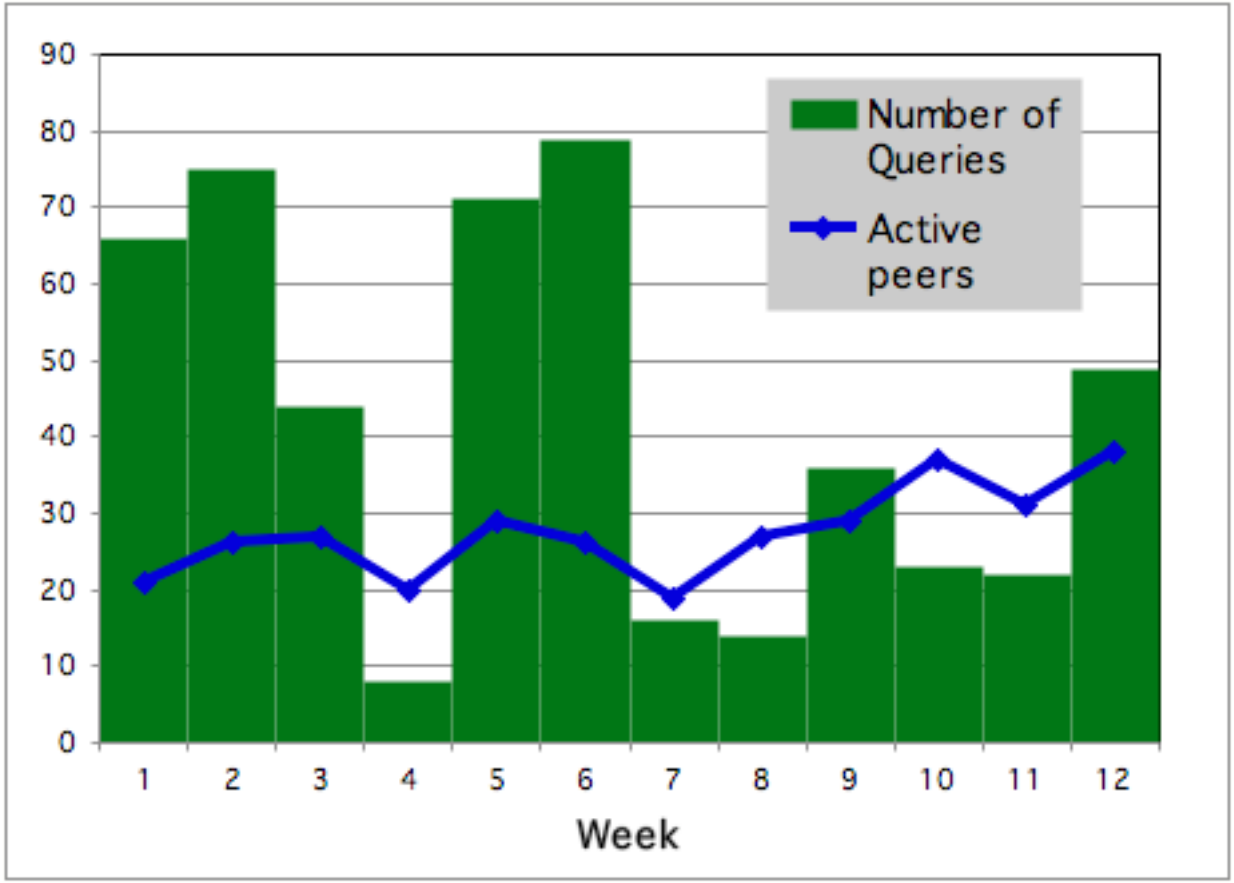
Semantic Similarity



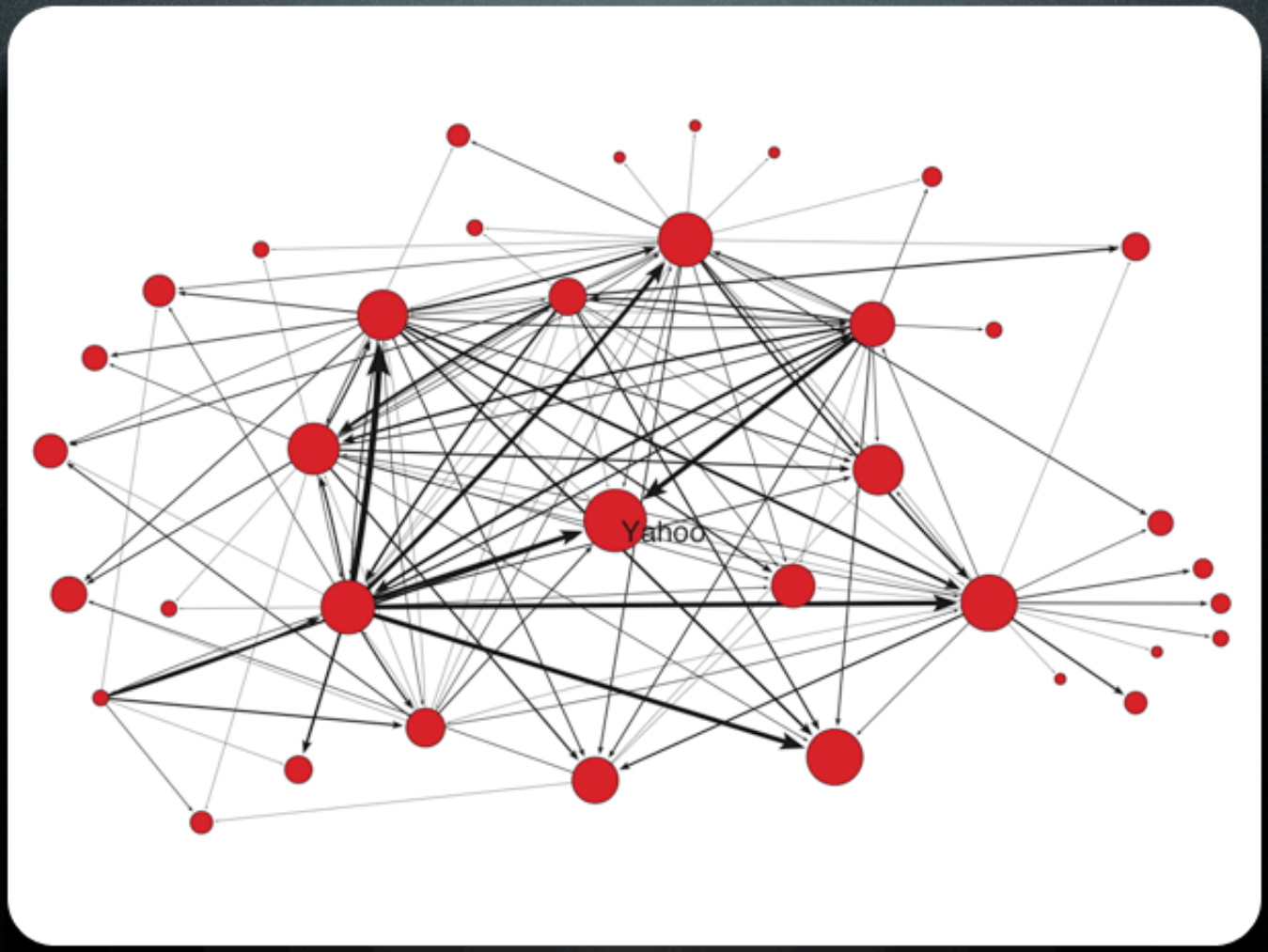


Ongoing Work

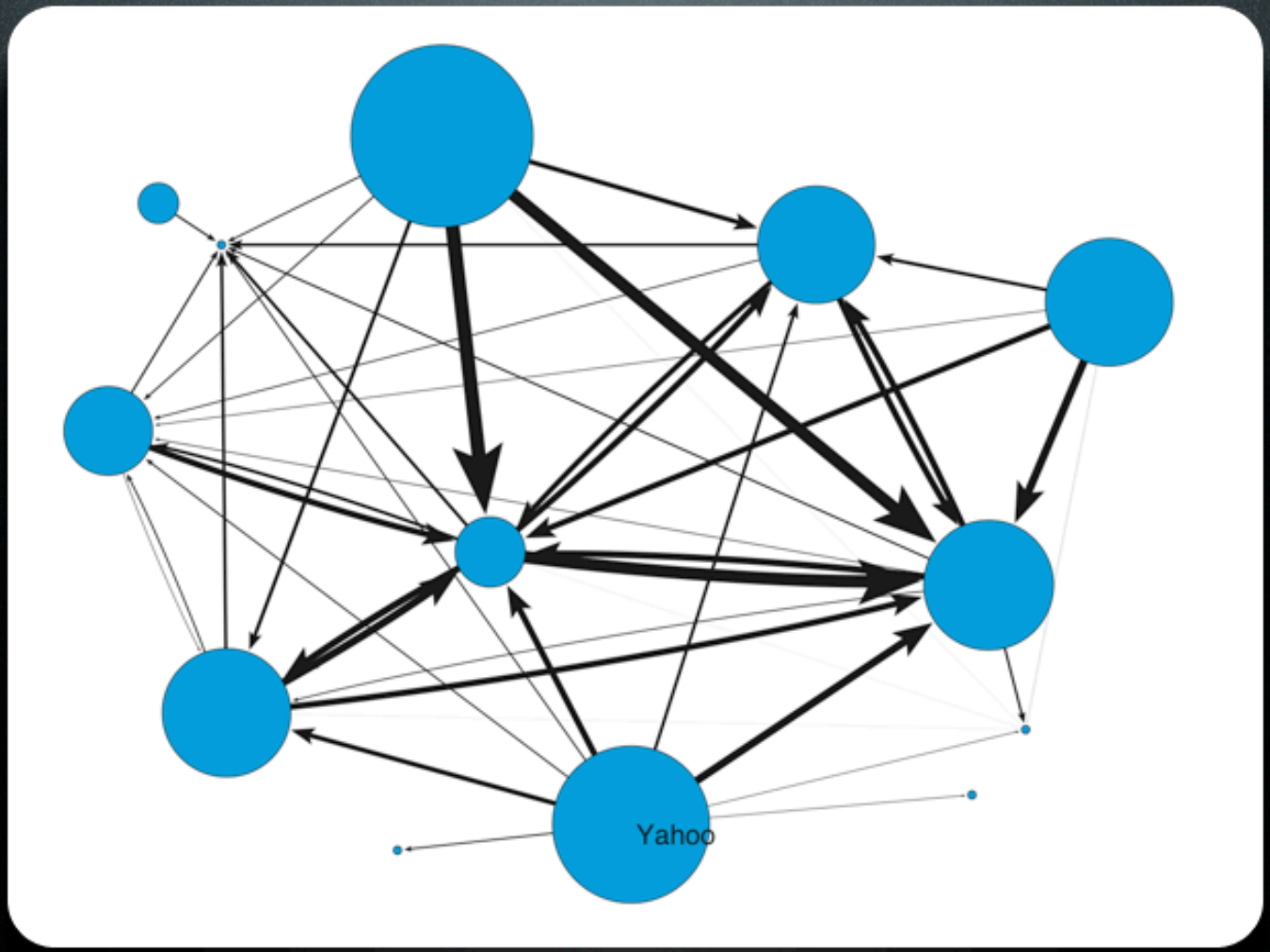
- Improve coverage/diversity in query routing algorithm
- Spam protection: trust/reputation subsystem
- User study with 6S application


90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
11
12
Week
Number of Queries
Active peers

User study

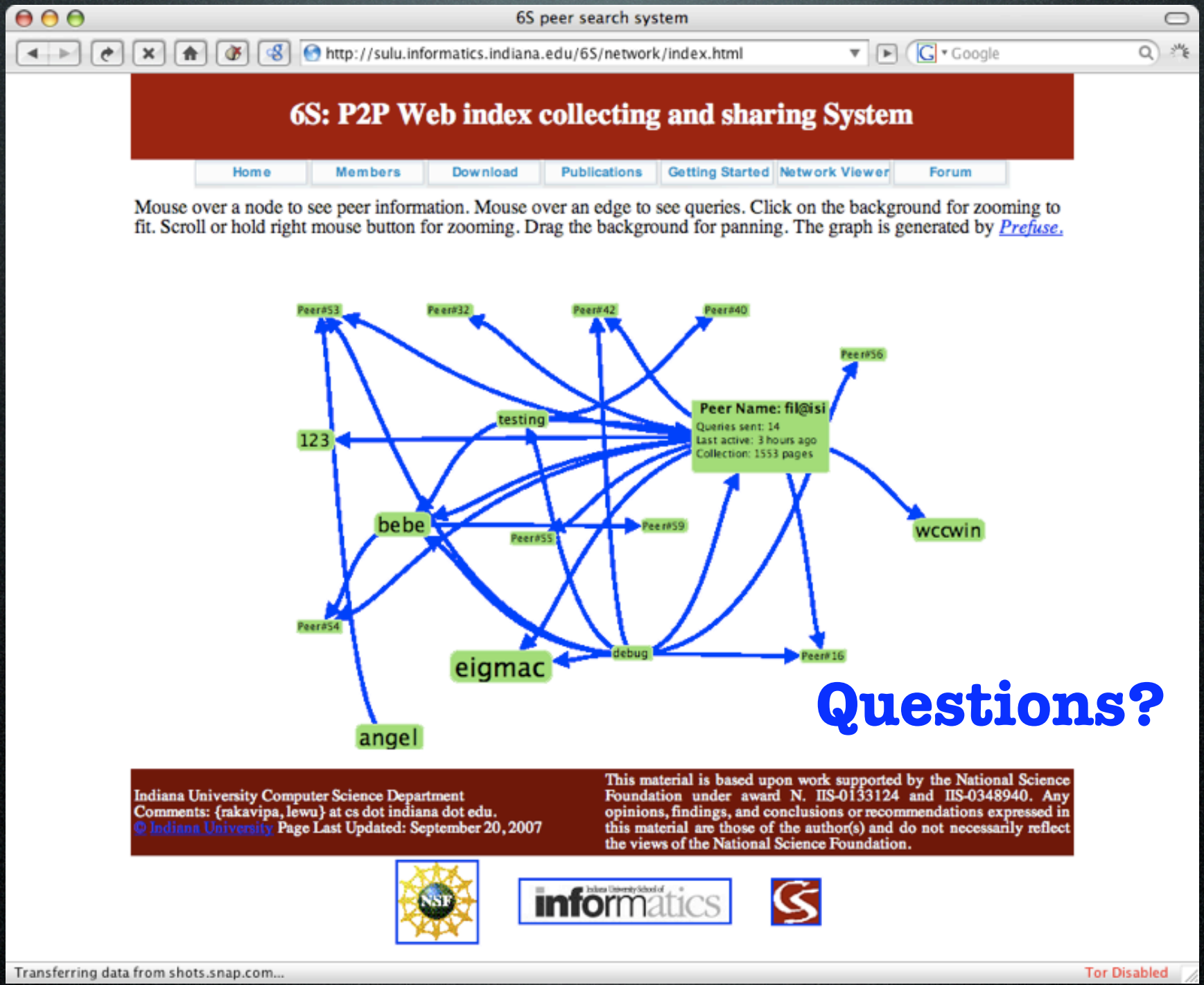


Query network



Result network

http://sixearch.org



Thank you!
Questions?

informatics.indiana.edu/fil

Indiana University School of informatics

Research supported by NSF
CAREER Award IIS-0348940