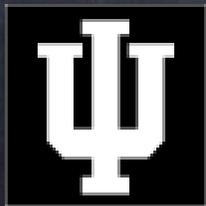


# Navigating the Web graph

Workshop on Networks and Navigation  
Santa Fe Institute, August 2008

Filippo Menczer  
Informatics & Computer Science  
Indiana University, Bloomington

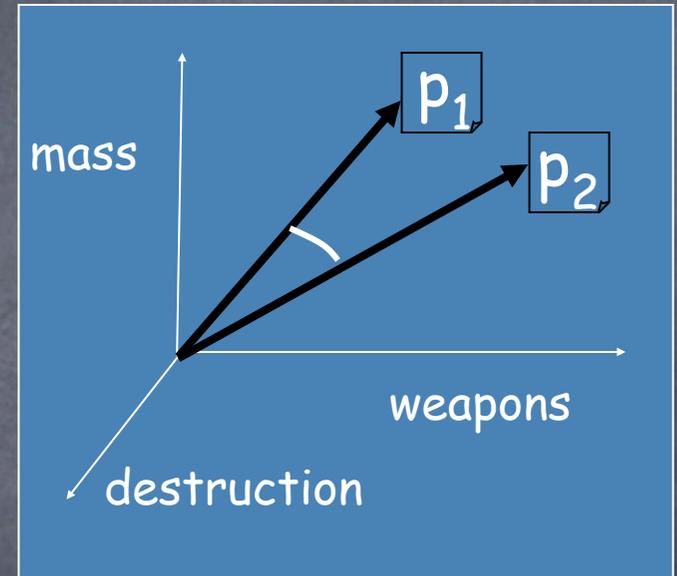


# Outline

- 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

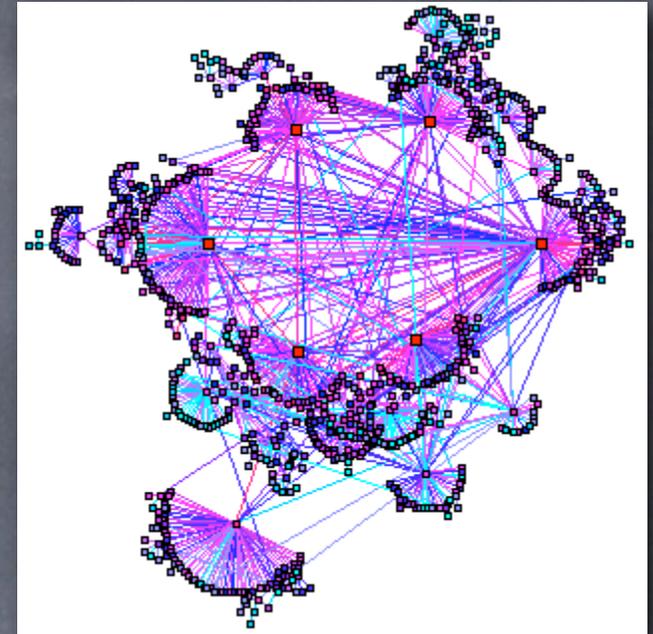
- Pages close in **word vector space** tend to be related
- Cluster hypothesis (van Rijsbergen 1979)
- The WebCrawler (Pinkerton 1994)
- The whole first generation of search engines



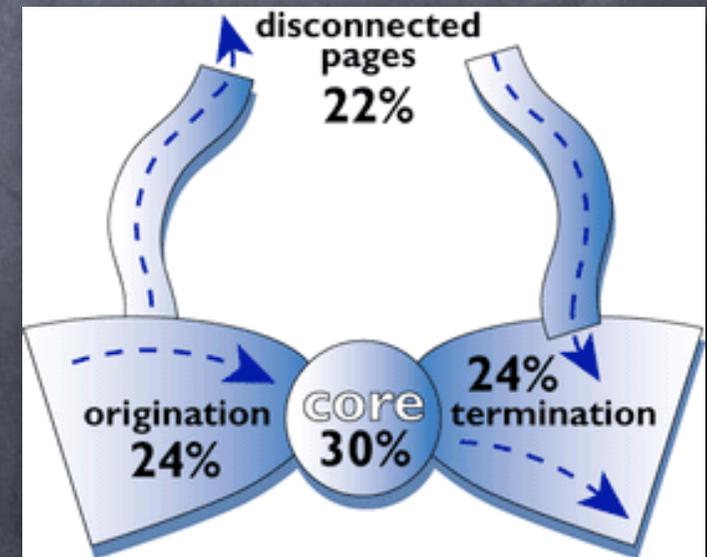
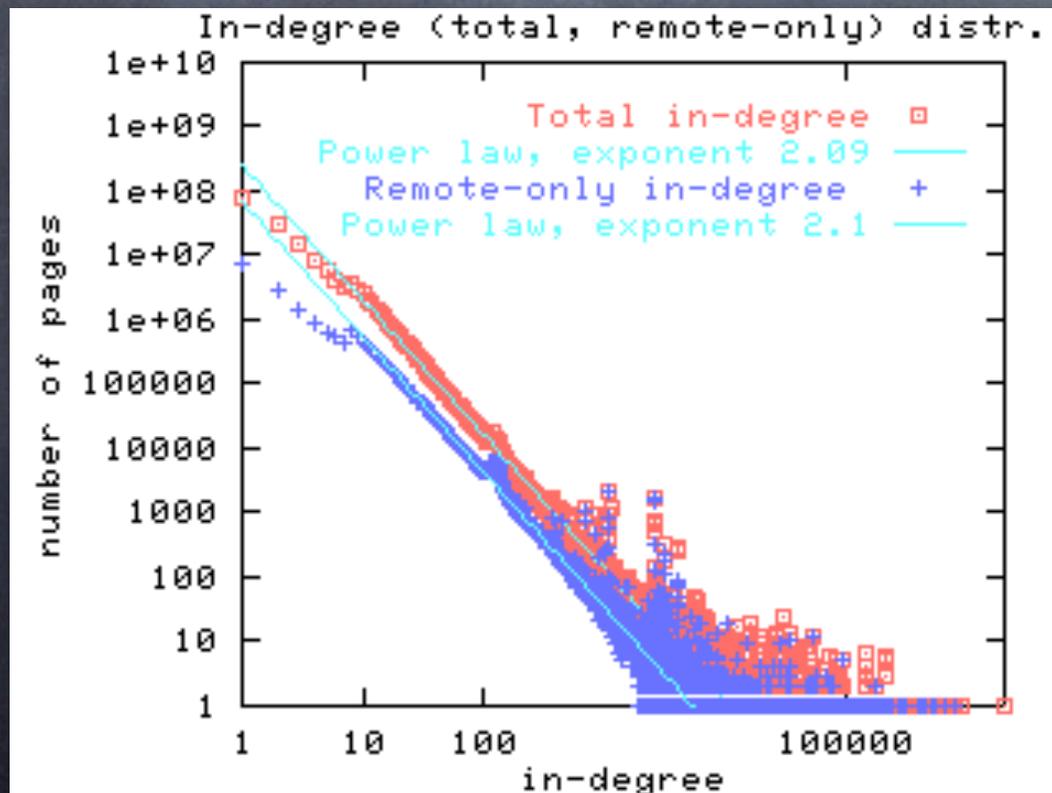
# Enter the Web's link structure

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

Brin & Page 1998

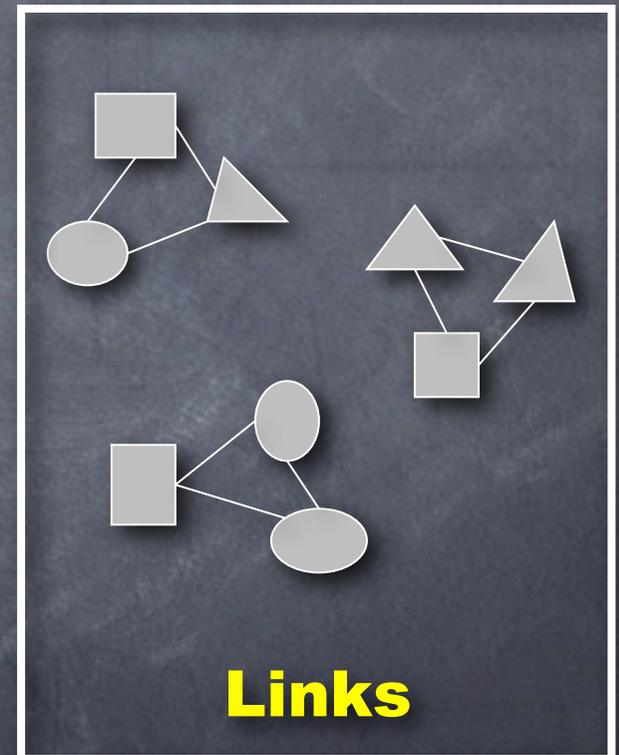
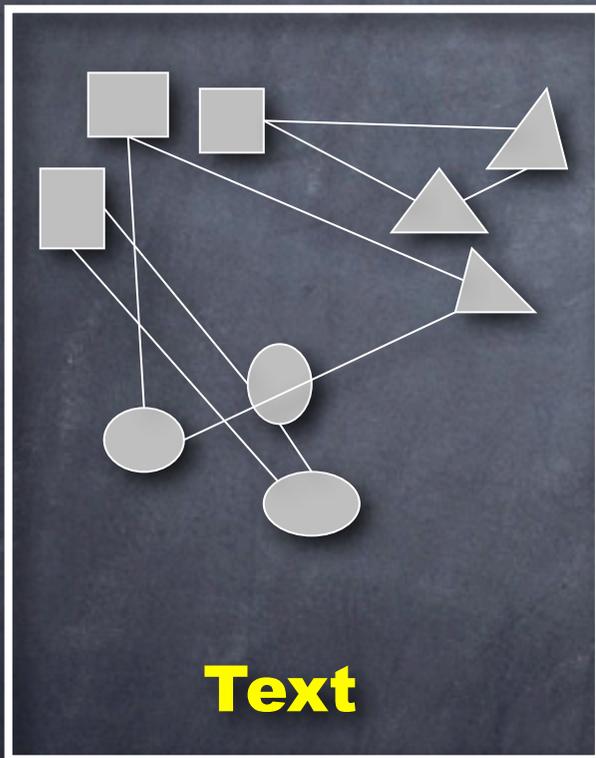


Barabasi & Albert 1999

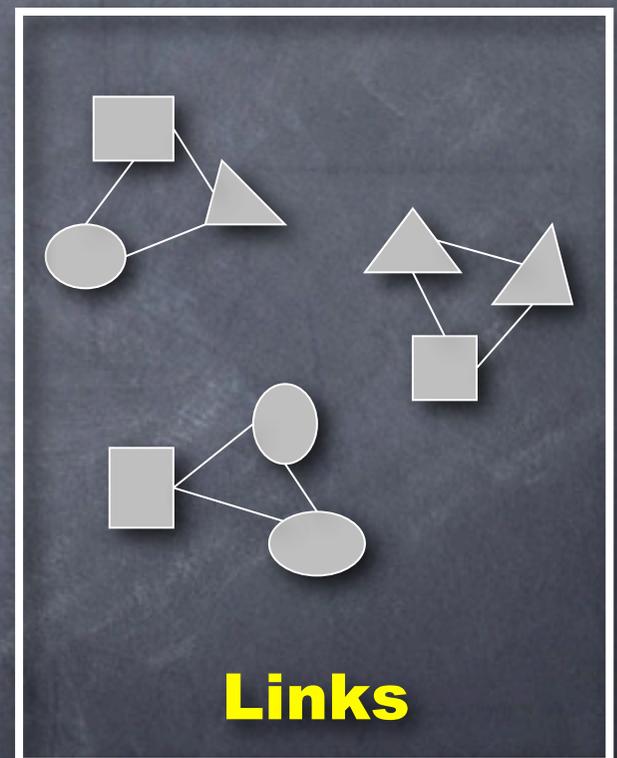
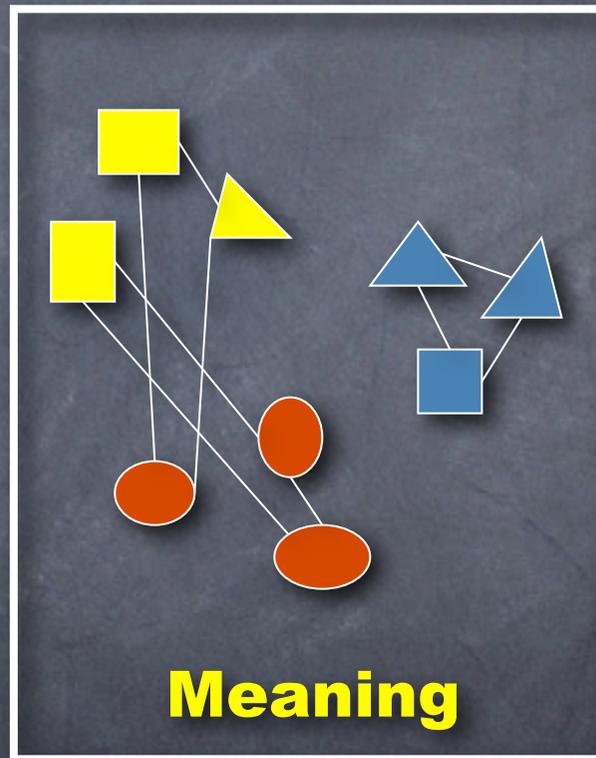
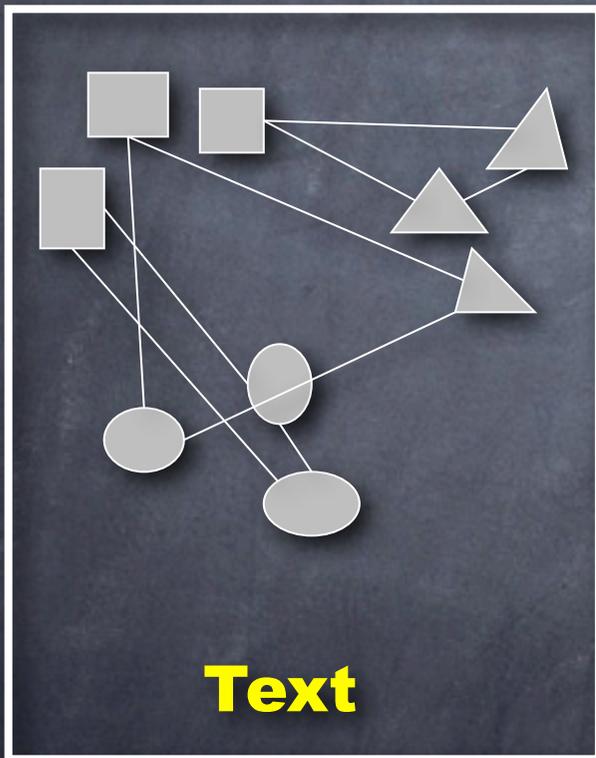


Broder & al. 2000

# Three network topologies



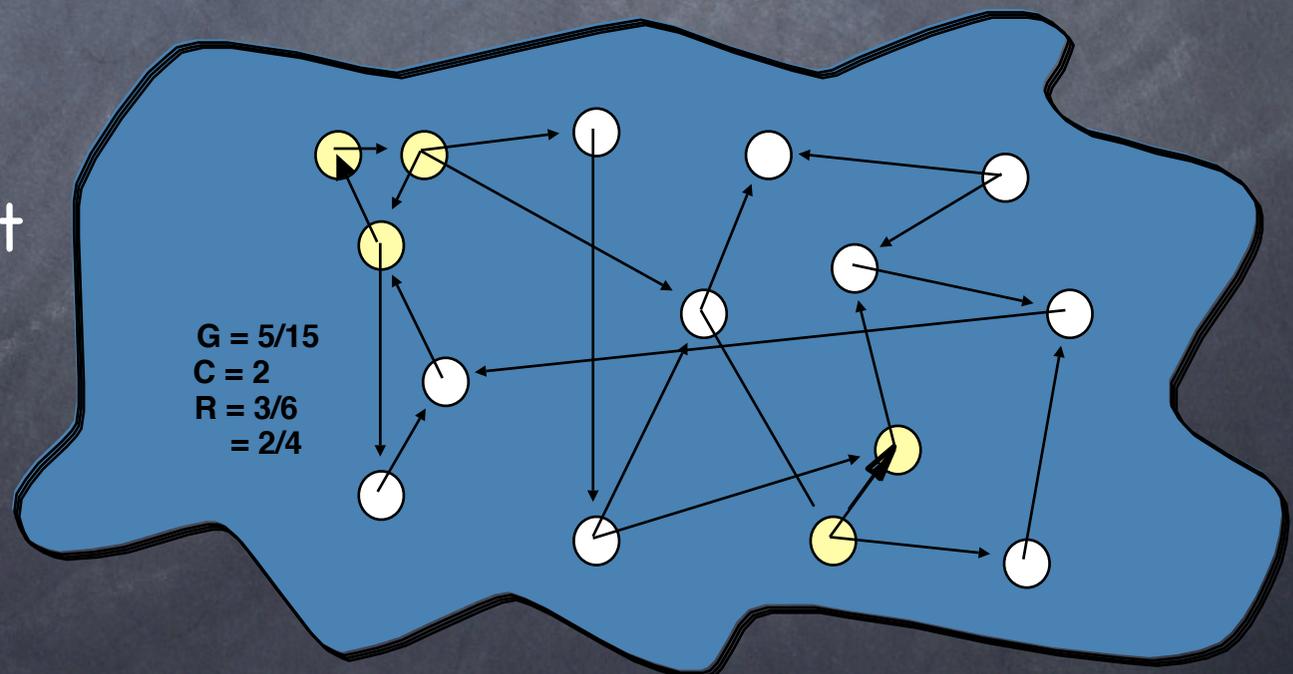
# Three network topologies



# The "link-cluster" conjecture

- Connection between **semantic** topology (topicality or relevance) and **link** topology (hypertext)
- $G = \Pr[\text{rel}(p)] \sim$  fraction of relevant pages (generality)
- $R = \Pr[\text{rel}(p) \mid \text{rel}(q) \text{ AND } \text{link}(q,p)]$
- Related nodes are "clustered" if  **$R > G$**  (modularity)

- Necessary and sufficient condition for a random crawler to find pages related to start points



# Link-cluster conjecture

- Stationary hit rate for a random crawler:

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

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

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

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

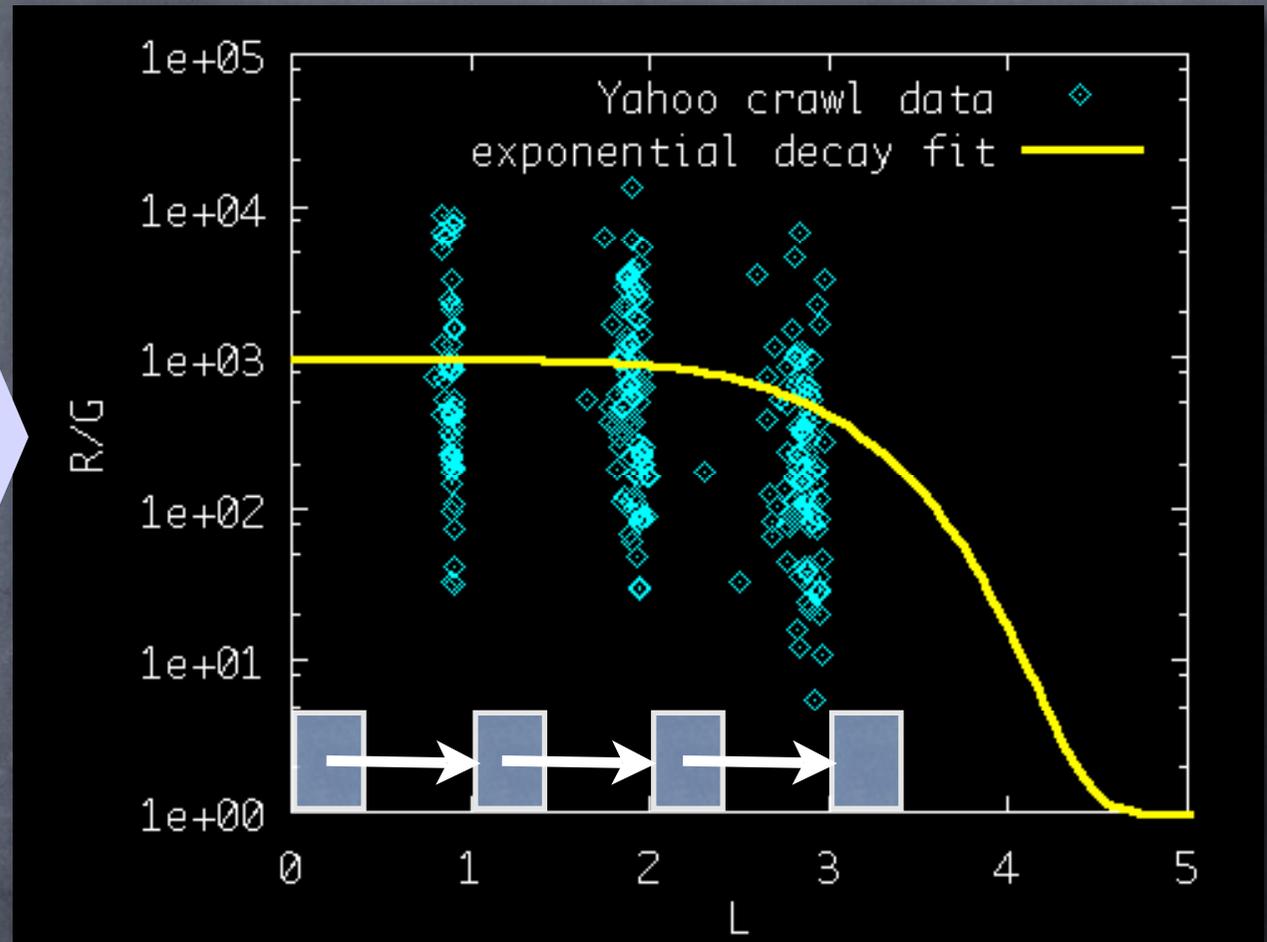
Conjecture

Value added

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

## Link-cluster conjecture

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

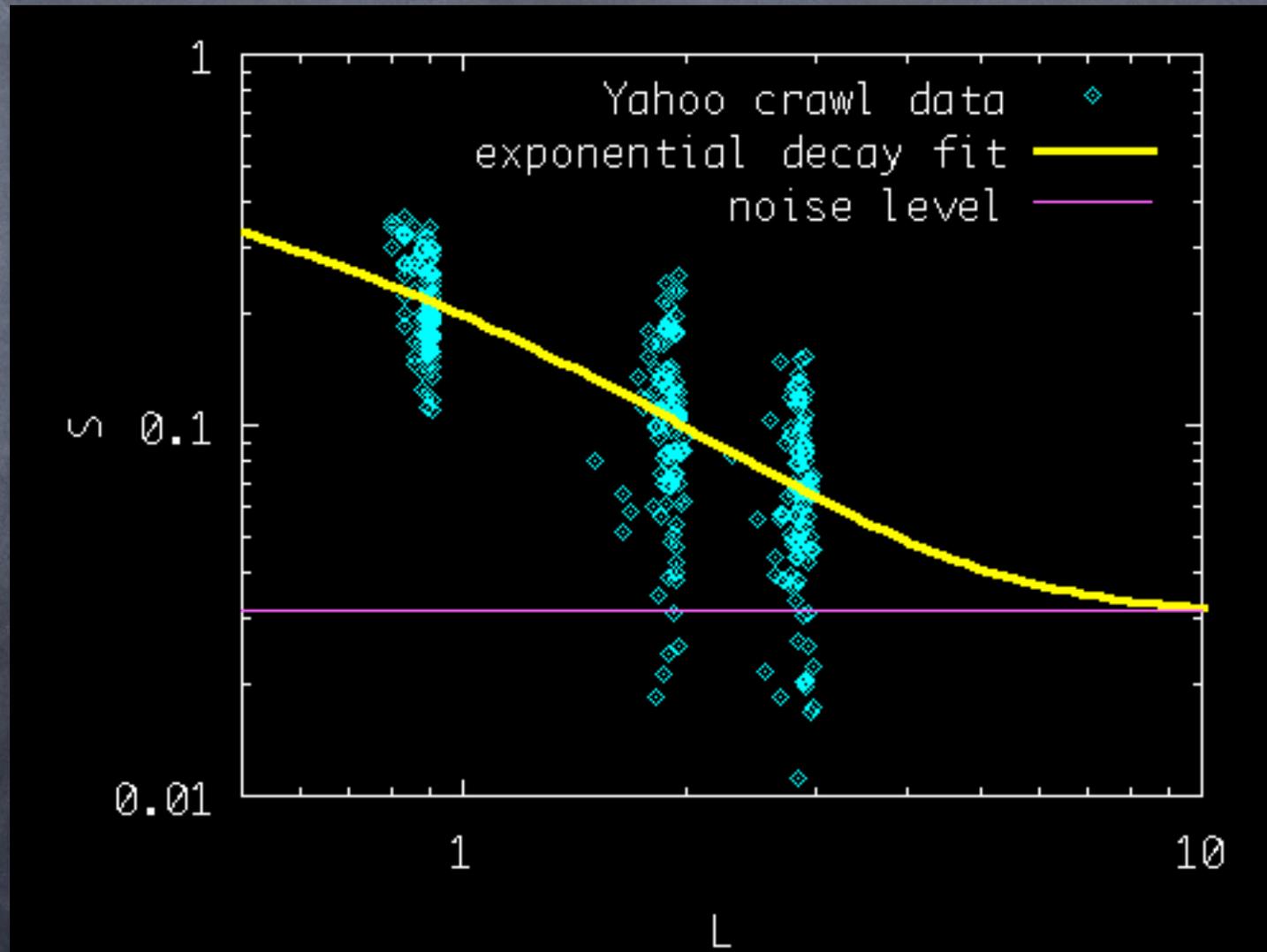


$$L(q, \delta) \equiv \frac{\sum_{\{p: \|path(q, p)\| \leq \delta\}} \|path(q, p)\|}{|\{p: \|path(q, p)\| \leq \delta\}|}$$

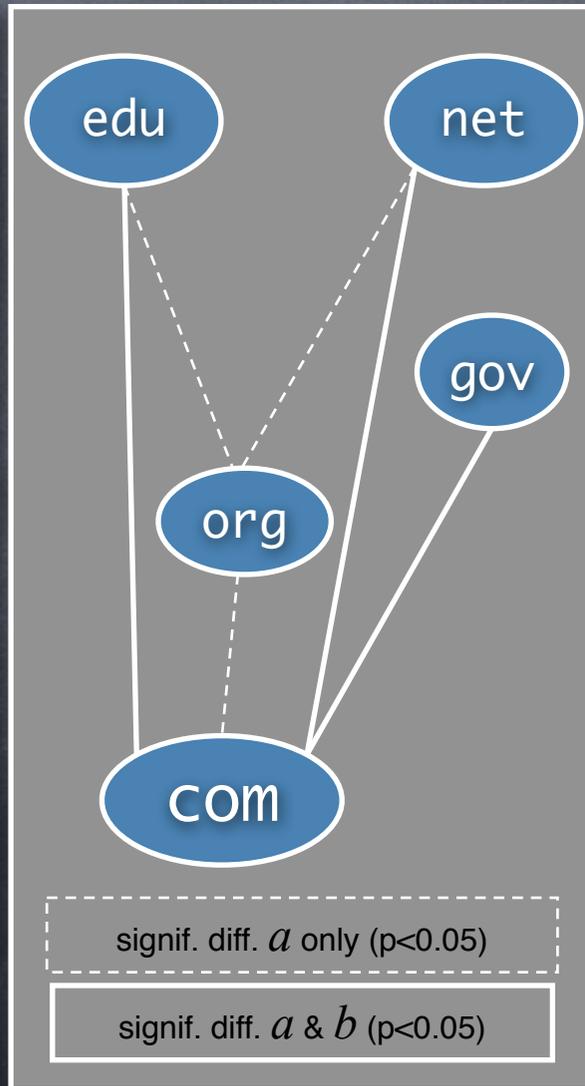
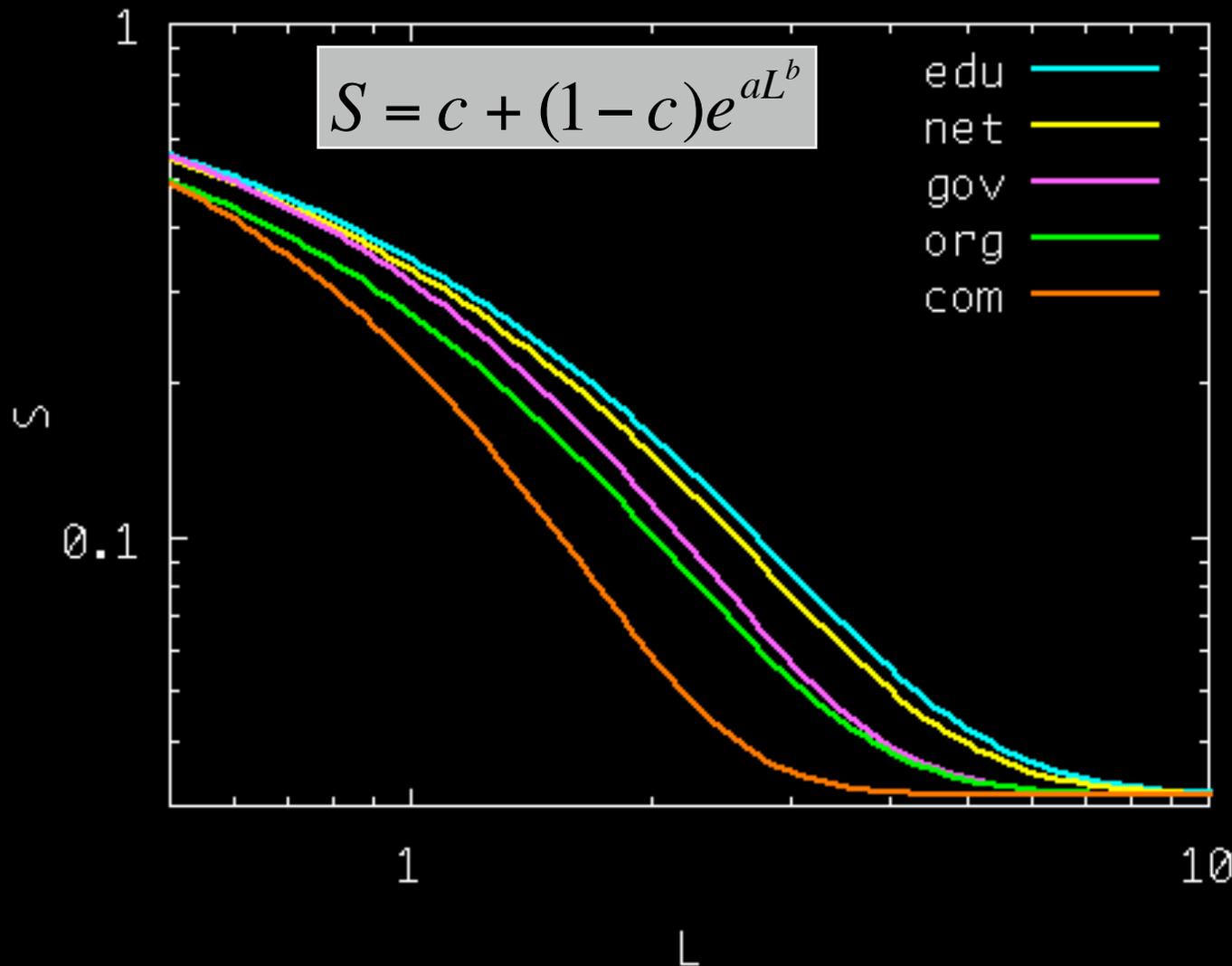
# The “link-content” conjecture

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

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



# Heterogeneity of link-content correlation

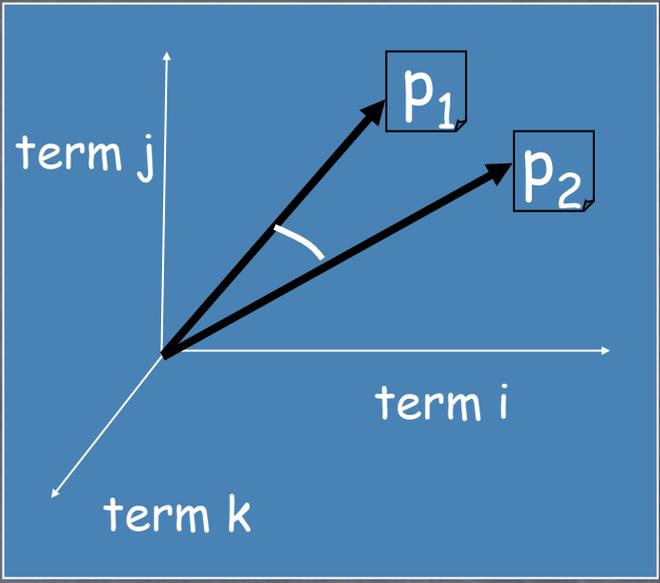


# Mapping the relationship between links, content, and **semantic** topologies

- 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
  - ~  $1.3 \times 10^{12}$  page pairs!

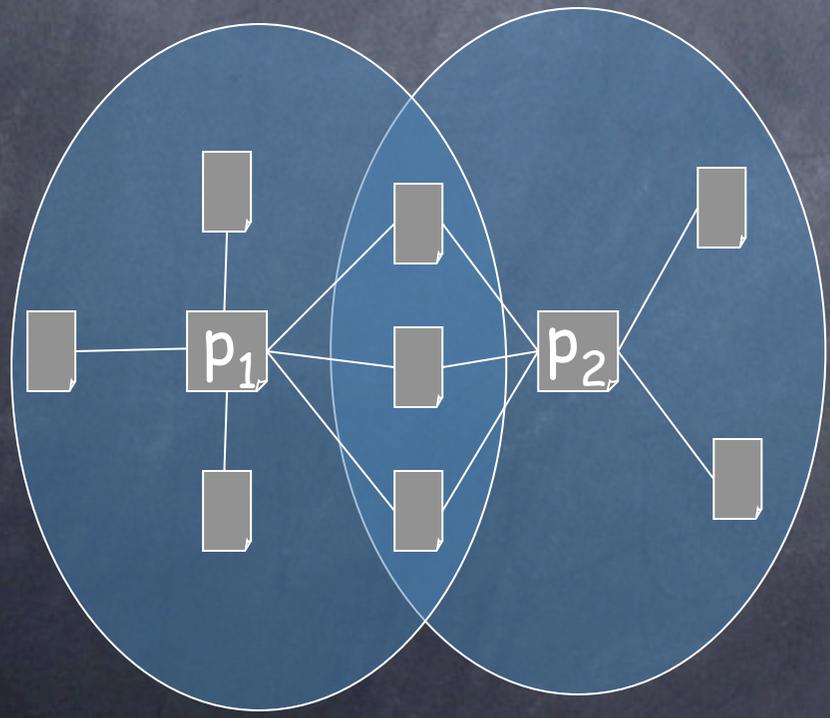
$$\sigma_c(\vec{p}_1, \vec{p}_2) = \frac{\vec{p}_1 \cdot \vec{p}_2}{\|\vec{p}_1\| \cdot \|\vec{p}_2\|}$$

Content similarity

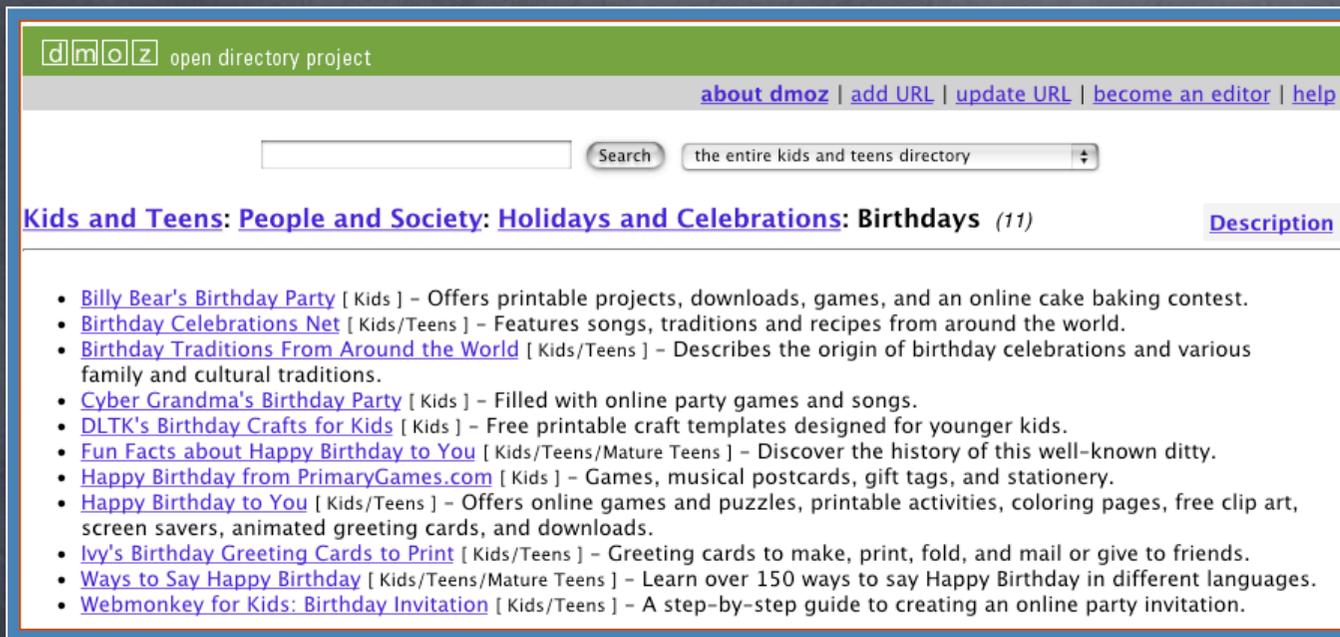


Link similarity

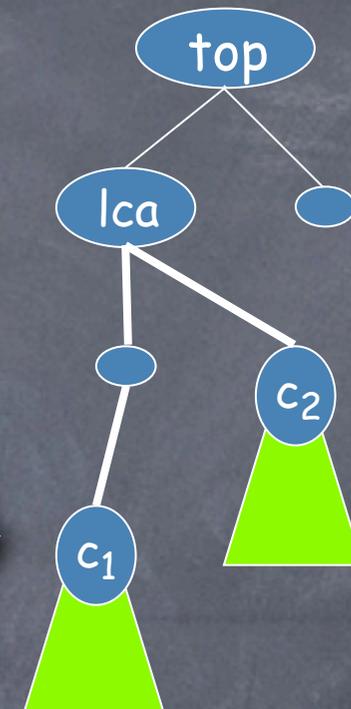
$$\sigma_l(p_1, p_2) = \frac{|U_{p_1} \cap U_{p_2}|}{|U_{p_1} \cup U_{p_2}|}$$



# Semantic similarity



The screenshot shows the DMOZ directory page for "Birthdays". The page title is "Kids and Teens: People and Society: Holidays and Celebrations: Birthdays (11)". The page contains a list of 11 links, each with a brief description of a birthday-related resource. A blue arrow points from the screenshot to the classification tree diagram on the right.

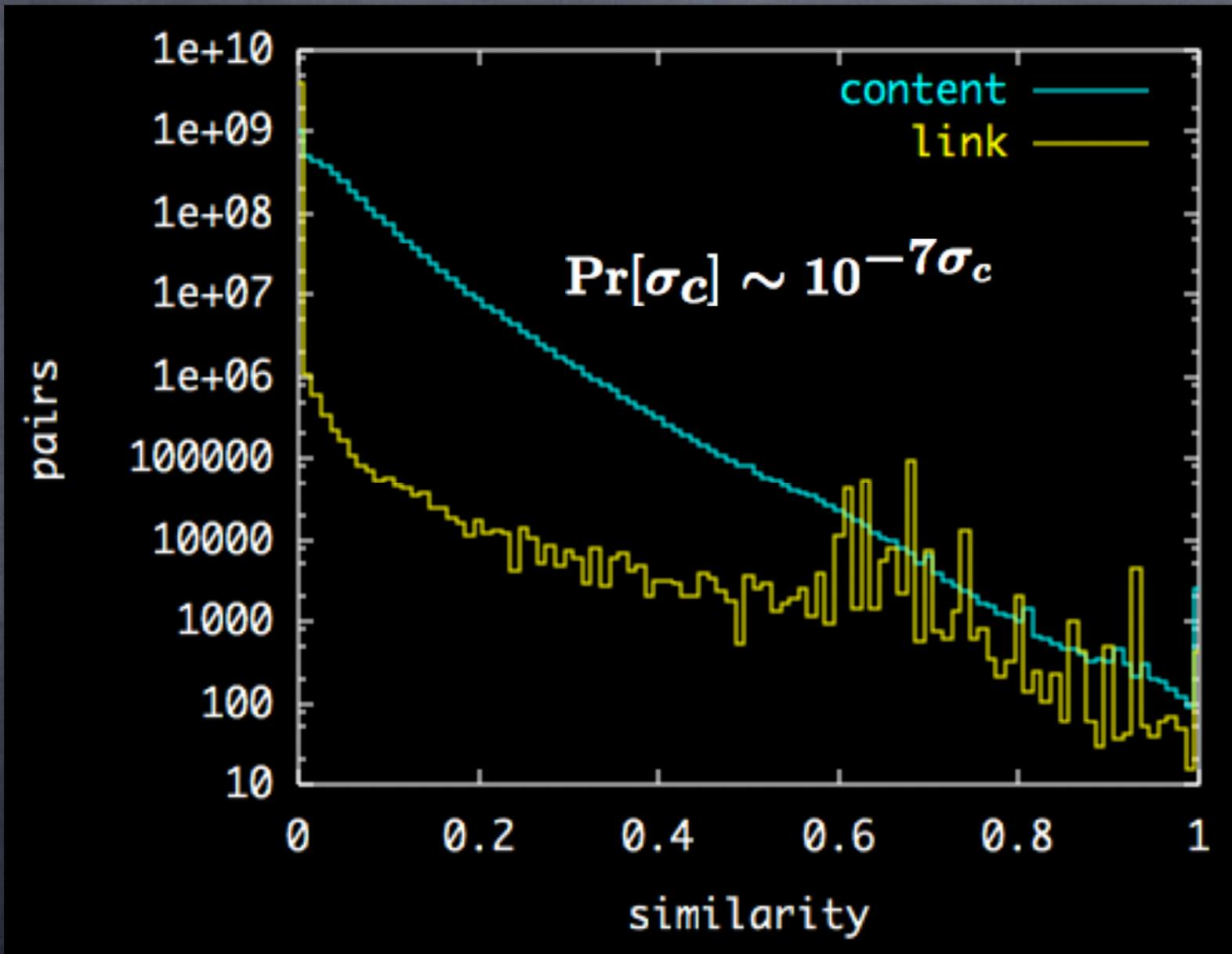


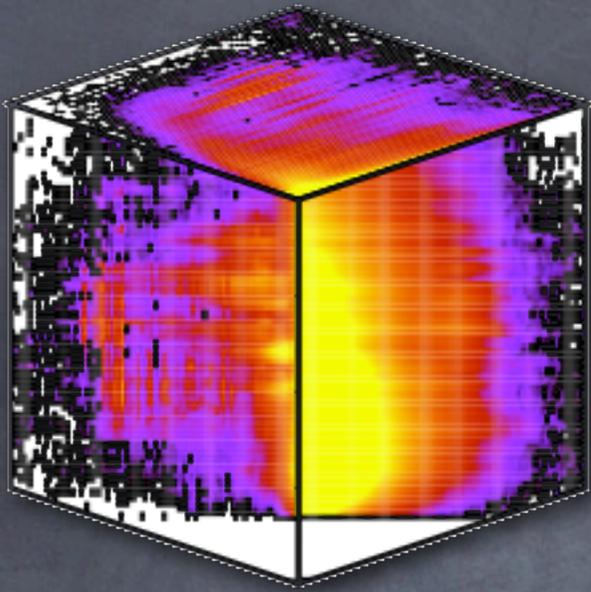
- Information-theoretic measure based on classification tree (Lin 1998)

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

- Classic path distance in special case of balanced tree

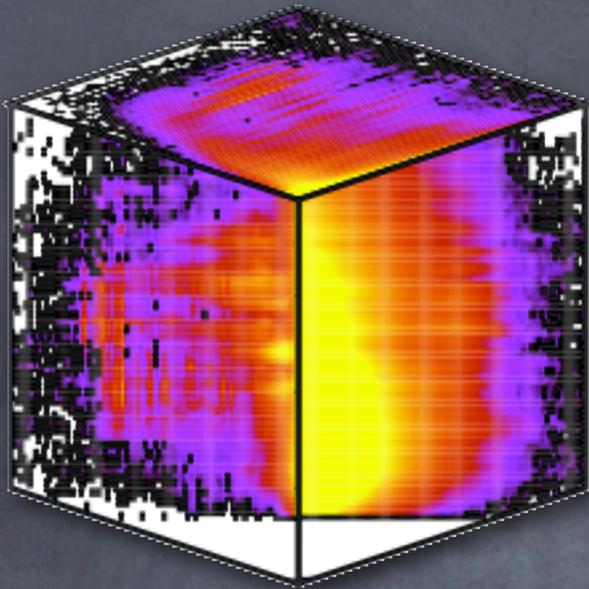
# Individual metric distributions





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

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



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

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

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

Averaging  
semantic  
similarity

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

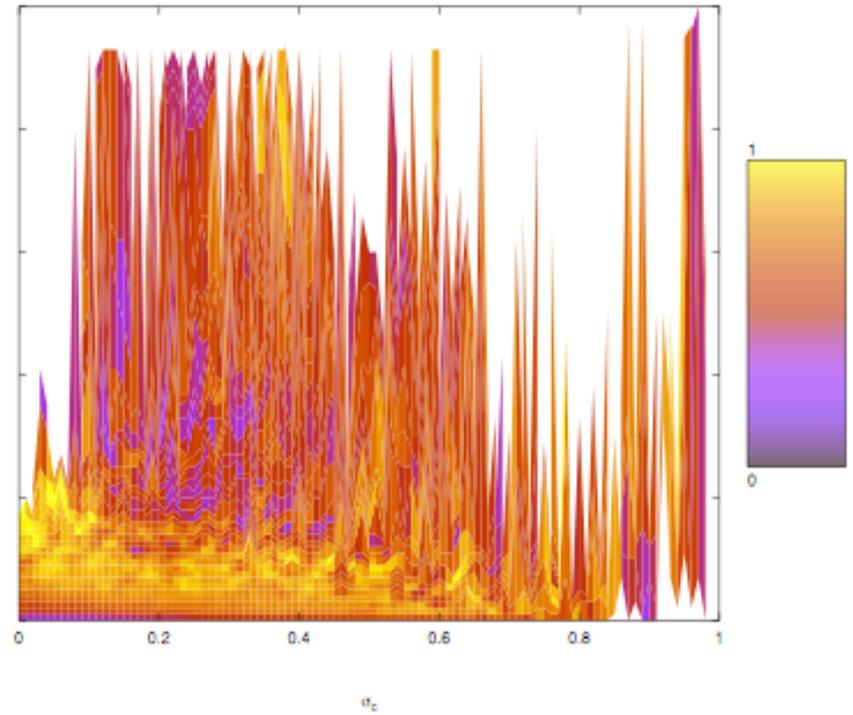
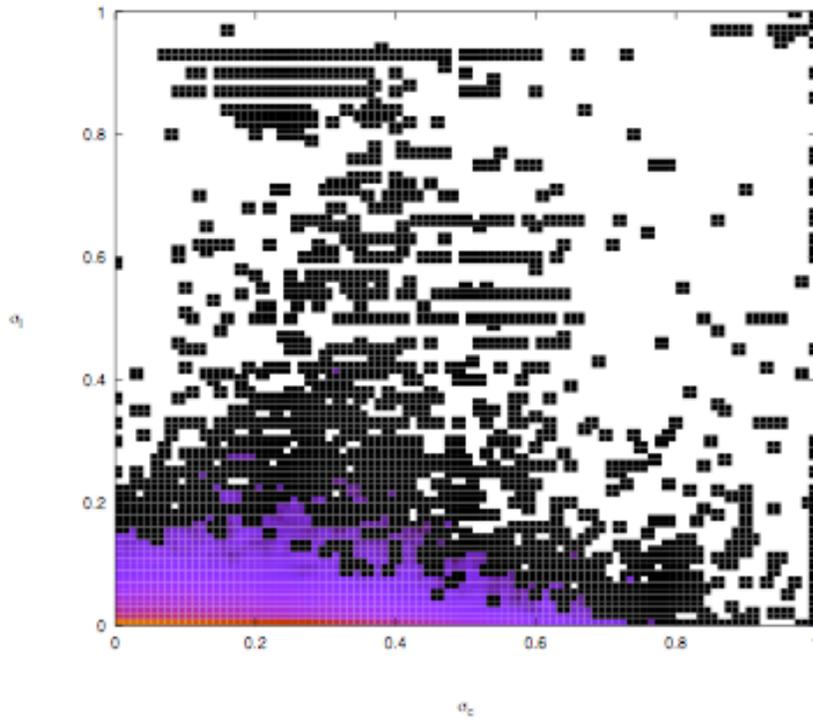
Summing  
semantic  
similarity

# Science

log Recall

Precision

$\sigma_\ell$



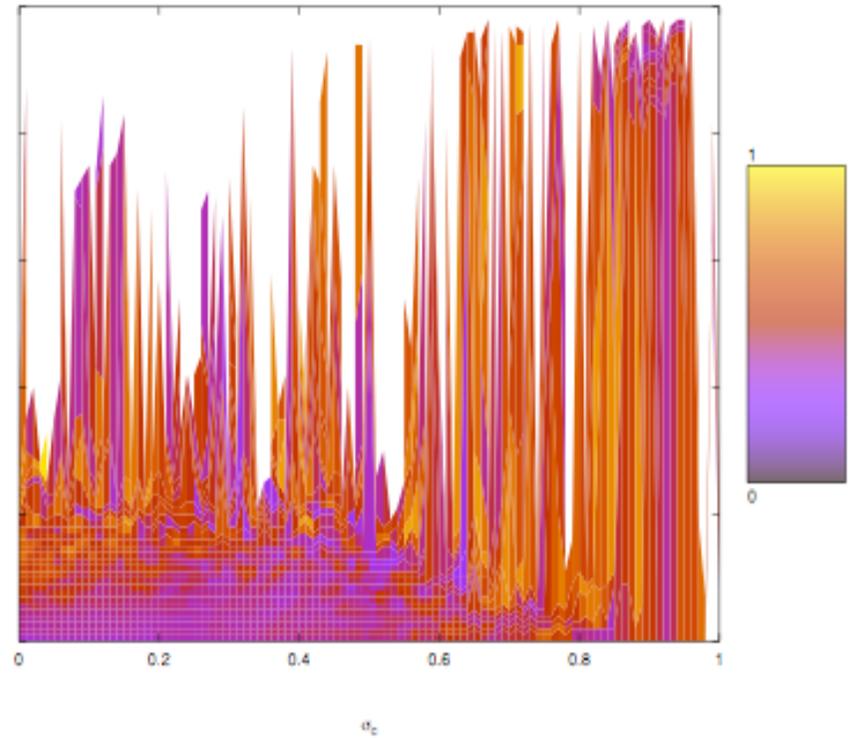
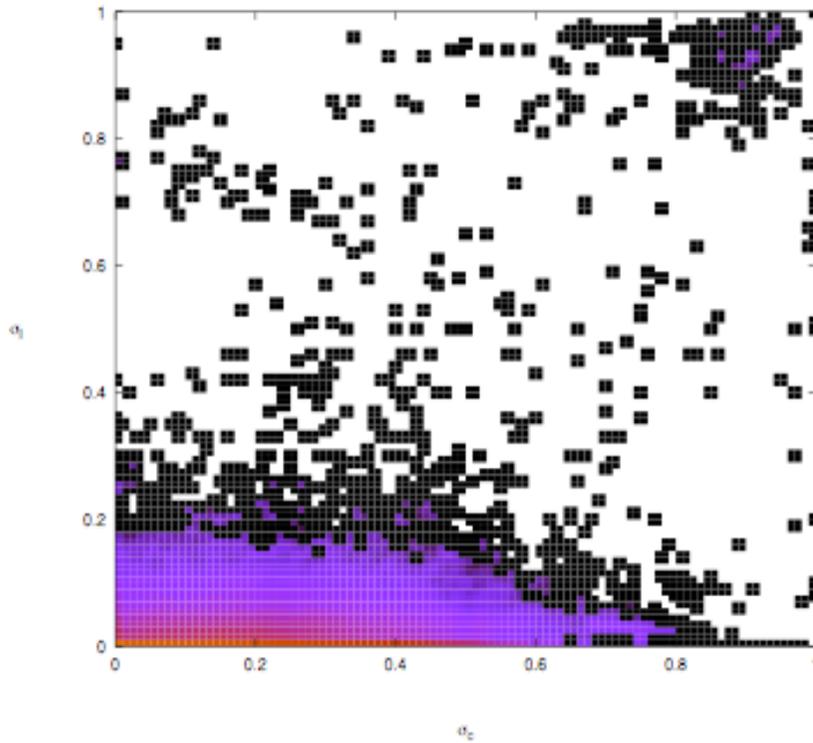
$\sigma_e$

# Adult

log Recall

Precision

$\sigma_\ell$



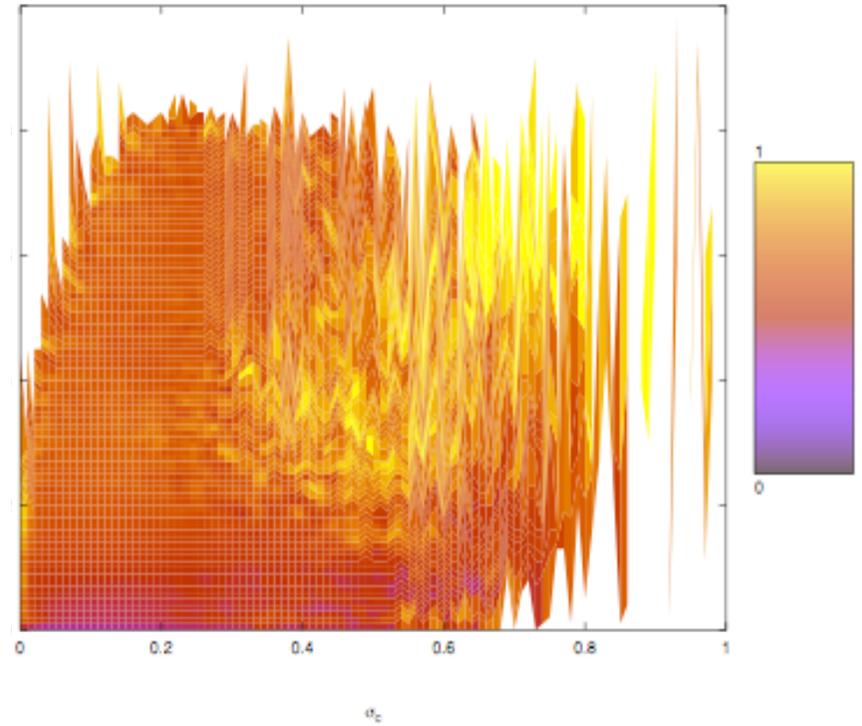
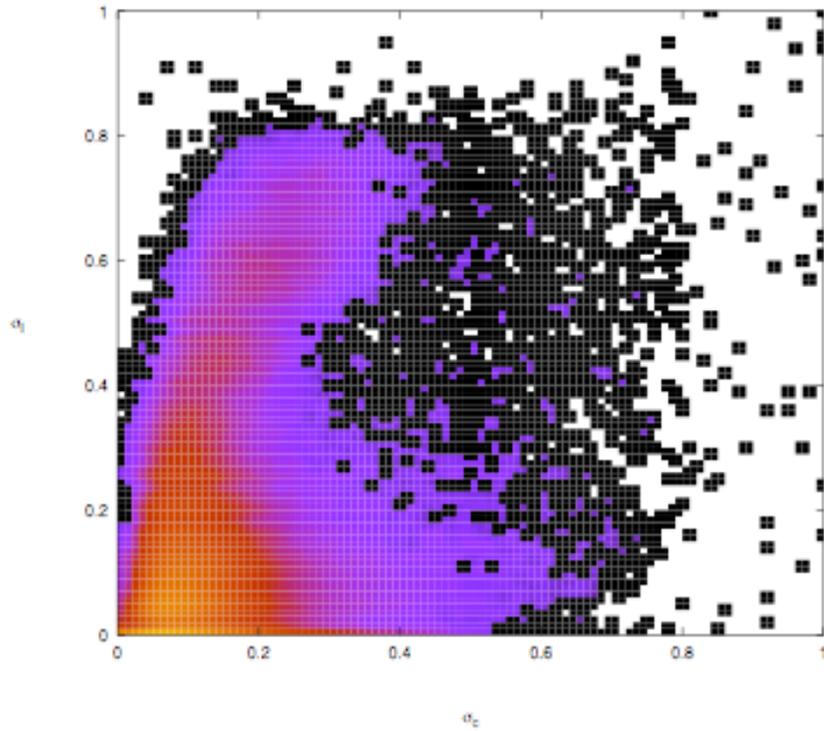
$\sigma_c$

# News

$\sigma_\ell$

log Recall

Precision



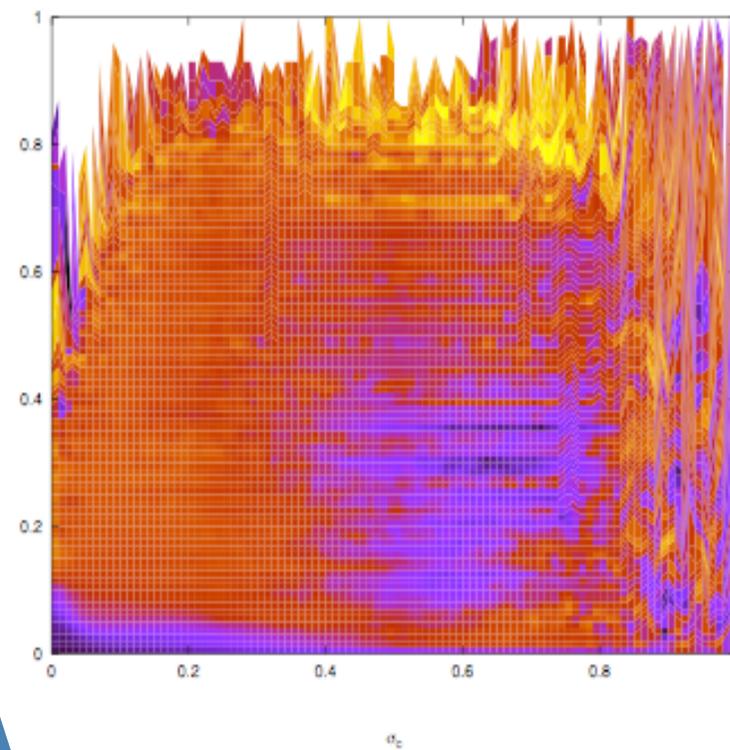
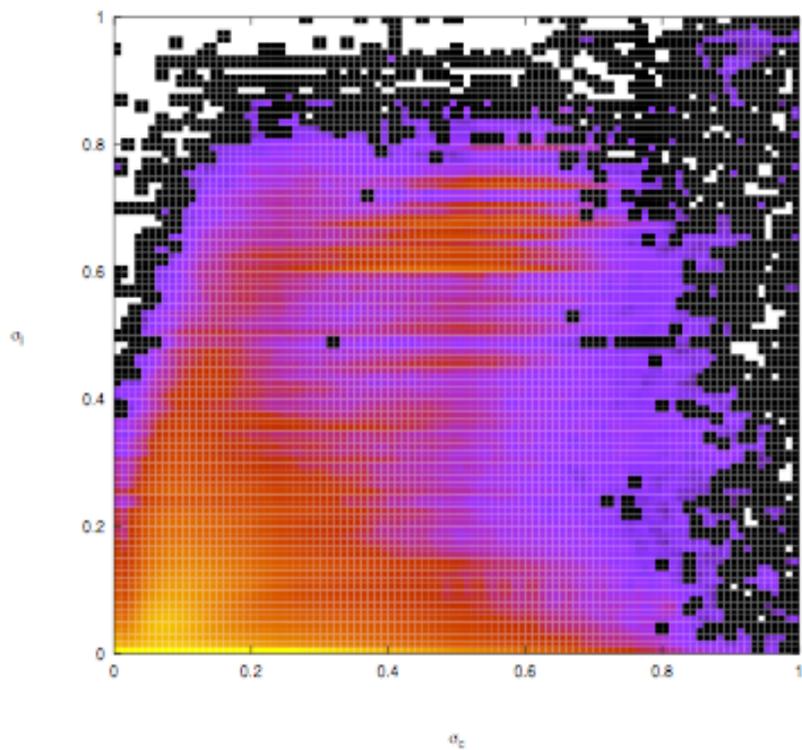
$\sigma_c$

# All pairs

$\sigma_e$

log Recall

Precision



$\sigma_e$

# Outline

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

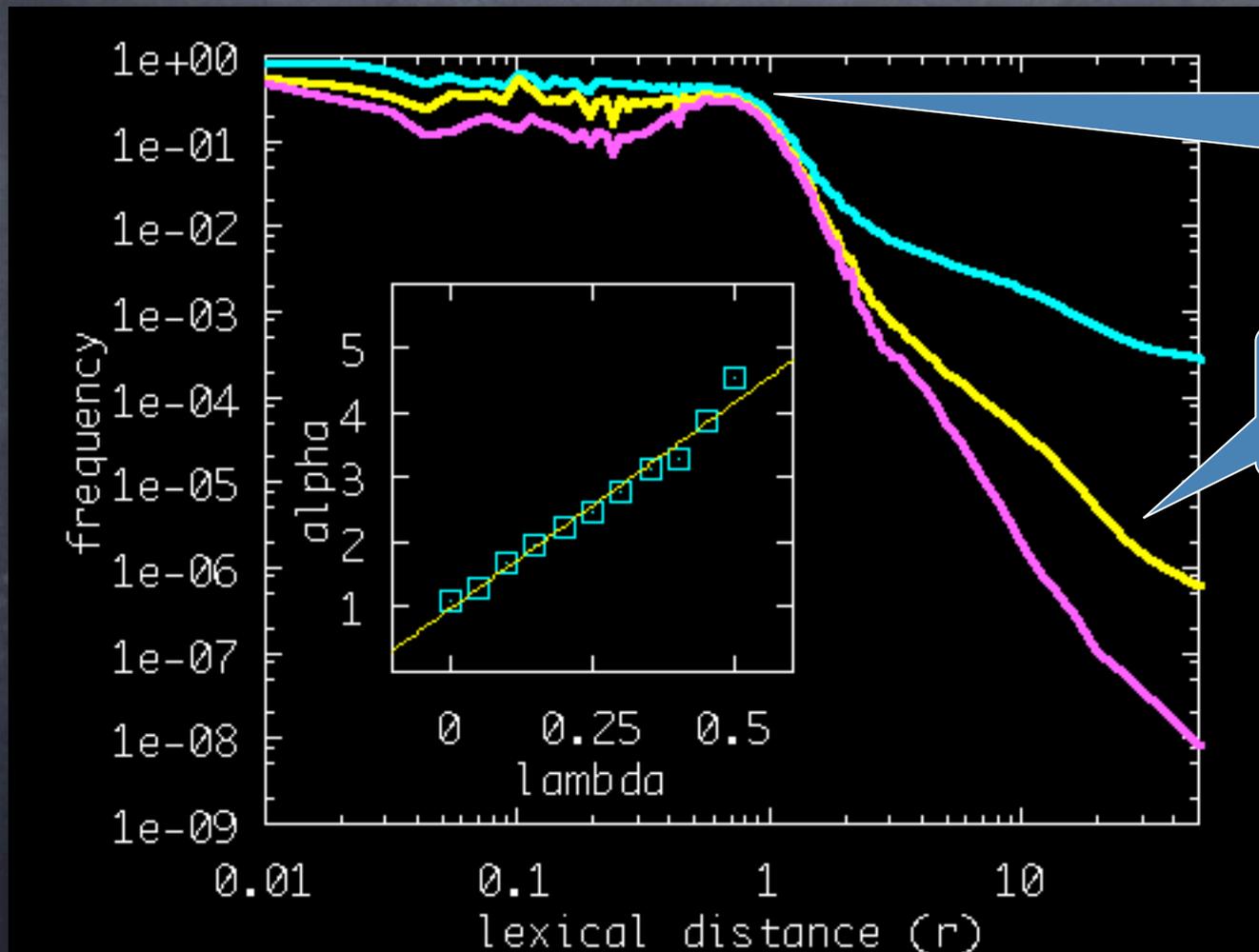
# Link probability vs lexical distance

$$r = 1/\sigma_c - 1$$
$$\Pr(\lambda | \rho) = \frac{|(p,q) : r = \rho \wedge \sigma_l > \lambda|}{|(p,q) : r = \rho|}$$

# Link probability vs lexical distance

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

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



Phase  
transition  
 $\rho^*$

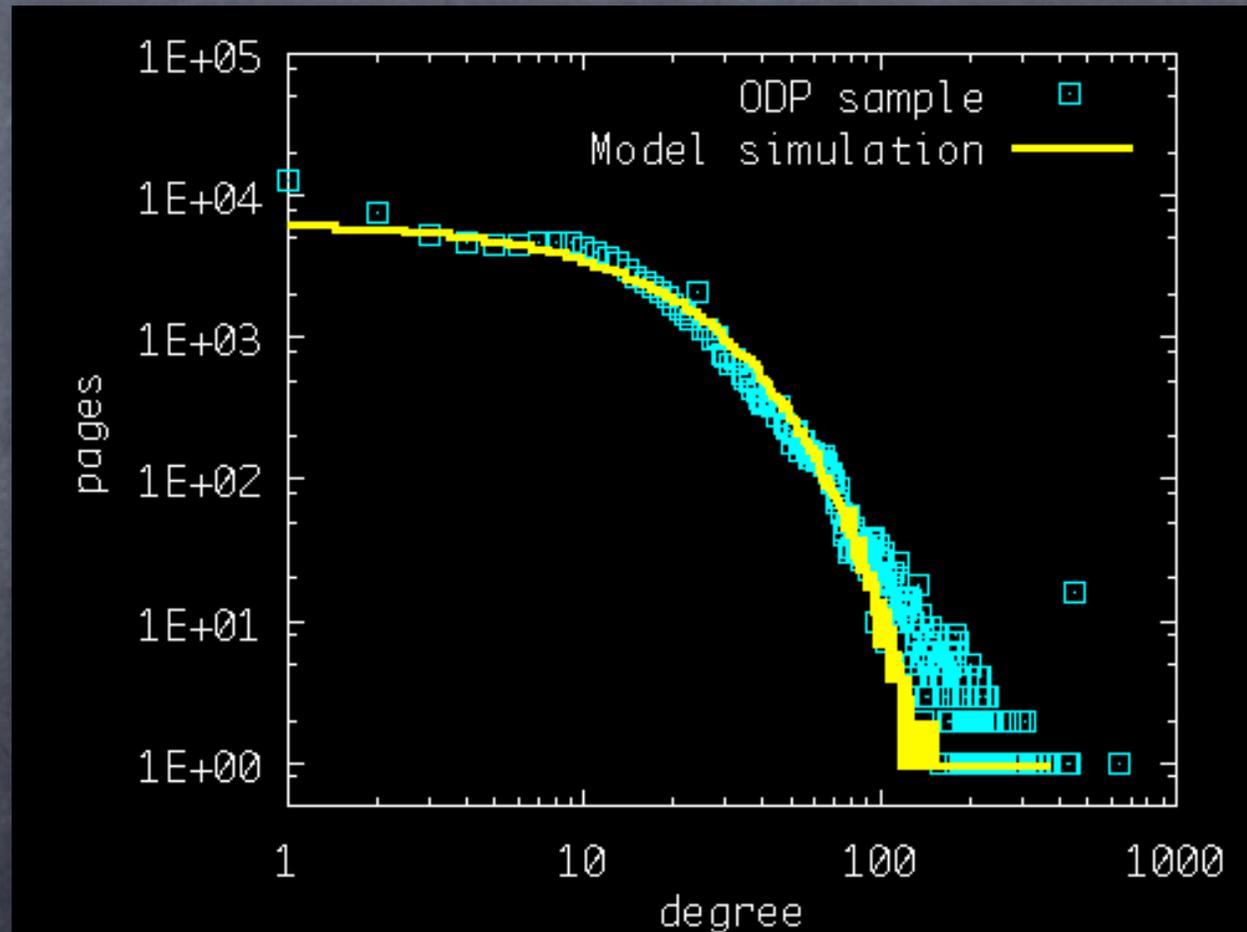
Power law tail  
 $\Pr(\lambda | \rho) \sim \rho^{-\alpha(\lambda)}$

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

# Local content-based growth model

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

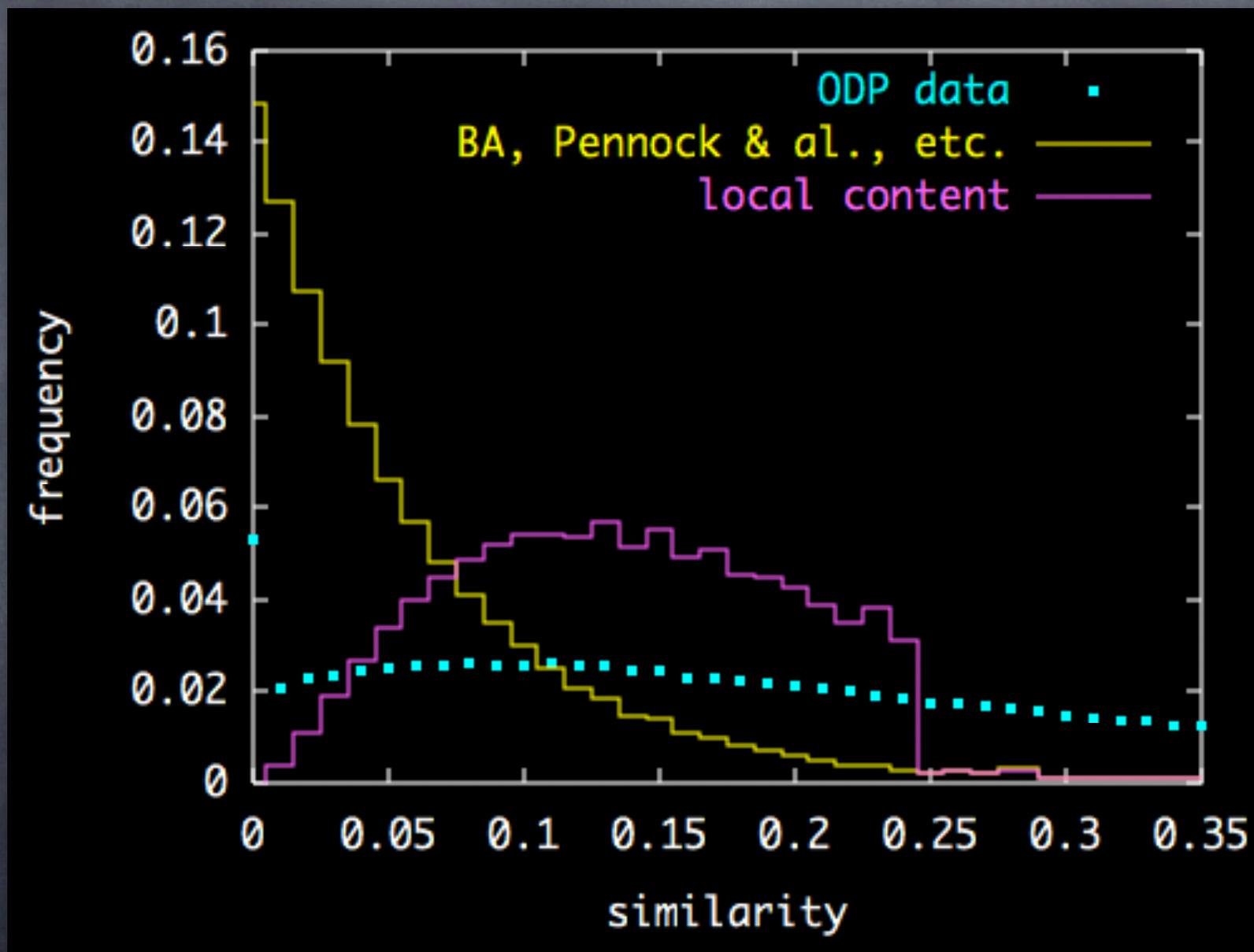
- Similar to preferential attachment (BA)
- Use degree info (popularity/ importance) only for nearby (similar/ related) pages



# So, many models can predict degree distributions...

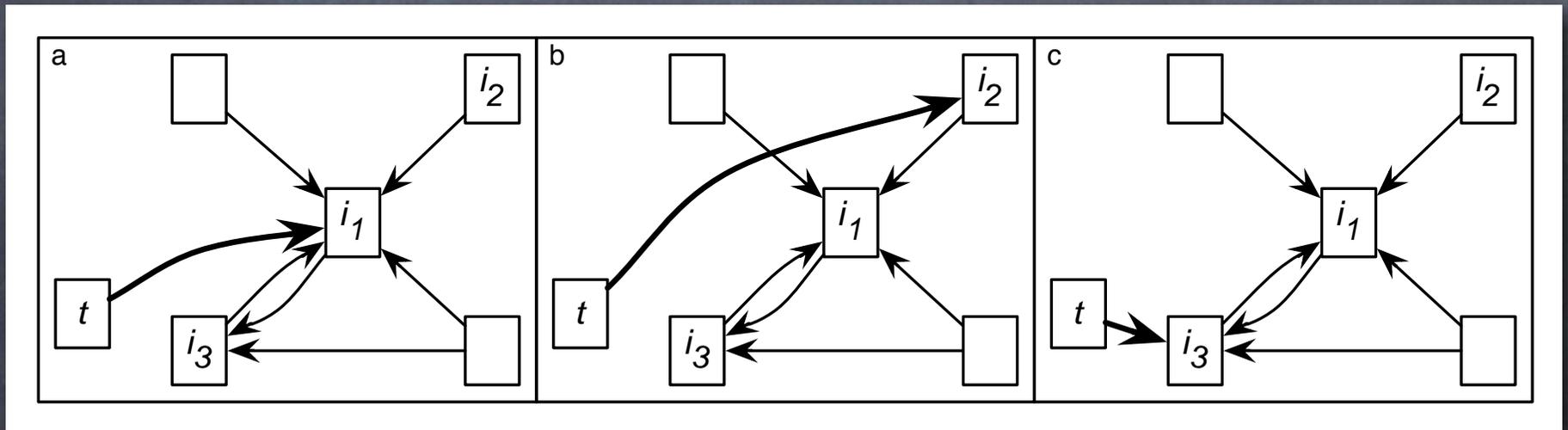
- Which is “right” ?
- Need an independent observation (other than degree) to validate models
- Distribution of content similarity across linked pairs

# None of these models is right!



# The mixture model

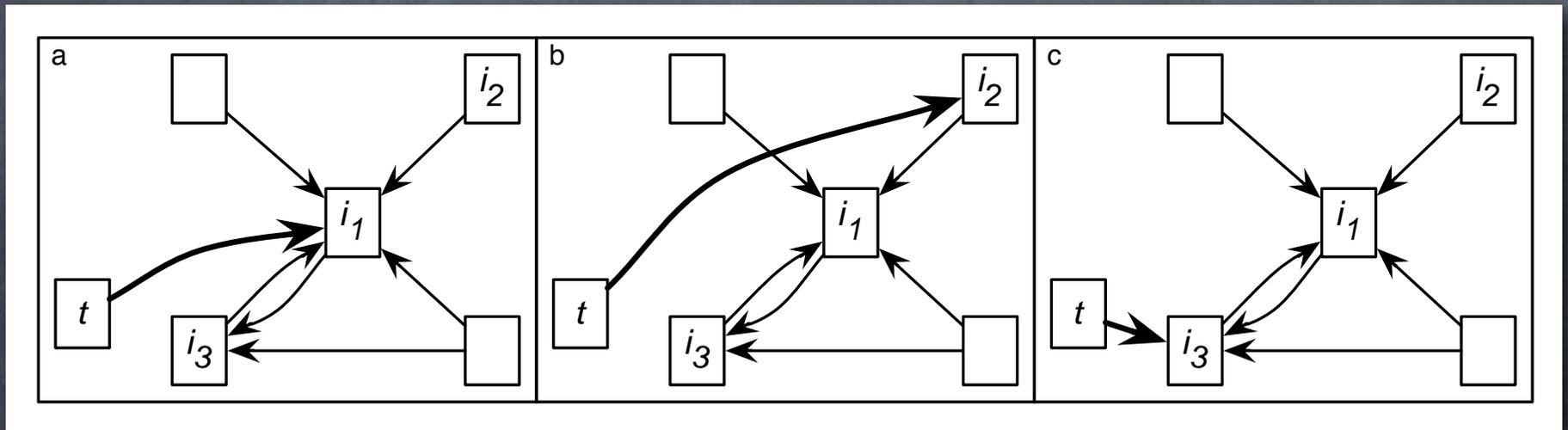
$$\Pr(i) \propto \underbrace{\psi \cdot \frac{1}{t}}_{\text{degree-uniform mixture}} + (1 - \psi) \cdot \frac{k(i)}{mt}$$



# 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

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

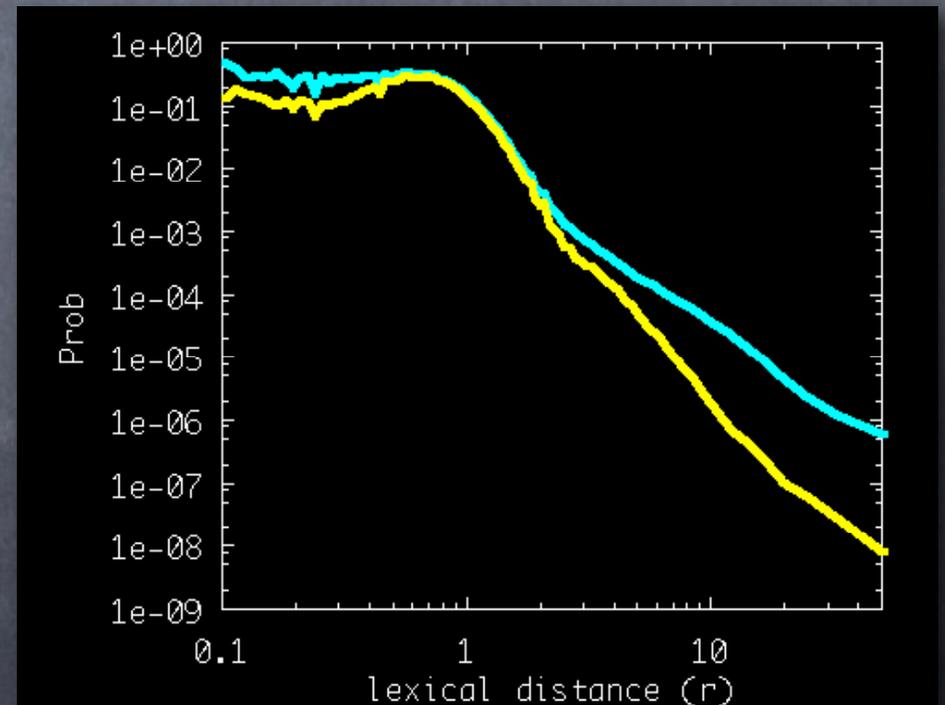
# Degree-similarity mixture model

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

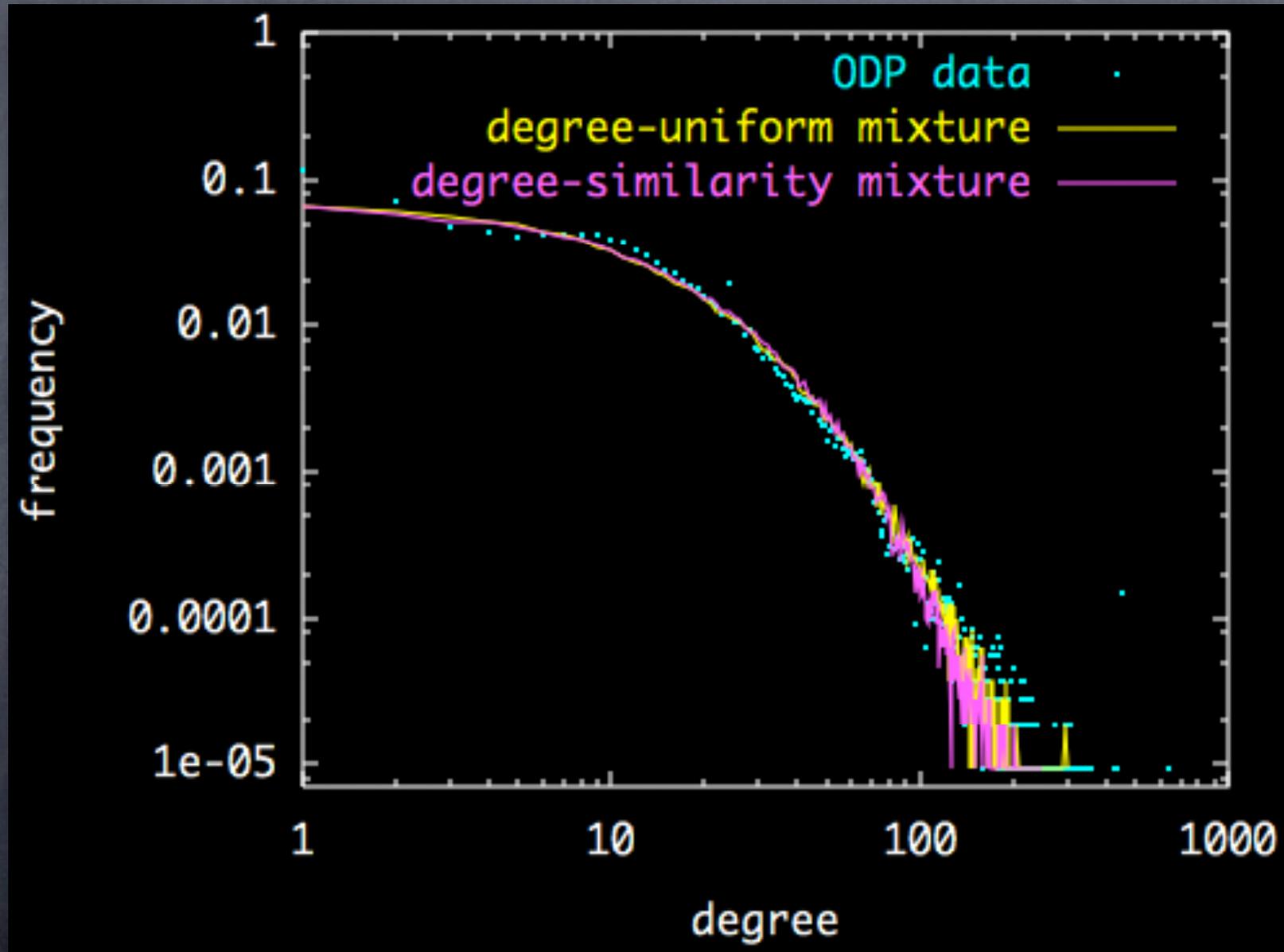


$$\hat{\text{Pr}}(i) \propto [r(i, t)]^{-\alpha}$$

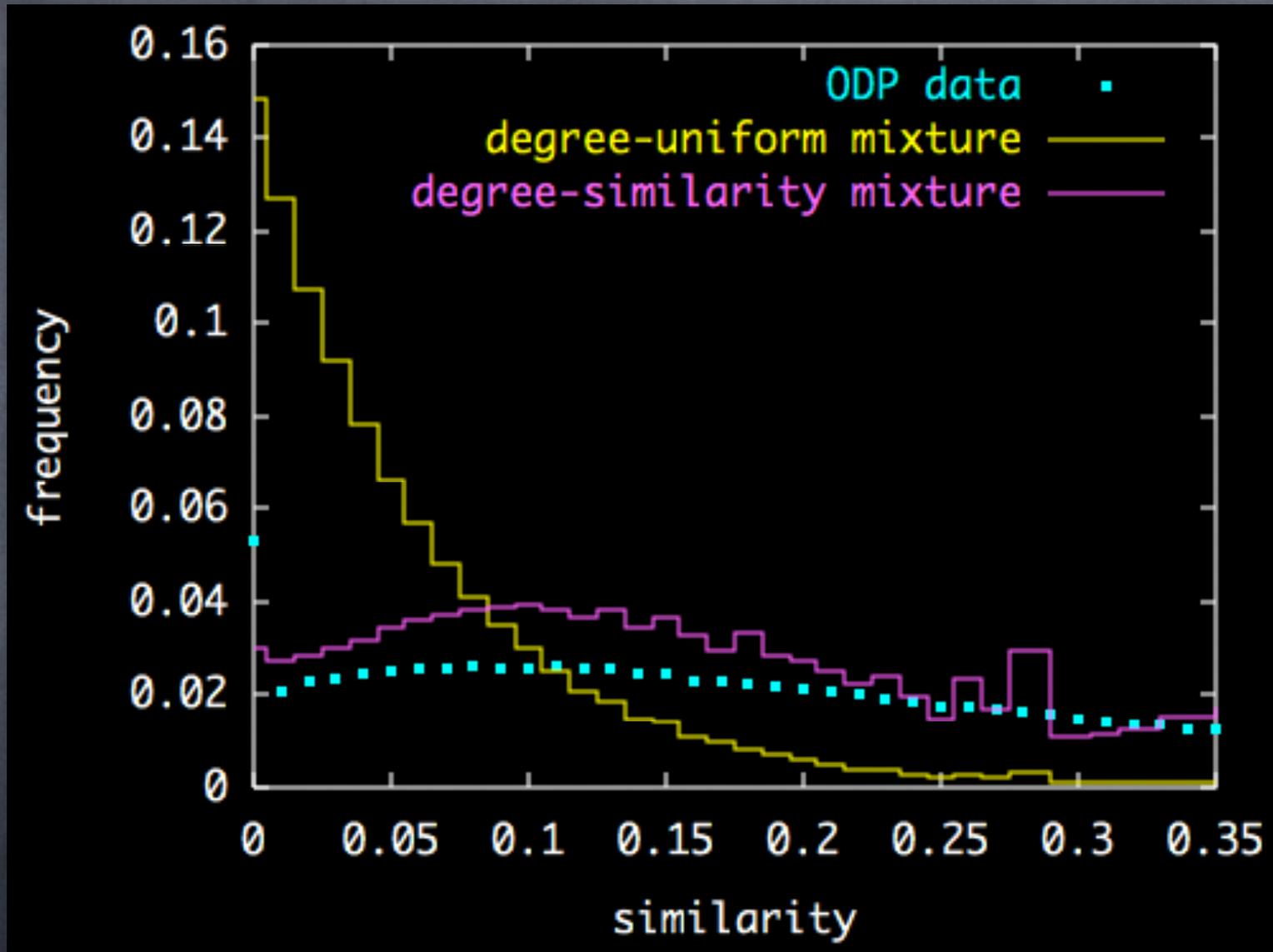
$$\psi = 0.2, \alpha = 1.7$$



Both mixture models get the degree distribution right...

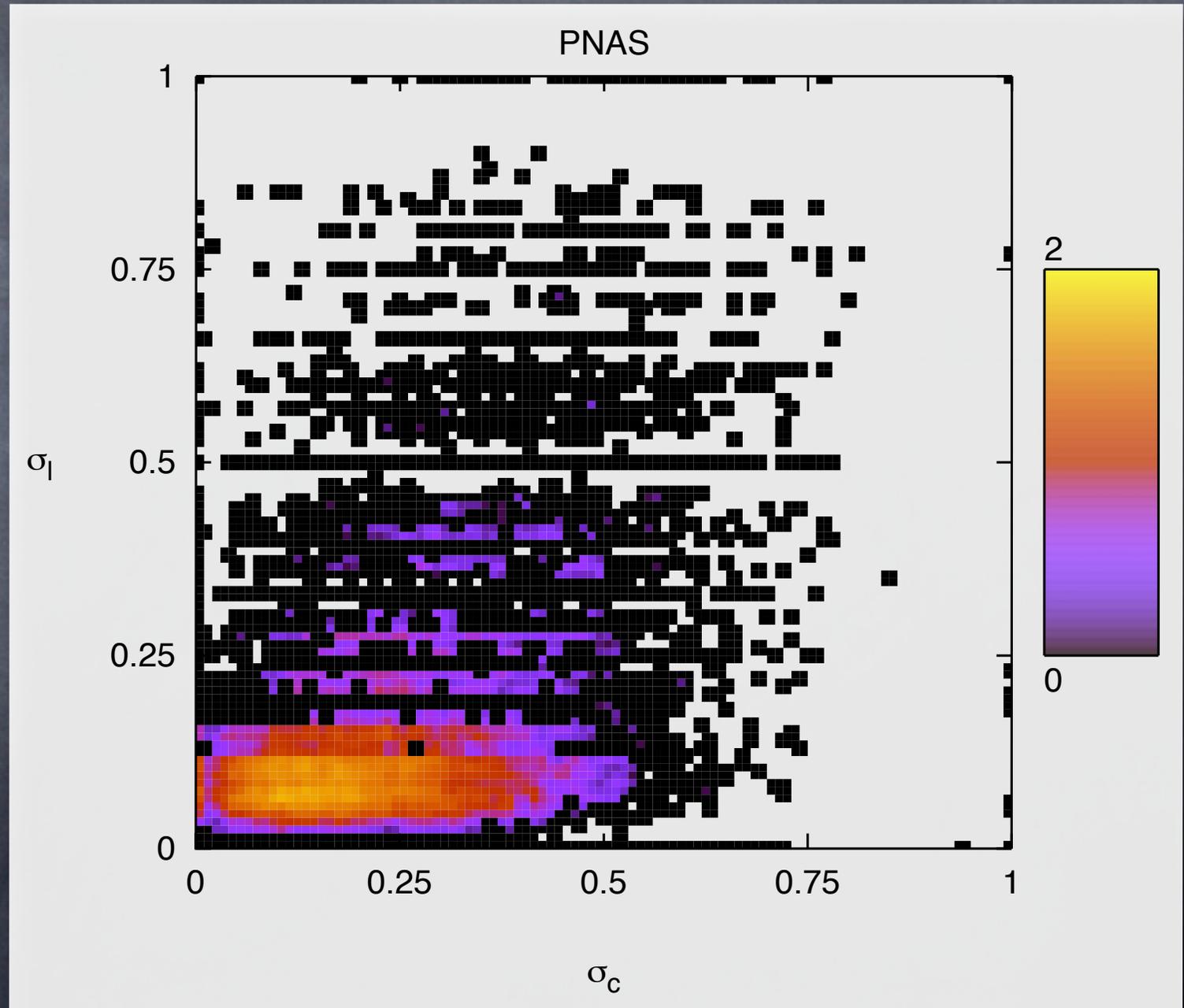


...but the degree-similarity mixture model predicts the similarity distribution better

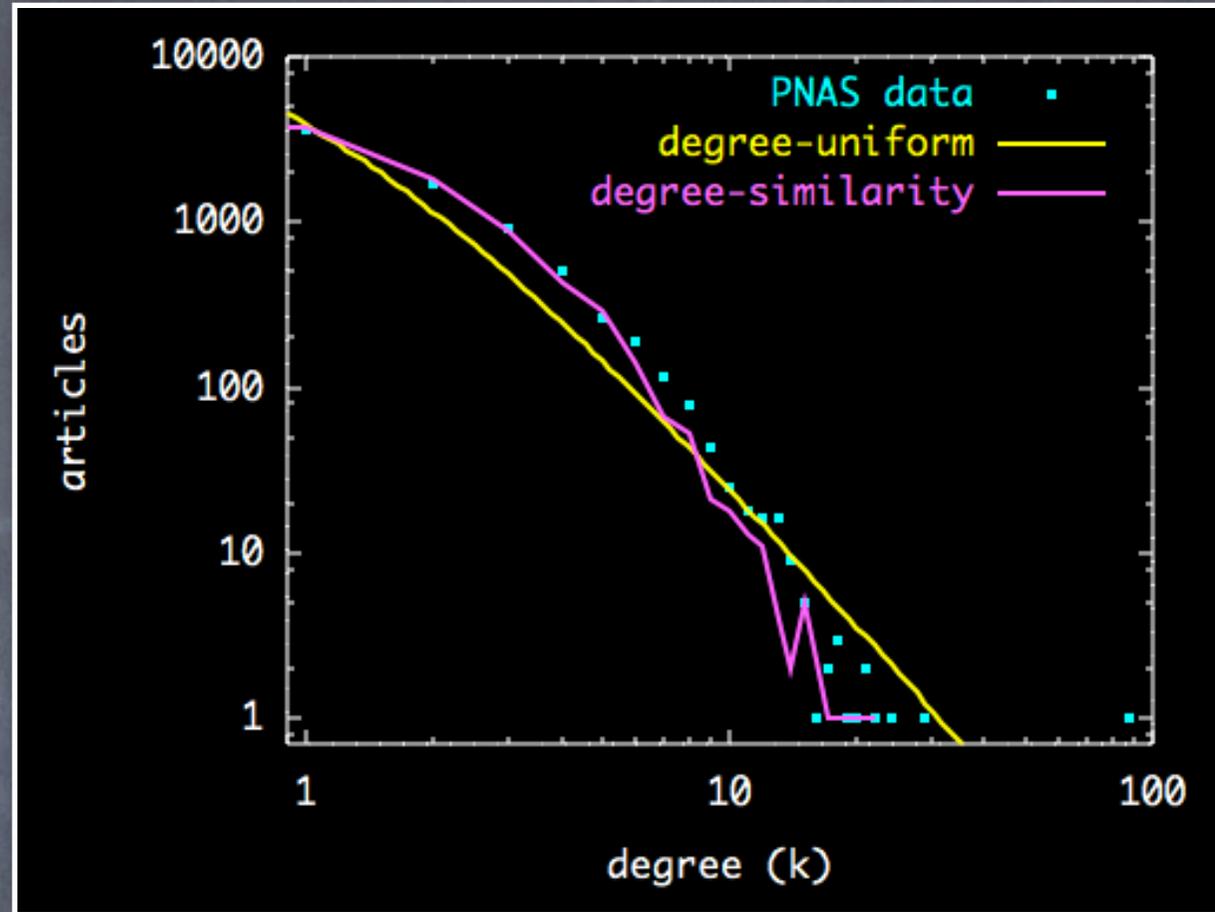


# Citation networks

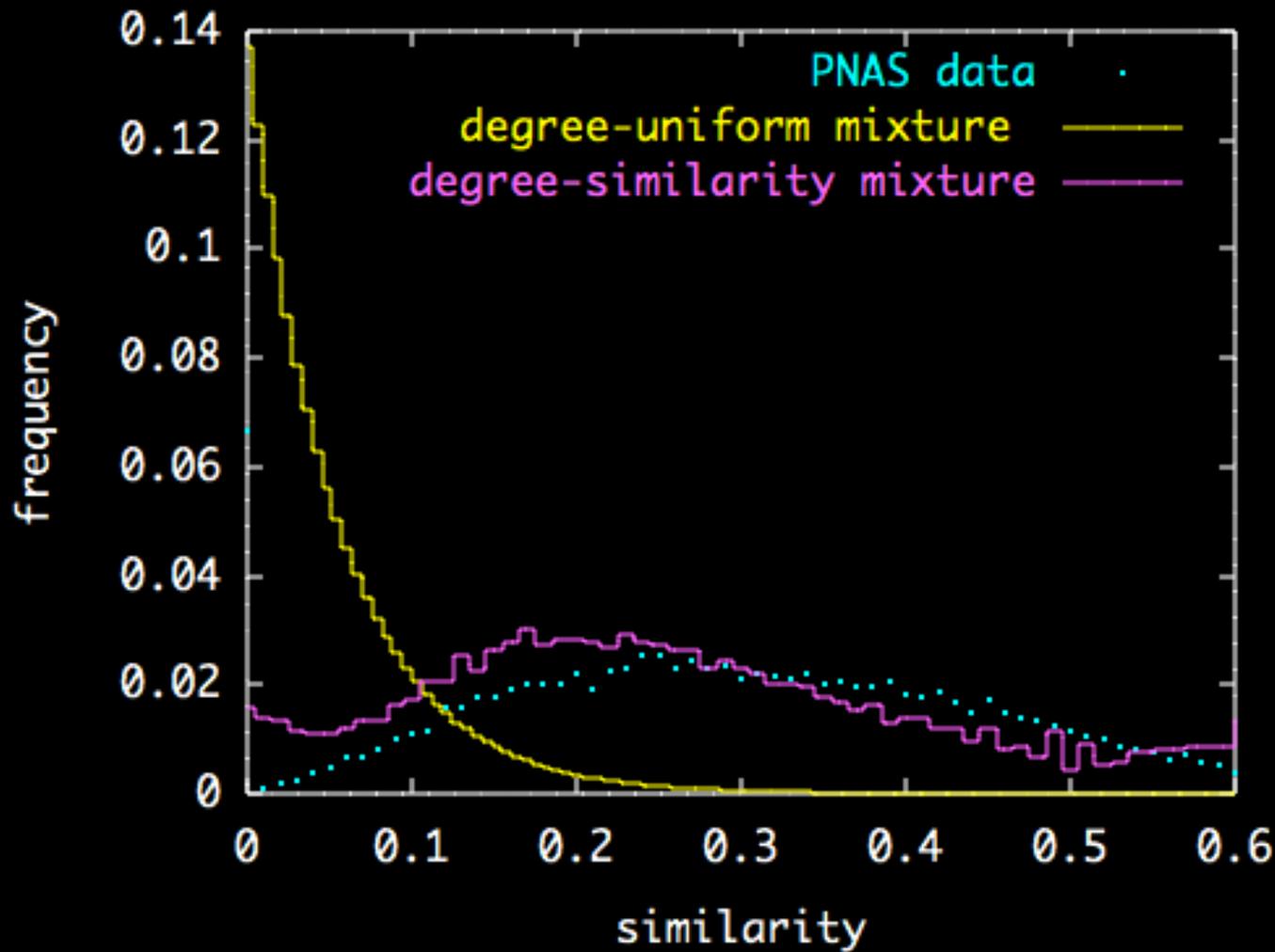
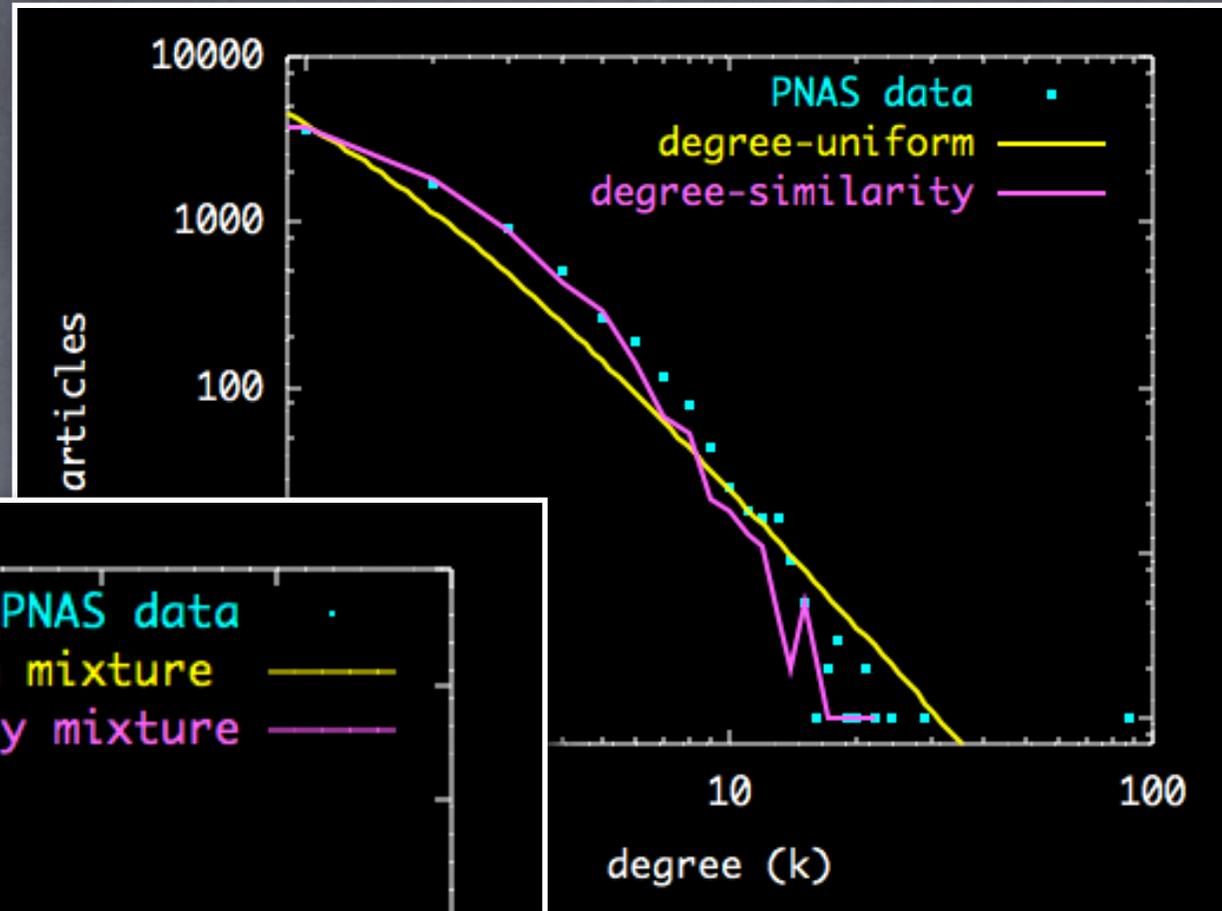
15,785  
articles  
published  
in PNAS  
between  
1997 and  
2002



# Citation networks



# Citation networks



# Open Questions

- ① Understand distribution of content similarity across all pairs of pages
- ① Growth model to explain co-evolution of both link topology and content similarity
- ① 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?

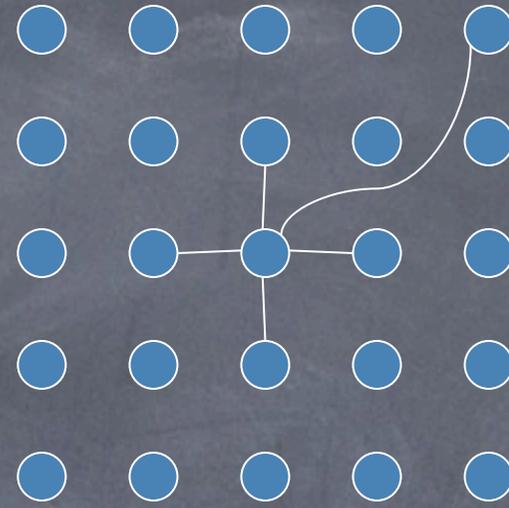
Theory: since the Web is a small world network, or has a scale free degree distribution, **short paths exist** between any two pages:

- ~  $\log N$  (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: ~  $\text{poly}(N)$  (Kleinberg 2000)
- Greedy algorithms based on degree in power law networks: ~  $N$  (Adamic, Huberman & al. 2001)

# Exception # 1



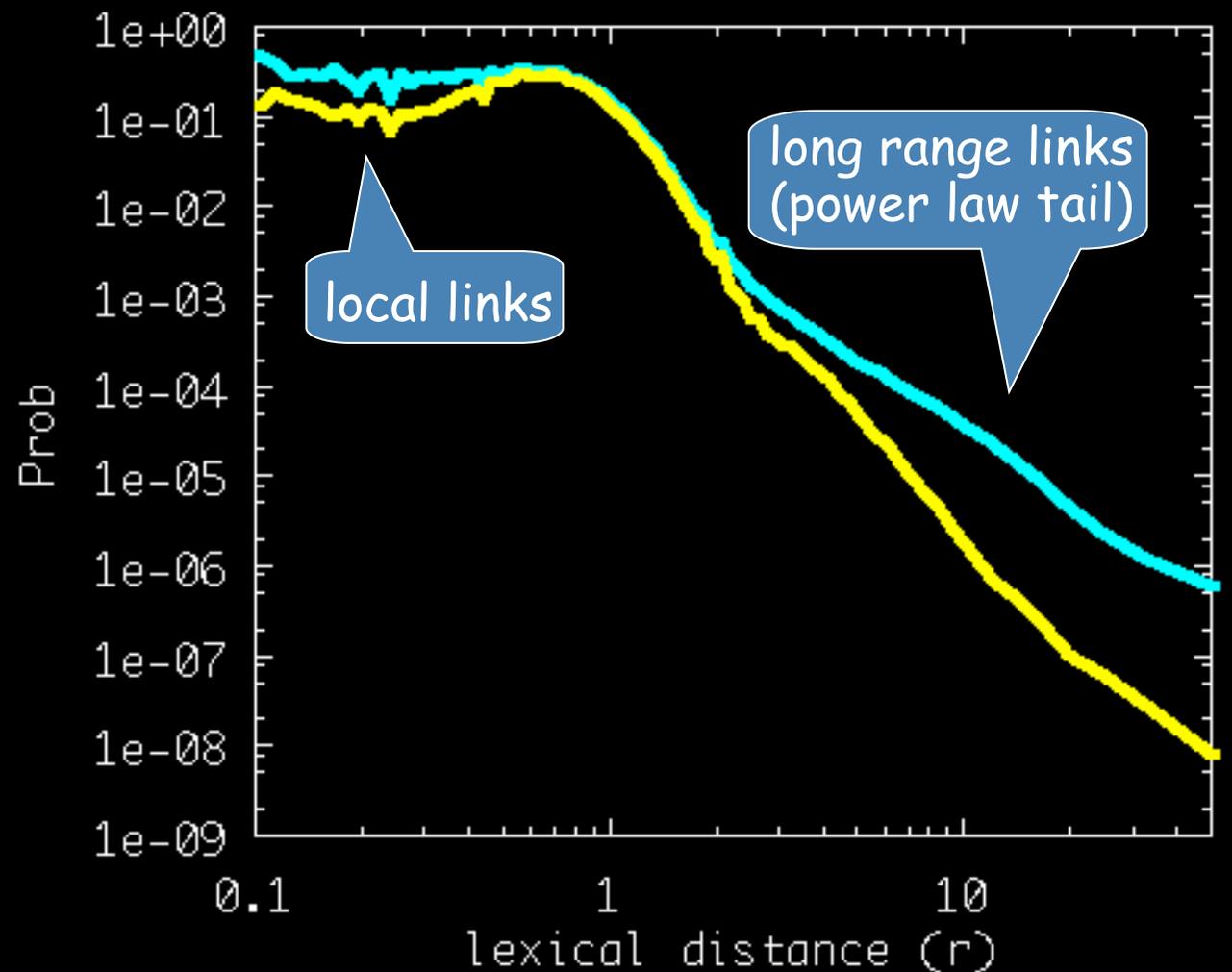
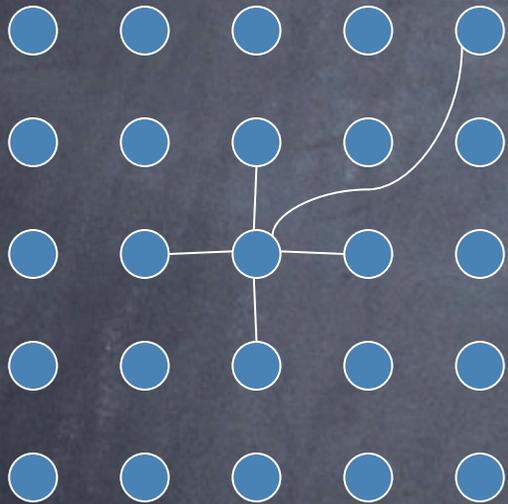
- Geographical networks (Kleinberg 2000)
  - Local links to all lattice neighbors
  - Long-range link probability distribution:  
power law  $P_r \sim r^{-\alpha}$ 
    - $r$ : lattice (Manhattan) distance
    - $\alpha$ : constant clustering exponent

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

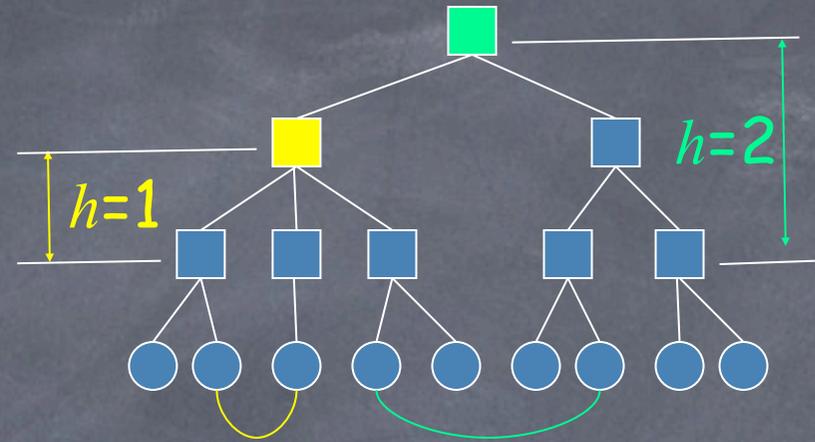
# Is the Web a geographical network?

Replace lattice distance by lexical distance

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



# Exception # 2



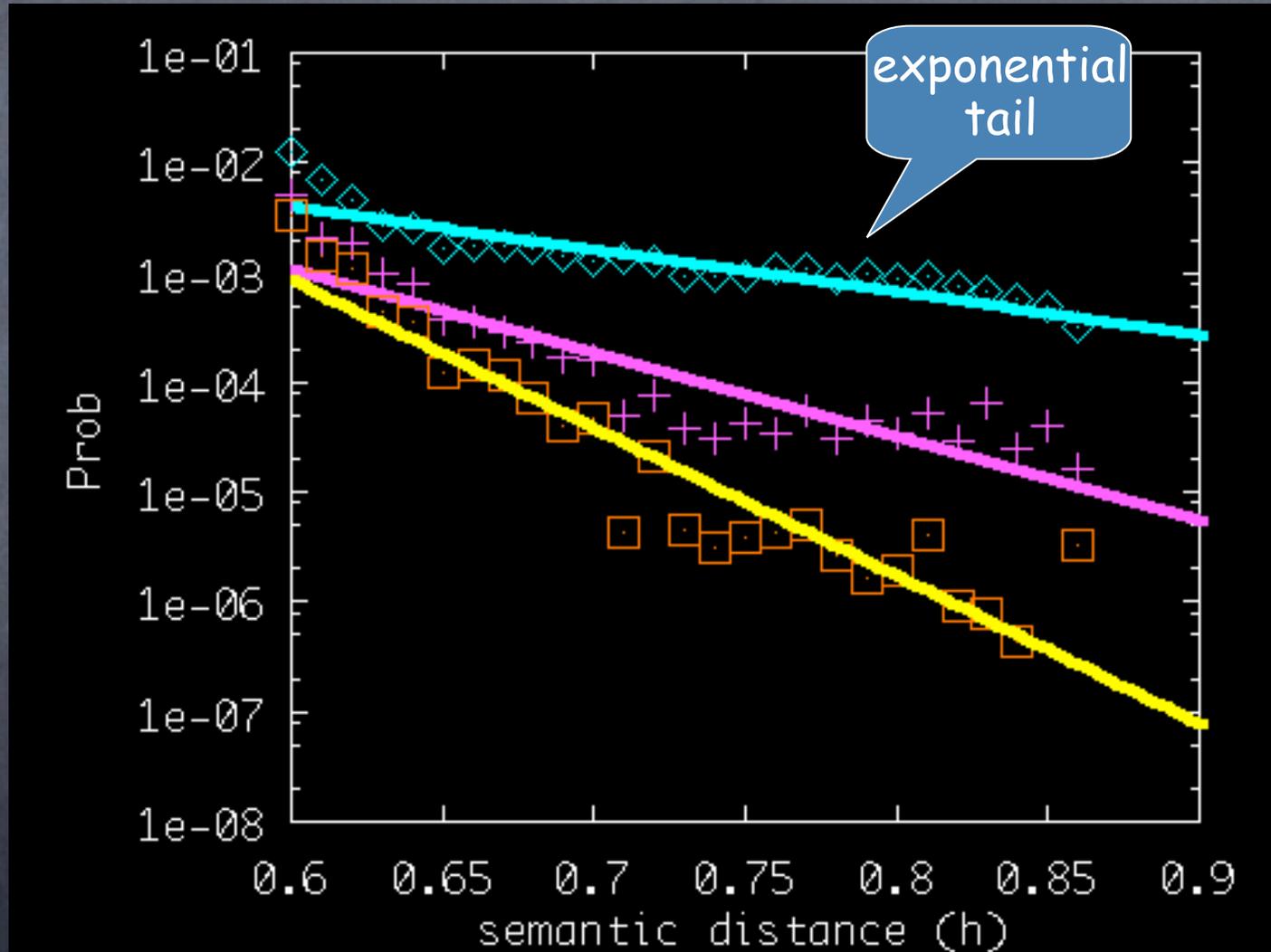
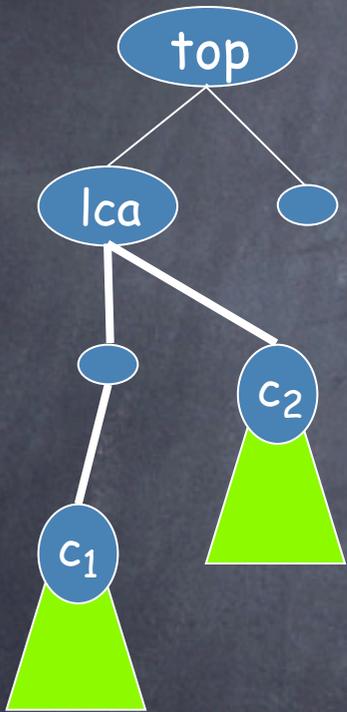
- Hierarchical networks  
(Kleinberg 2002, Watts & al. 2002)
  - Nodes are classified at the leaves of tree
  - Link probability distribution: exponential tail  
 $Pr \sim e^{-h}$ 
    - $h$ : tree distance (height of lowest common ancestor)

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

# 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](http://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.

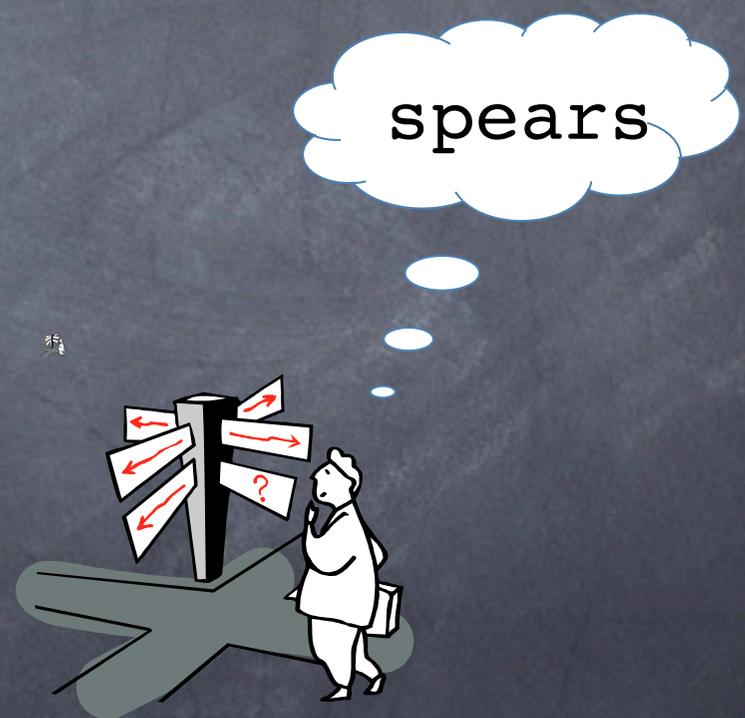
['Bone and Antler Working'](#) - Article.

['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.



spears

# 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

Keywords

Description

Targets

dmoz open directory project

Home: [Cooking](#): [Baking and Confections](#): [Cookies](#): [Chocolate Chip](#) (6)

- [The Big Chocolate Chip Cookie Page](#) - Devoted to the chocolate chip cookie.
- [Chocolate Chip Cookies](#) - Various recipes for cookies with morsels of chocolate.
- [Chocolate Chip Cookies from Allrecipes](#) - Include regular, nuts, white chocolate.
- [In the Chips](#) - Cookies, cakes, candy, muffins, etc. using chocolate chips.

Copyright © 1998-2001 Netscape

[Terms of Use](#)

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

Recipe and Tips R

A chocolate chip cookie recipe that uses Karo syrup in it.  
A recipe using metric measurements?  
Any recipes that don't use eggs?  
A chocolate chip cookie recipe that you bake in mini muffin pans w  
How does one avoid dry, 'cakey' cookies?  
Any recipes for Chocolate Chip Coolie Pies?  
The recipe for chocolate chip cookies in a jar. All the dry ingredien

Recipe	Rating
<b>Absolutely Excellent Oatmeal Cookies</b> Submitted by: <b>Marylou</b>	★★★★★ 46 Ratings 25 Reviews
These are chewy, healthy oatmeal cookies which can be prepared in a number of variations, just add nuts, raisins, chocolate chips, coconut, candied fruit or any other additions.	
<b>Absolutely Sinful Chocolate Chocolate Chip Cookies</b> Submitted by: <b>Marsha</b>	★★★★★ 63 Ratings 49 Reviews
chocolate chips -- made with sour cream.	

## *Cookies that are out of this world...*



In the kitchen of a Whitman Massachusetts country inn, the first chocolate chip cookie emerged in 1937. Simple experiments led to a recipe combining bits of chocolate candy with a kind of butter cookie cookie dough resulting in a delicious mixture that offered the crunchiness of a cookie with a taste of chocolate candy in every bite. Obviously the cookies were a hit at the Inn and wherever else the recipe spread. Chocolate chip cookies have remained an American homemade treat.

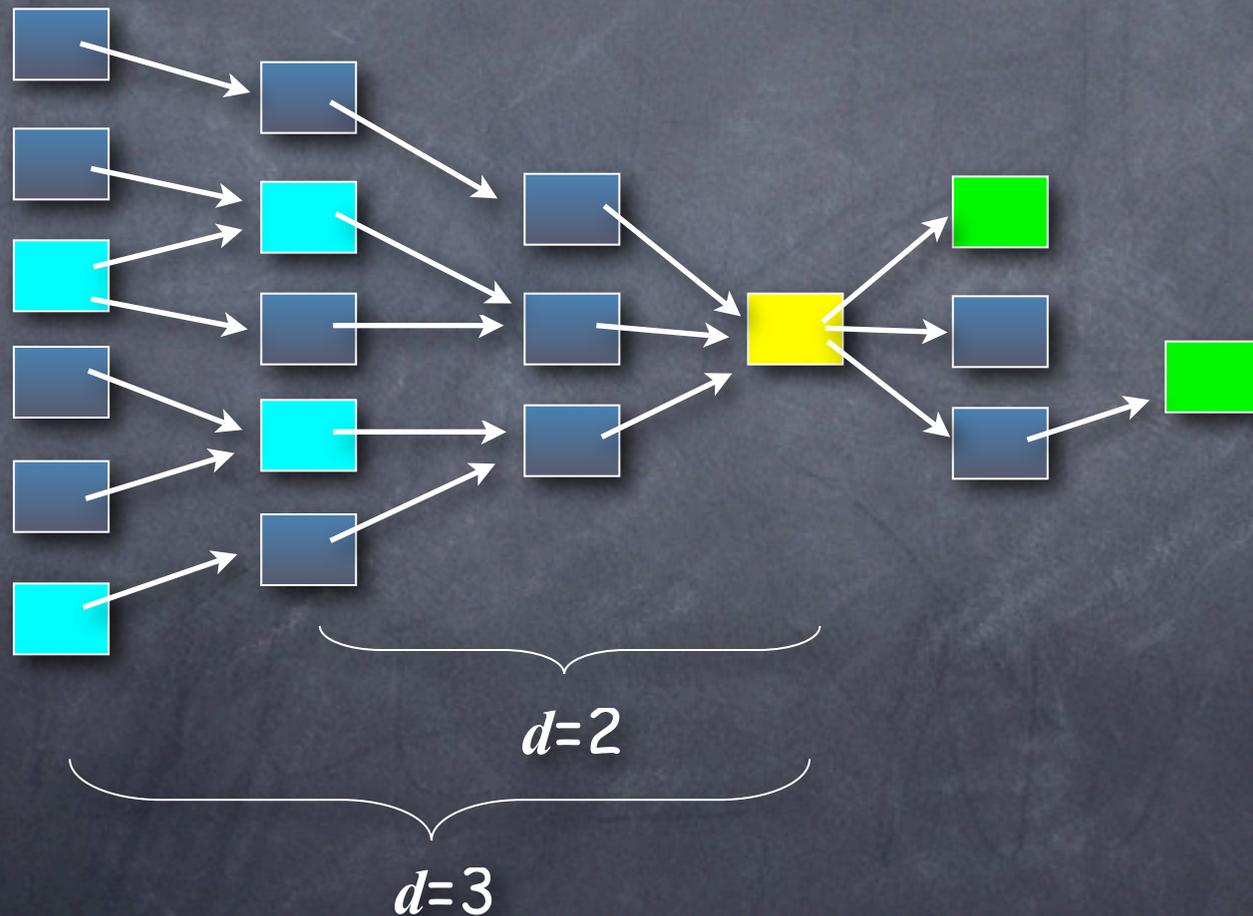
## CHOCOLATE CHIP COOKIES

### RECIPE INDEX

[BLACK AND WHITE CHOCOLATE CHIPPERS](#)  
[CLASSIC CHOCOLATE CHIP COOKIES](#)  
[COW CHIP COOKIES](#)  
[DEVIL'S FOOD CHOCOLATE CHIP COOKIES](#)  
[GOTTA HAVE EM' NOW! COOKIES](#)  
[MINT CHOCOLATE SANDWICH COOKIES](#)  
[NEIMAN MARCUS CHOCOLATE CHIP COOKIES](#)  
[OLD FASHIONED CHOCOLATE CHIPPERS](#)

# 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

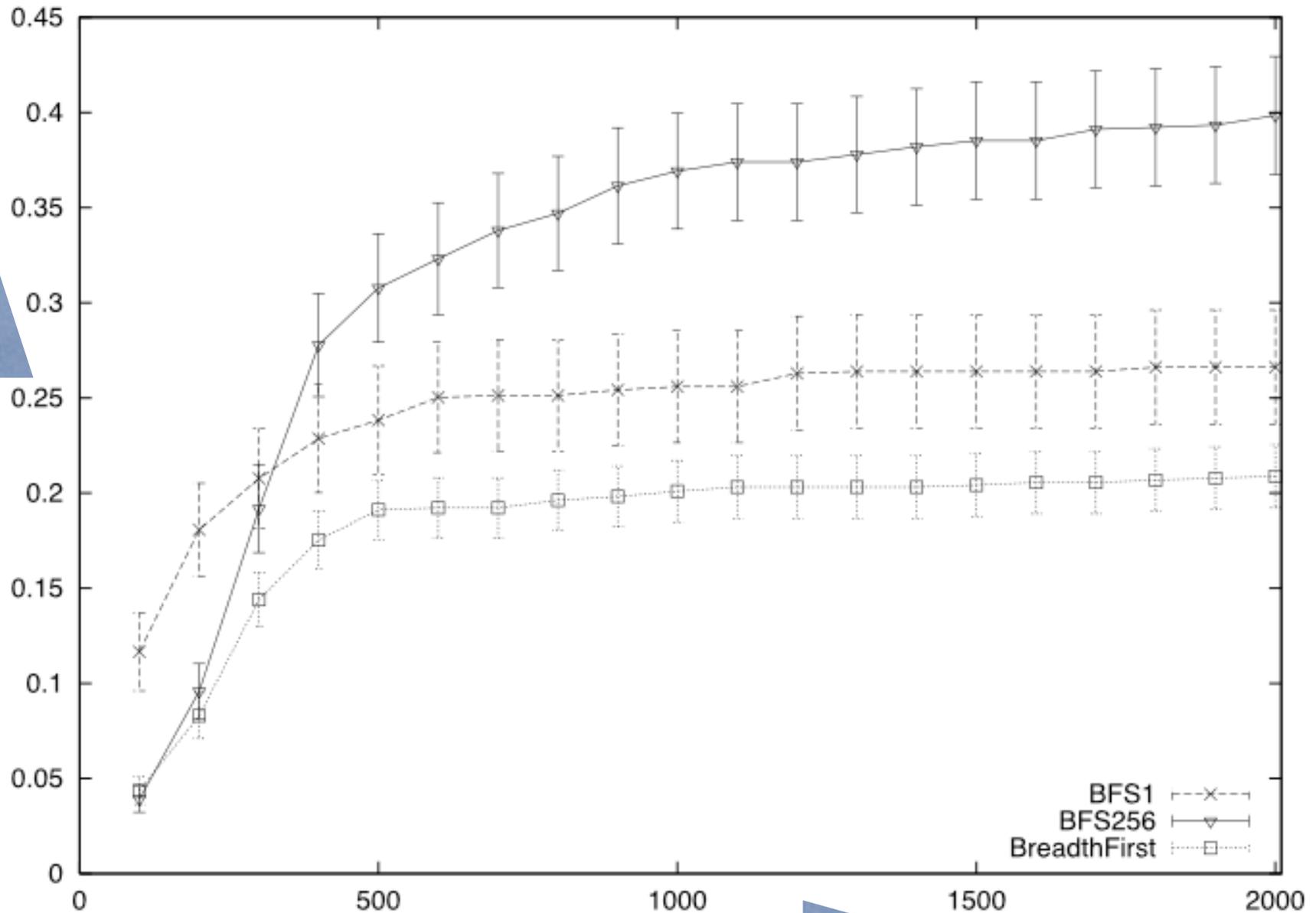
- **SharkSearch**

- Priority queue sorted by combination of similarity, anchor text, similarity of parent, etc.

- **InfoSpiders**

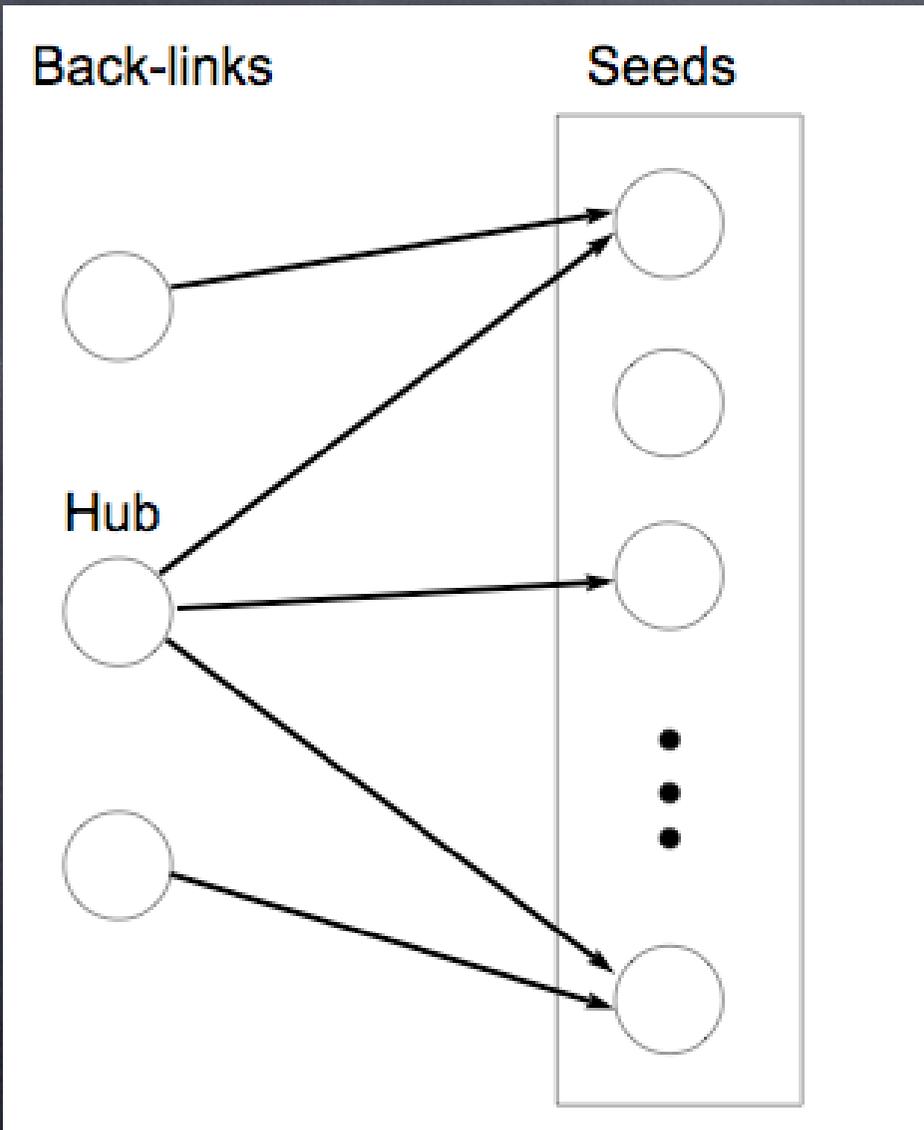
# Exploration vs. Exploitation

Avg target recall

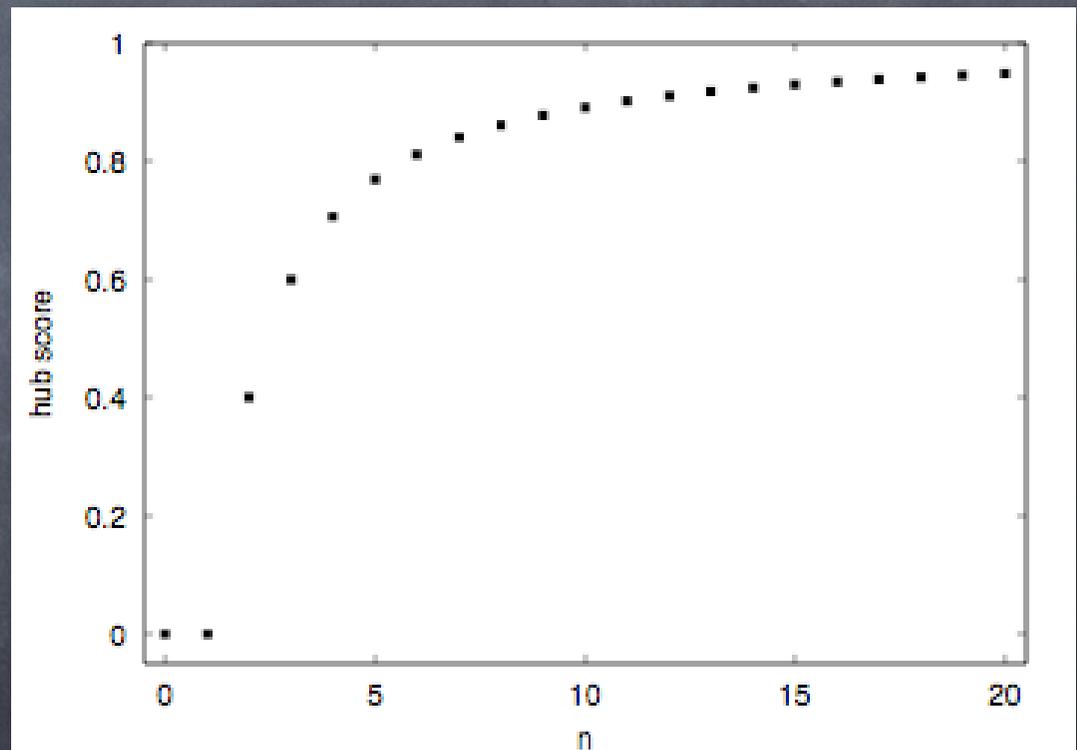


Pages crawled

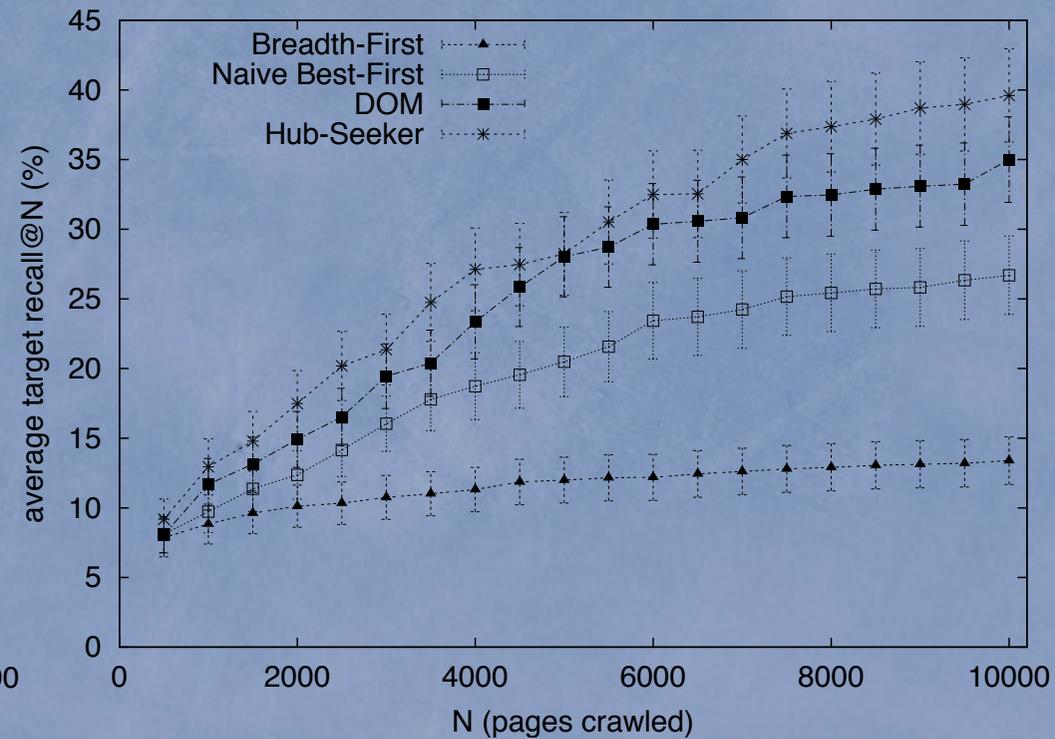
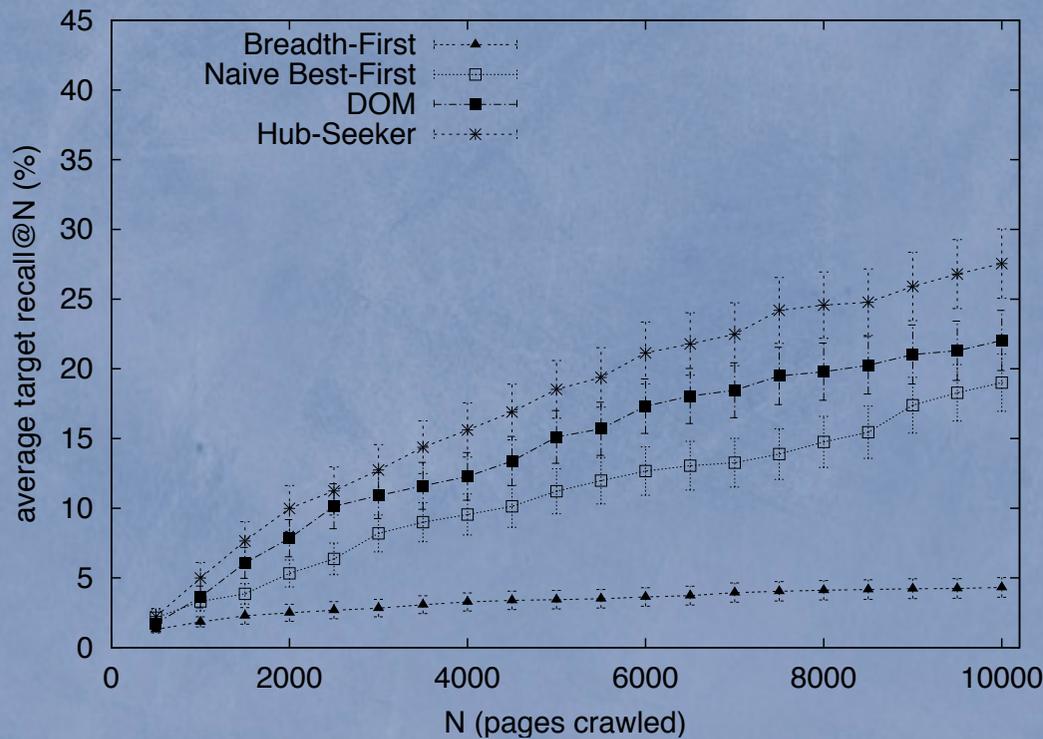
# Co-citation: hub scores



Link score<sub>hub</sub> = linear combination between link and hub score



# Recall (159 ODP topics)

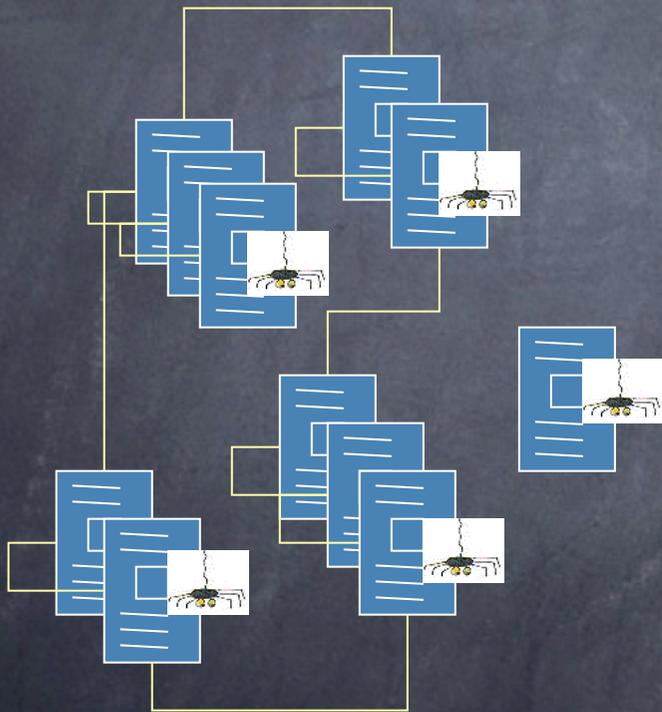


Split ODP URLs  
between seeds and  
targets

Add 10 best hubs to  
seeds for 94 topics

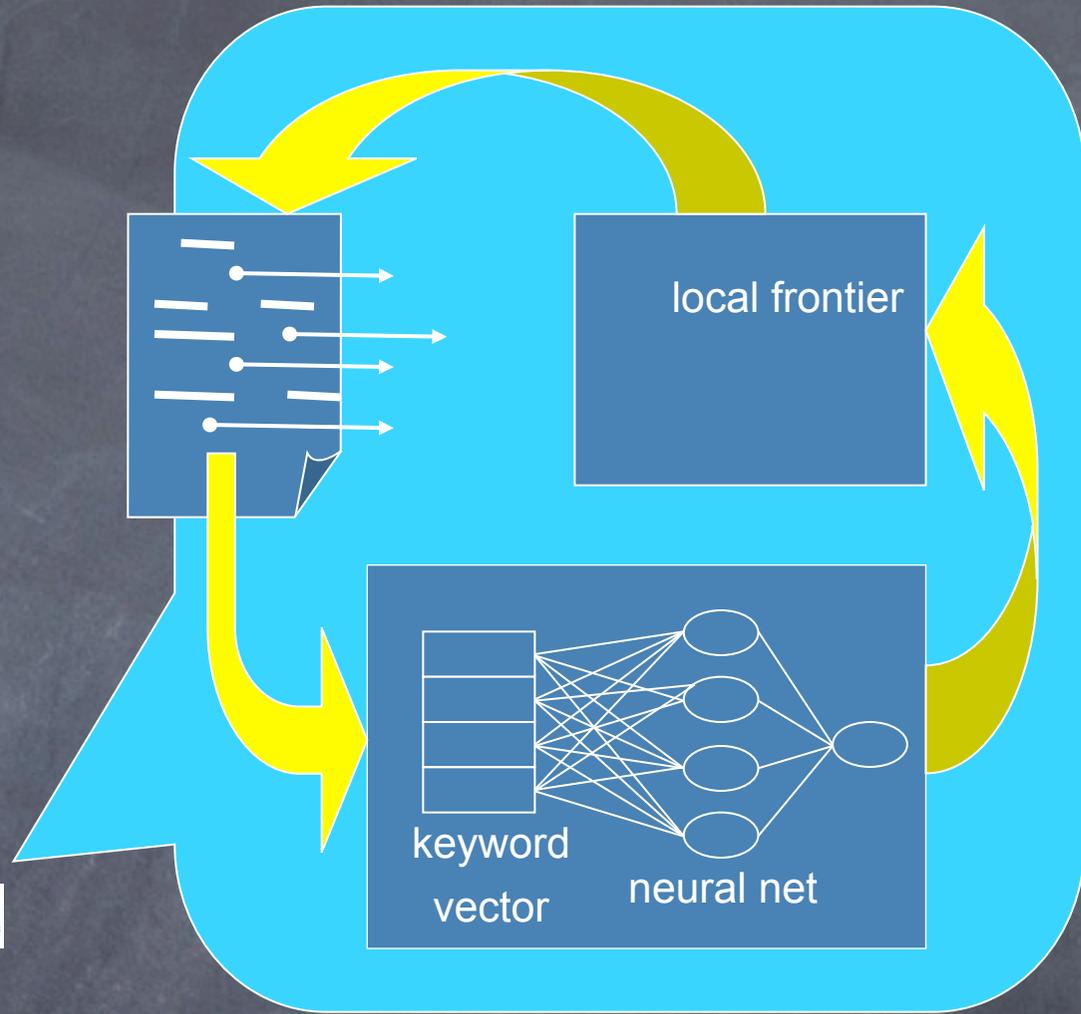
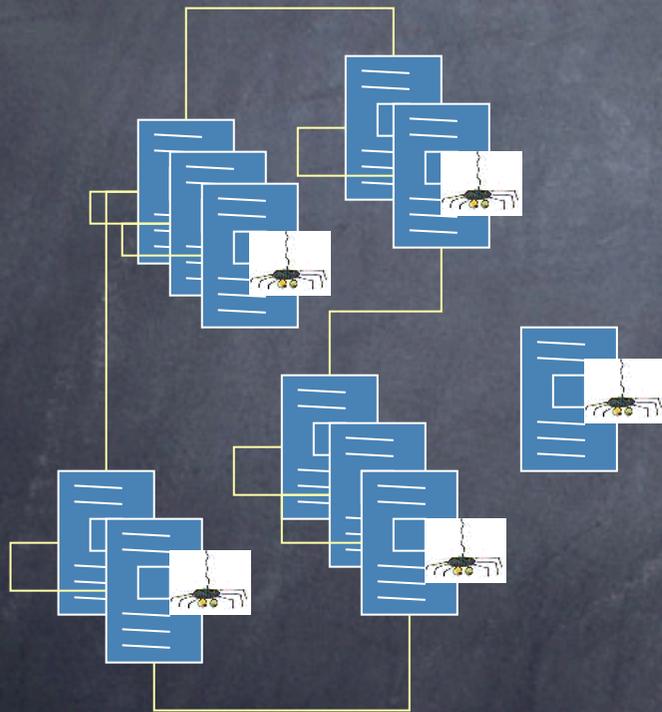
# InfoSpiders

adaptive distributed  
algorithm using an  
evolving population of  
learning agents



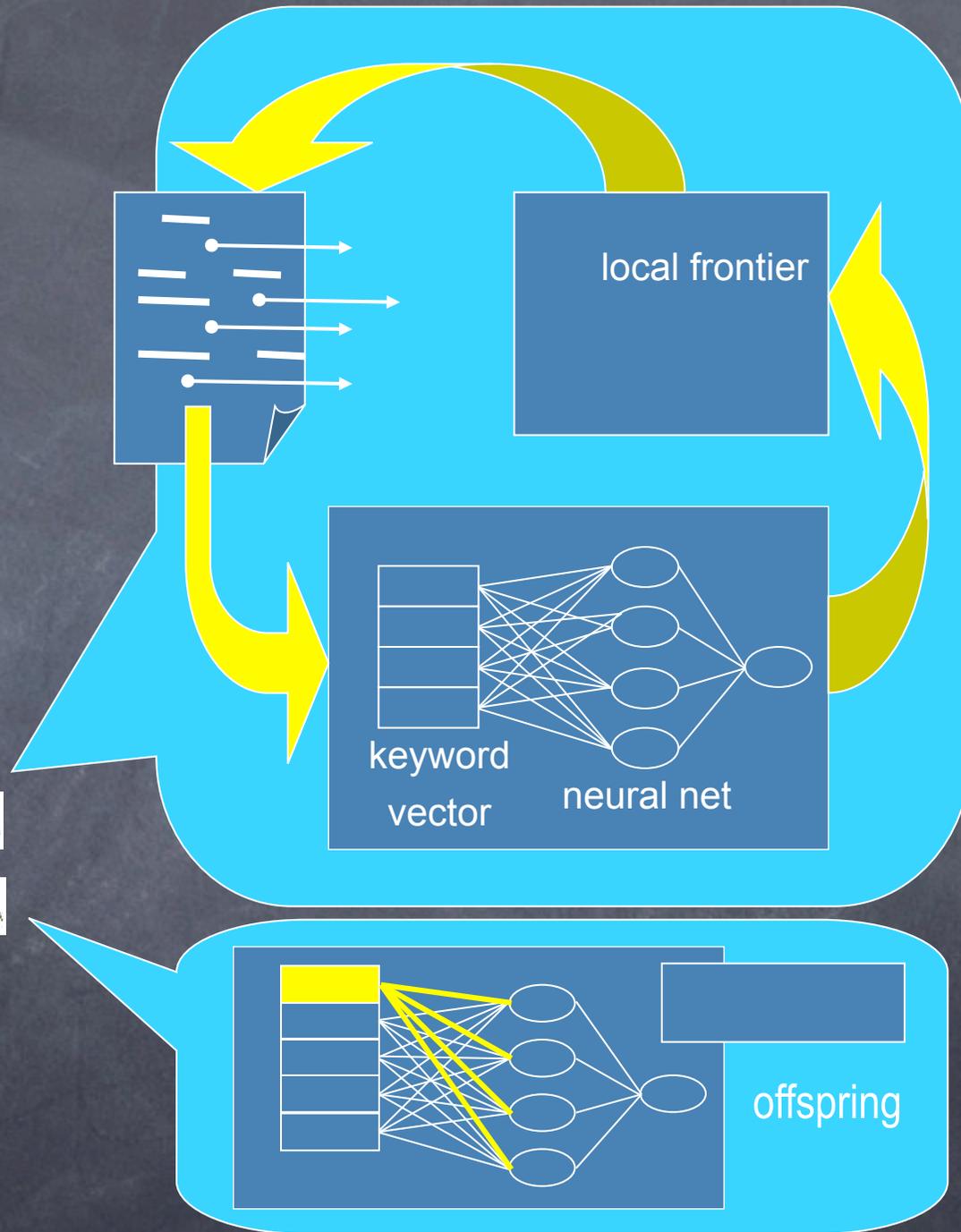
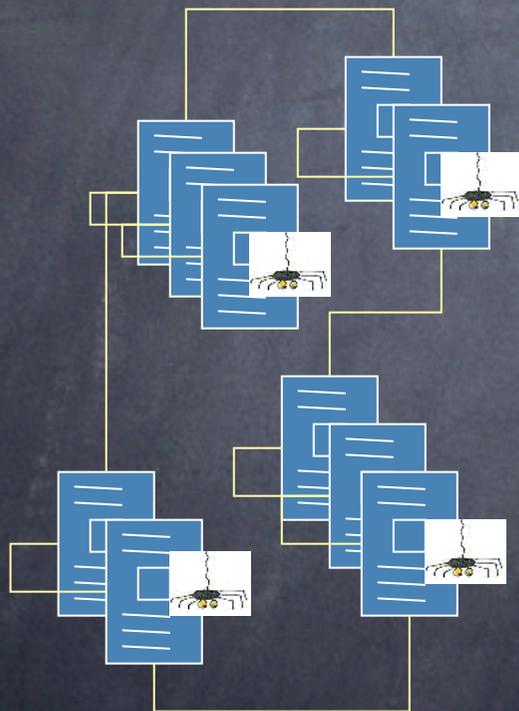
# InfoSpiders

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 preeminence 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 reminiscent 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 censorship 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' censorship - it would be useful to be able to compare results by country where 'business' censorship 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)

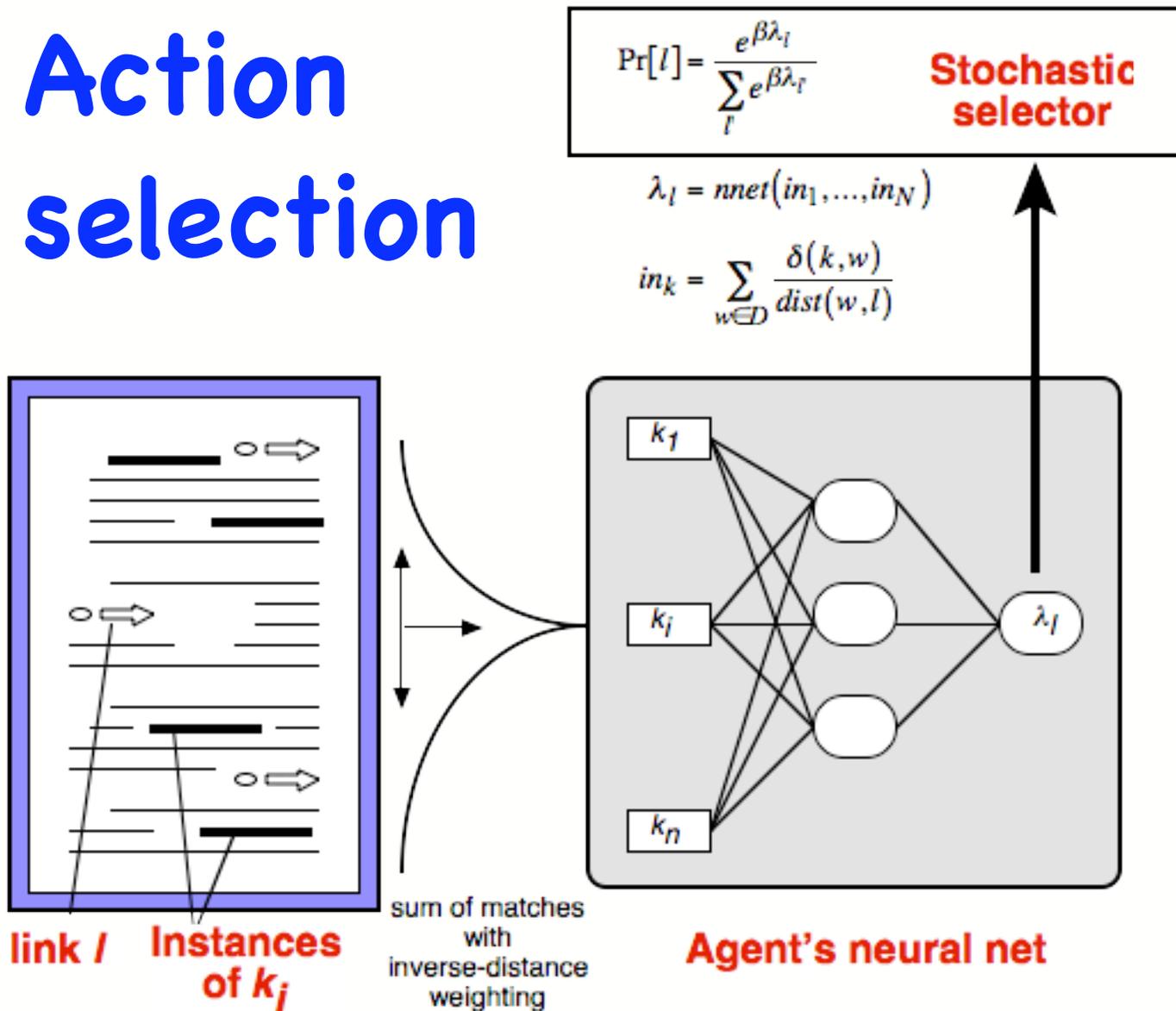
```
Foreach agent thread:  
  Pick & follow link from local frontier  
  Evaluate new links, merge frontier  
  Adjust link estimator  
   $E := E + \text{payoff} - \text{cost}$   
  If  $E < 0$ :  
    Die  
  Elsif  $E > \text{Selection\_Threshold}$ :  
    Clone offspring  
    Split energy with offspring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

reinforcement  
learning

match  
resource  
bias

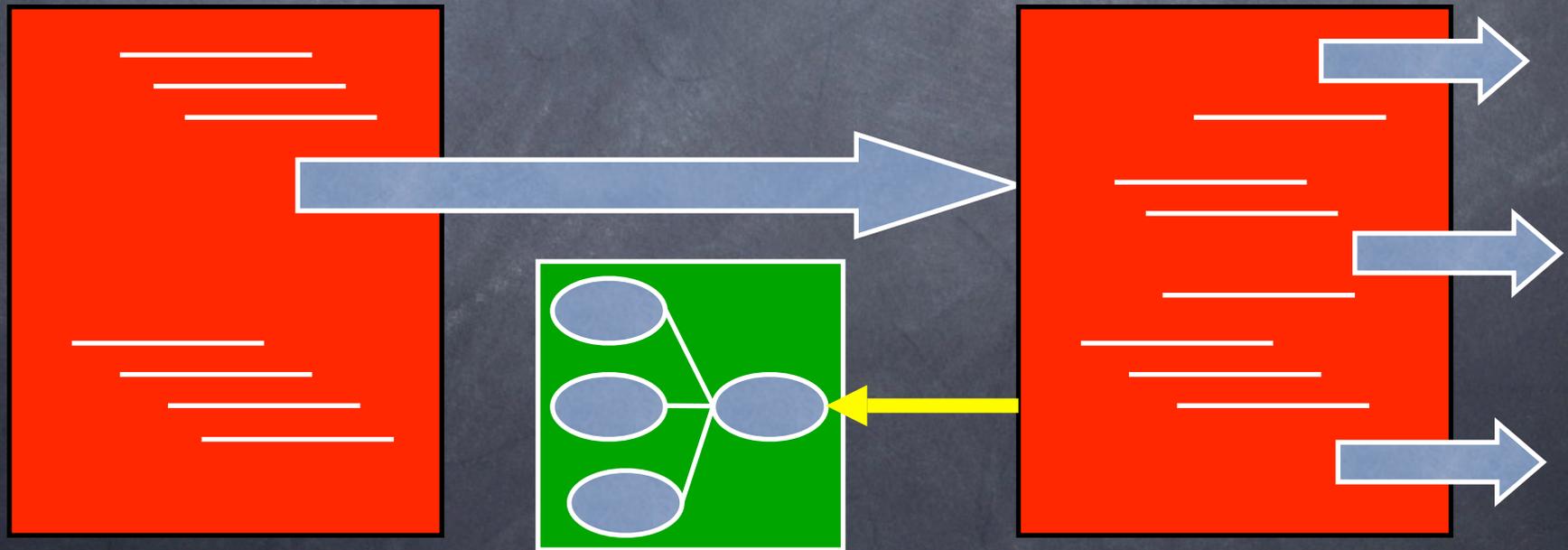
selective  
query  
expansion

# Action selection

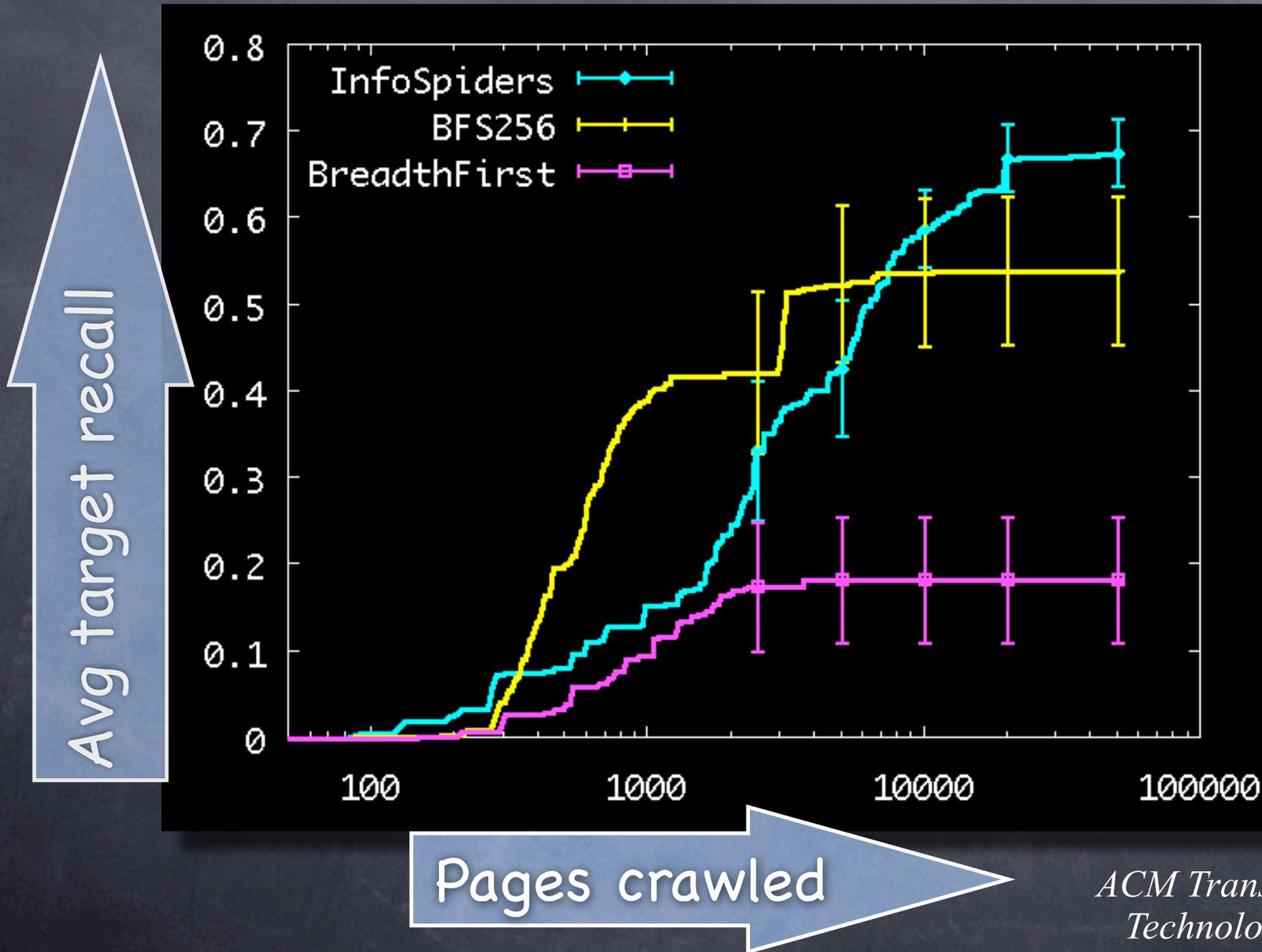


# Q-learning

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



# Performance





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](http://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/](http://www.parishilton.com/) - 8k - [Cached](#) - [Similar pages](#)

## [Paris Hilton](#)

**Paris Hilton** on IMDb: Movies, TV, Celebs, and more...  
[www.imdb.com/name/nm0385296/](http://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/](http://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/](http://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](http://www.askmen.com/women/models_150/171_paris_hilton.html) - [Similar pages](#)

## Sponsored Links

### [Paris Hilton](#)

Video News, Articles, Blogs & More.  
Get 24/7 Entertainment News at TMZ!  
[TMZ.com](#)

### [Love Music? Review it](#)

Read other opinions on Music  
and Register to Voice Yours  
[www.HeyNielsen.com](http://www.HeyNielsen.com)

### [Paris Hiltons Diary](#)

Read **Paris Hilton's** daily prison diary. Daily updates and more.  
[www.ParisHiltonsPrisonDiary.com](http://www.ParisHiltonsPrisonDiary.com)

### [Hot Paris Hilton](#)

Hot Pics of **Paris Hilton**  
Visit "Girls of Maxim" Gallery Now  
[MaximOnline.com](http://MaximOnline.com)

**News results for paris**

ITN

**Paris Hilton**  
By Cathy R  
American o  
Korea Time  
Hallmark F**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™ \(8\)](#)

▼ advertisement

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

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

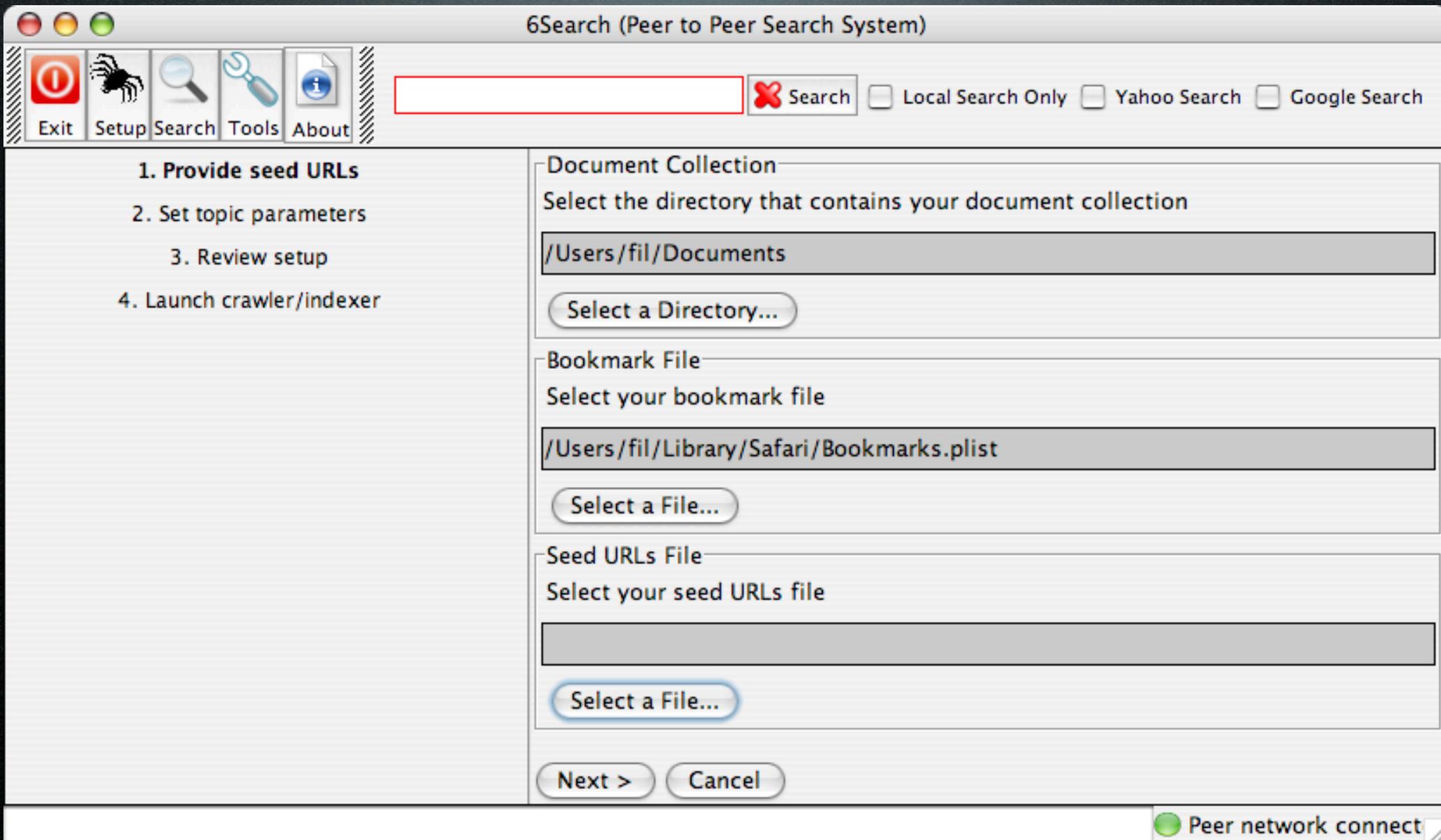
[▶ START HERE](#)**Paris Hilton Travel Deals***Sponsored links \****Compare all deals on Paris Hilton****Paris Hilton: Special autumn and winter rates!** SmartDeal**Hilton.com** Our Best Rates. Guaranteed. Book now!**Paris Hilton: Great rooms from \$282** SmartDeal**Expedia.com** Get the lowest price 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, France1 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://sixsearch.org>



<http://sixearch.org>

6Search (Peer to Peer Search System)

Exit Setup Search Tools About

Local Search Only  Yahoo Search  Google Search

1. Provide seed URLs
2. Set topic parameters
3. Review setup
4. Launch crawler/indexer

Crawling Progress  
11/09 14:53:34

Indexing Progress  
11/09 14:56:28

Information Summary

Please verify the information you entered.

Document Collection:  
/Users/fil/Documents/Homepages/Informatics

Bookmark File:  
/Users/fil/Library/Safari/Bookmarks.plist

Crawling Topic:  
web mining modeling search peer crawl networks agents

Seed URLs File:

Number of Pages:  
100

Your Email:  
fil@indiana.edu

Tag for URLs:  
web mining modeling search peer crawl networks agents

Peer network connect

<http://sixearch.org>

The search results of 6S system

http://localhost:1999/result/ 6S web mining

Mozilla.org Plug-in FAQ CSN givealink.org Oncourse OneStart CenSEARCHip MySpiders Windows Live Google Calendar CX

[KDnuggets: Data Mining, Web Mining, and Knowledge Discovery Guide](#)  
Newsletter on the data mining and knowledge industries, offering information on data mining, knowledge discovery, text mining, and web mining software, courses, jobs, publications, and meetings.  
<http://www.kdnuggets.com/>  
Contributors: Main6S#0,wls\_angel#3,wls\_iceman#1  
[Similar pages \(Power by GiveALink\)](#)

[Data Mining - Home Page \(Misc\)](#)  
... in an unrestricted hands-on web. Software Information on **Data Mining** Software Evaluate Software, Read and ... OnLine Analytical Processing (OLAP) , **Data Mining** ...  
<http://www.the-data-mine.com/>  
Contributors: Unknown#7,wls\_angel#3,wls\_iceman#1  
[Similar pages \(Power by GiveALink\)](#)

[Data Mining Software in the Yahoo! Directory](#)  
... for IT professionals focusing on **data mining**, **data** analysis, and reports ... to Business > Computers > Software > Databases > **Data** ...  
[http://dir.yahoo.com/Business\\_and\\_Economy/Business\\_to\\_Business/Computers/Software/Databases/Data\\_Mining/](http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Computers/Software/Databases/Data_Mining/)  
Contributors: Unknown#7,wls\_iceman#1  
[Similar pages \(Power by GiveALink\)](#)

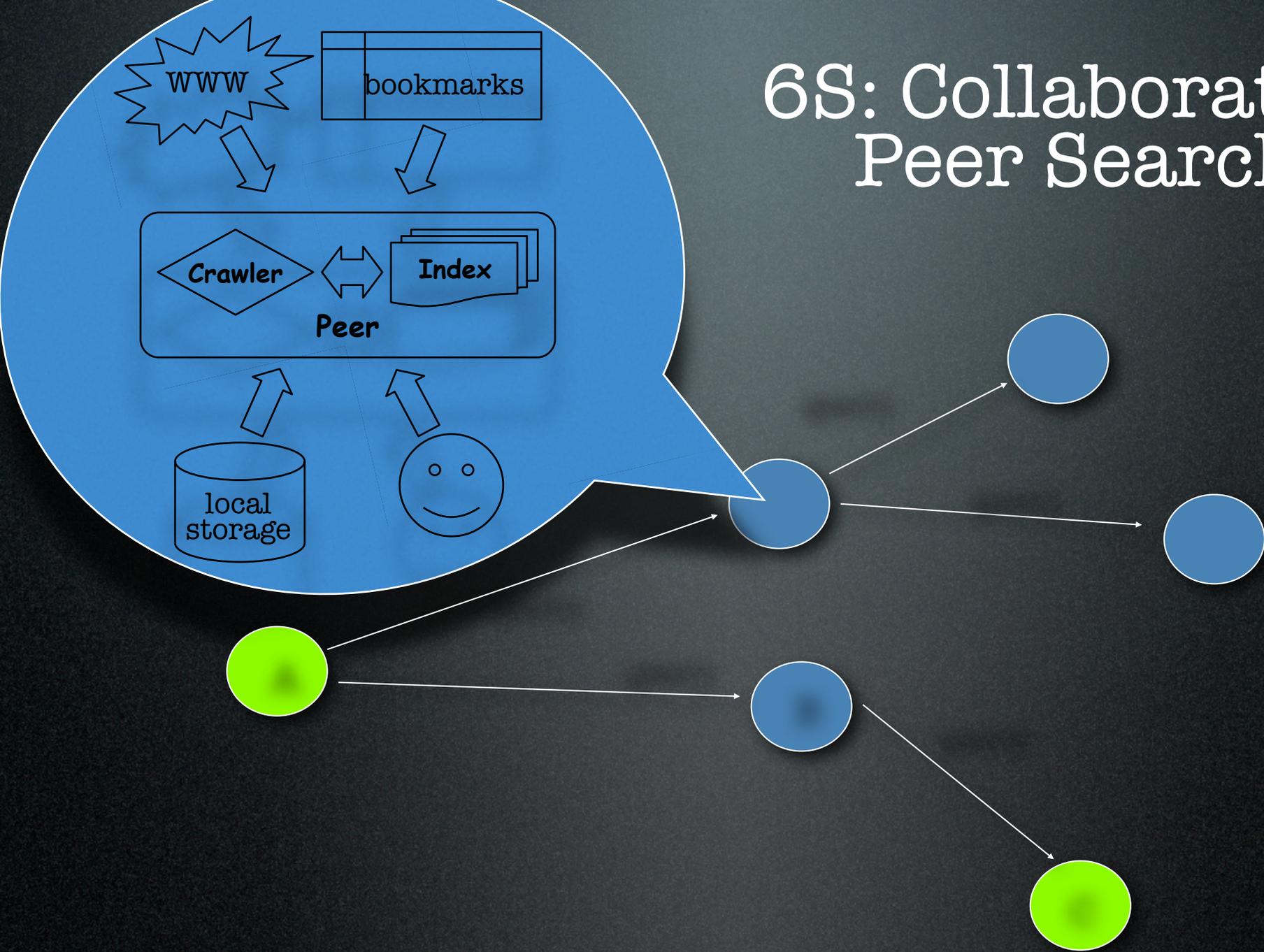
[UMBC Agent Web -- news and information on software agent technology](#)  
... UMBC Agent Web -- news and information ...  
[http://agents.umbc.edu/Topics/Related\\_Topics/Information\\_retrieval\\_and\\_knowledge\\_management/](http://agents.umbc.edu/Topics/Related_Topics/Information_retrieval_and_knowledge_management/)  
Contributors: Unknown#7,wls\_iceman#1  
[Similar pages \(Power by GiveALink\)](#)

[BUBL LINK: Information retrieval](#)  
... and Forward Knowledge Approach **Data Mine: Data Mining** and Knowledge Discovery ... the Gaps KD Mine: **Data** ...  
<http://bubl.ac.uk/link/i/informationretrieval.htm>  
Contributors: Unknown#7,wls\_iceman#1  
[Similar pages \(Power by GiveALink\)](#)

[Library and Information Science > Information Retrieval in the Yahoo! Directory](#)

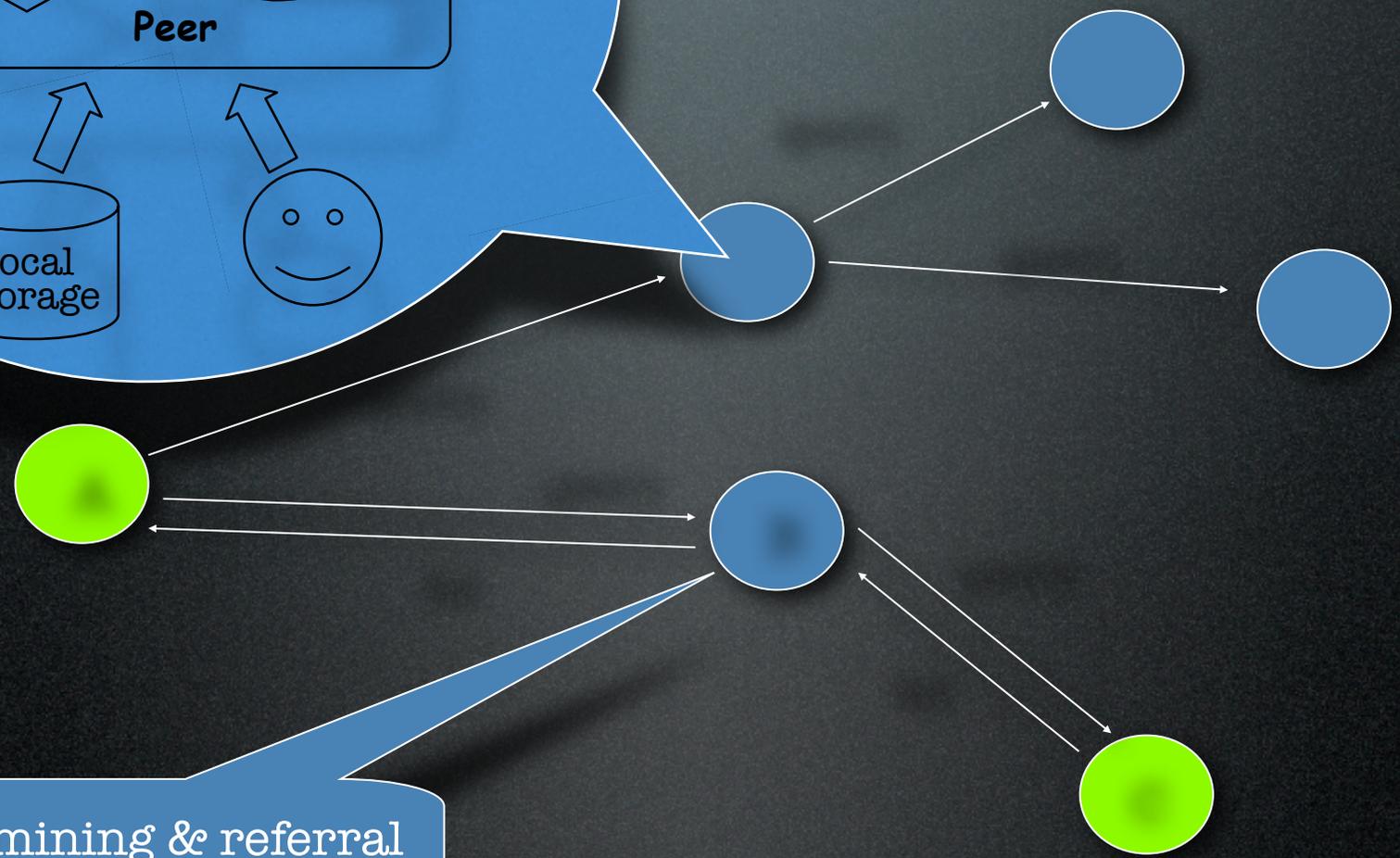
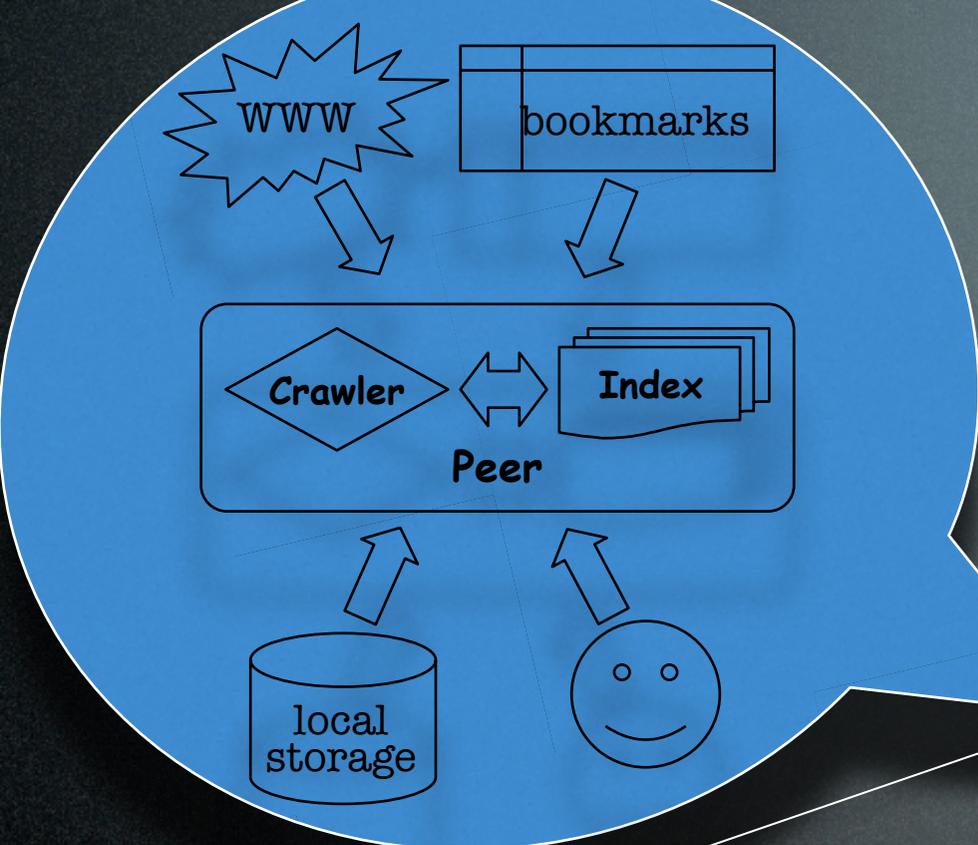
Done

# 6S: Collaborative Peer Search



WWW2004  
WWW2005  
WTAS2005  
P2PIR2006

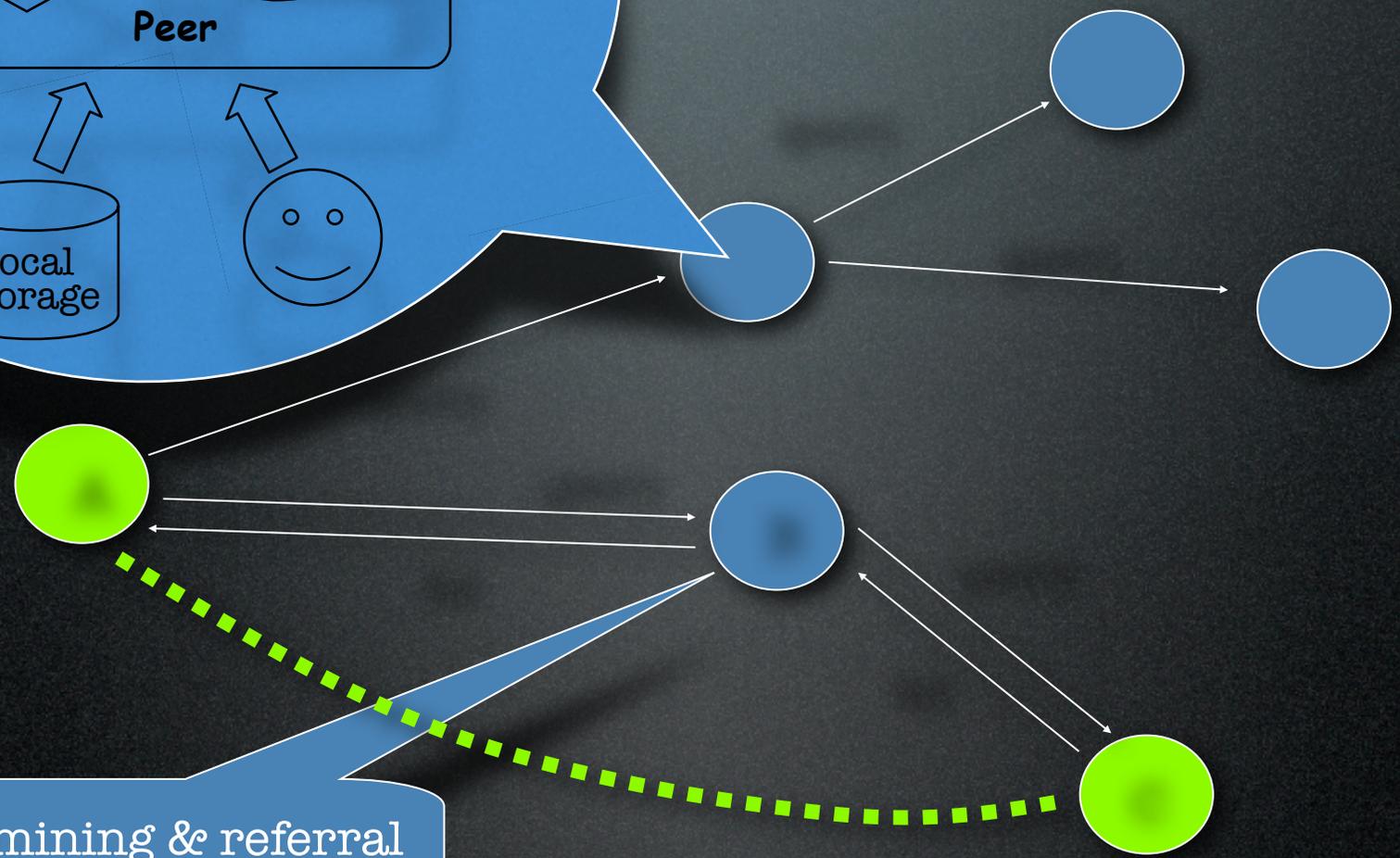
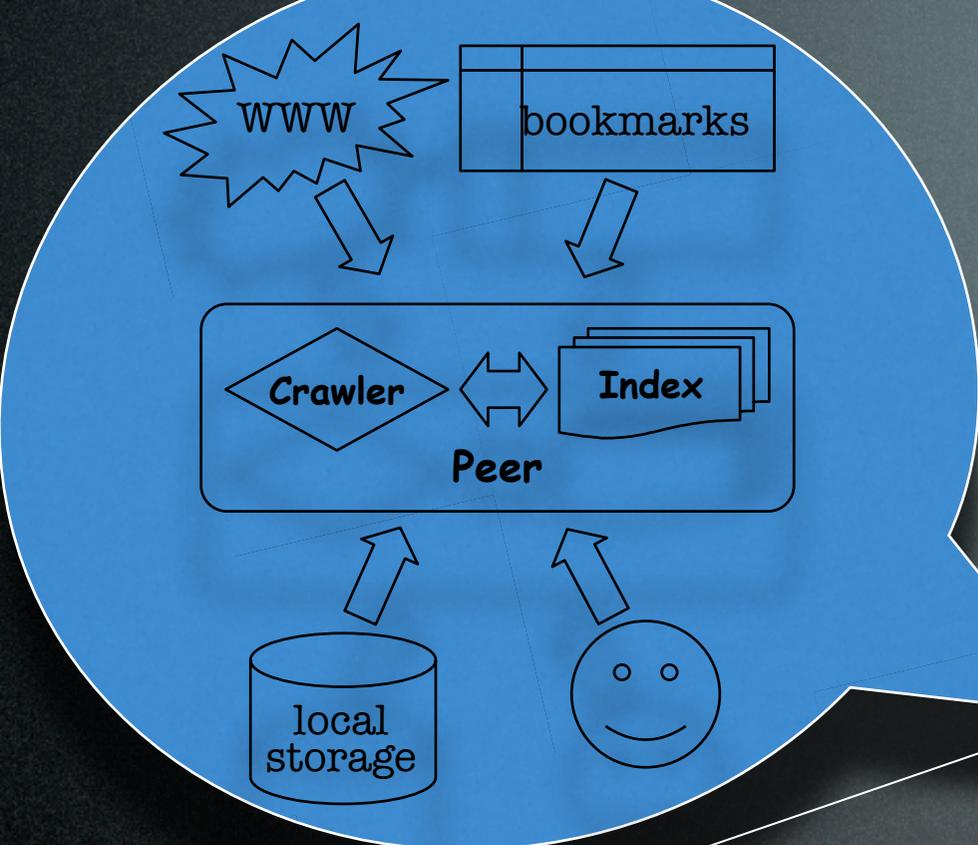
# 6S: Collaborative Peer Search



Data mining & referral opportunities

WWW2004  
WWW2005  
WTAS2005  
P2PIR2006

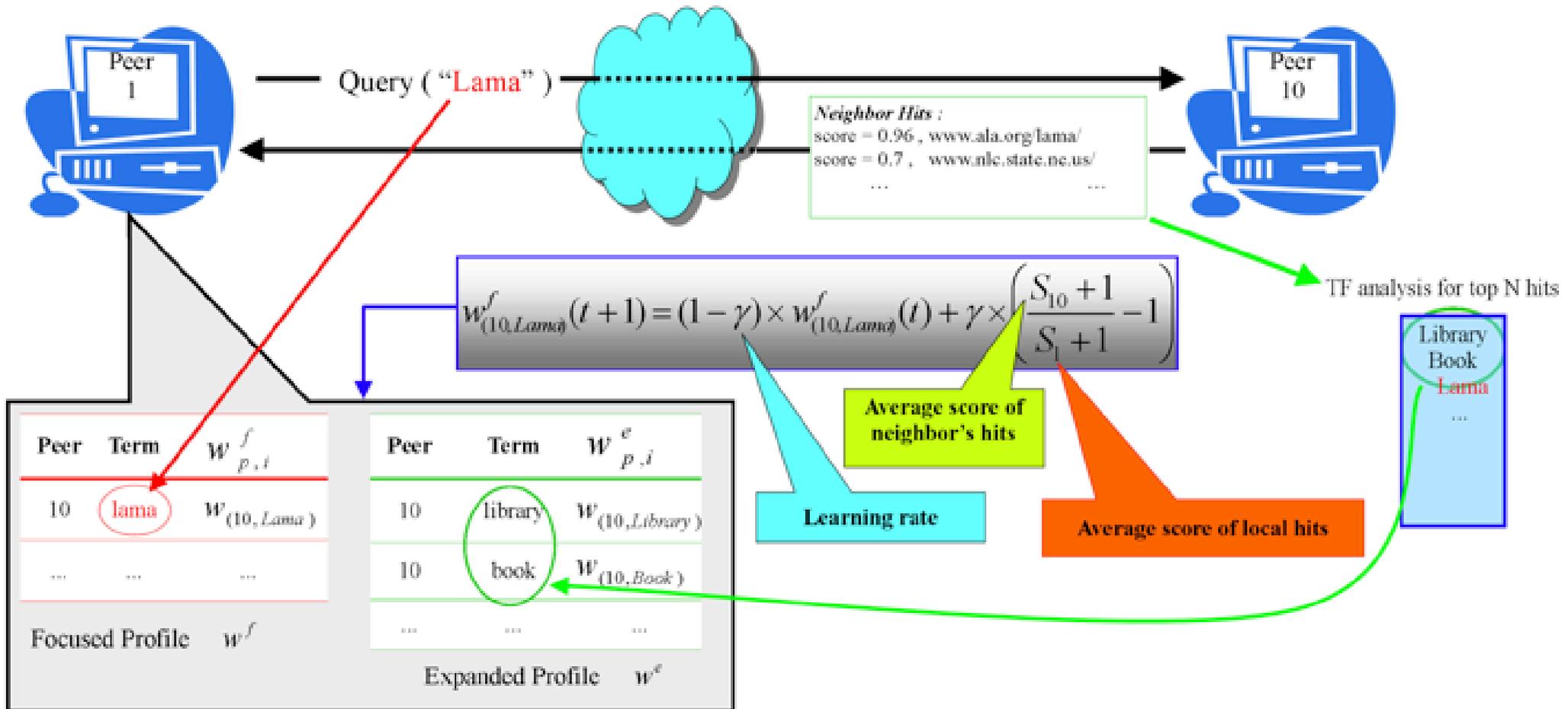
# 6S: Collaborative Peer Search



Data mining & referral opportunities

WWW2004  
WWW2005  
WTAS2005  
P2PIR2006

# Reinforcement Learning



# Query Routing



Query ( "Lama library book" )

$$\sigma(p, Query) = \sum_{i \in Query} [\alpha \cdot w_{p,i}^f + (1 - \alpha) \cdot w_{p,i}^e]$$

Peer	Term	$w_{p,i}^f$
10	lama	.98
...	...	...

Focused Profile  $w^f$

Peer	Term	$w_{p,i}^e$
10	library	.87
10	book	.67
...	...	...

Expanded Profile  $w^e$

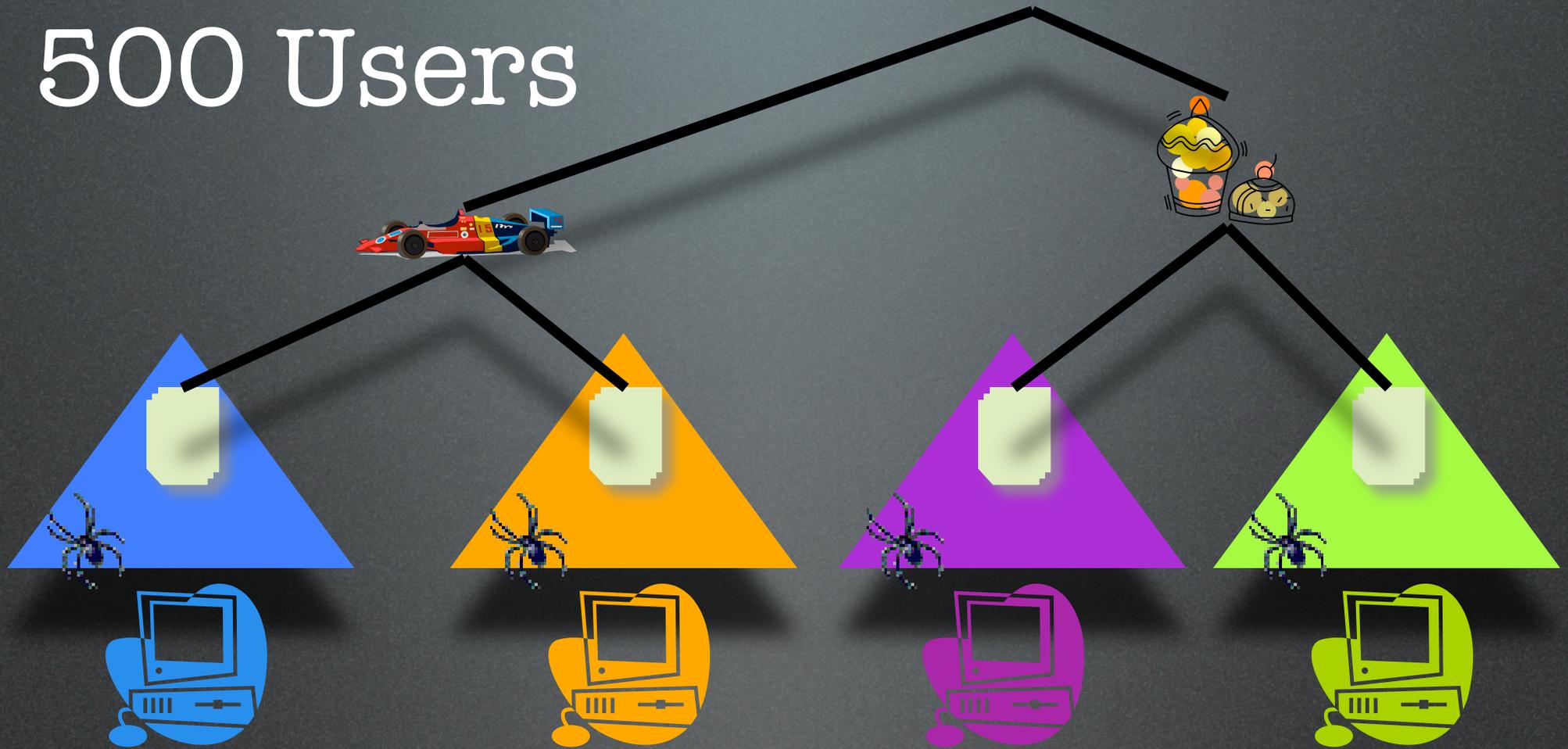
N top ranked among known peers are selected as neighbors and sent Query.

*Forwarding the query*



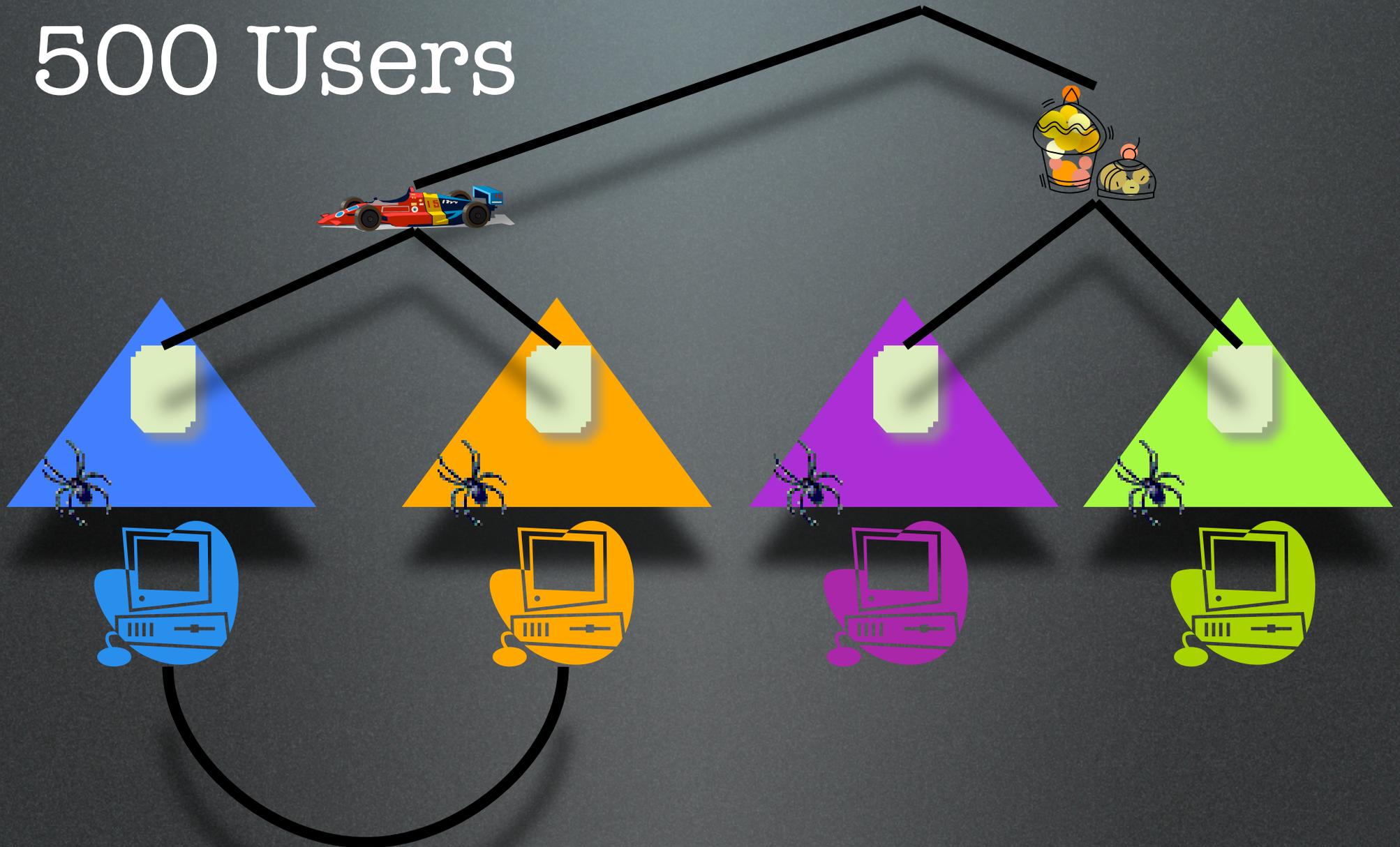
# Simulating 500 Users

ODP (dmoz.org)



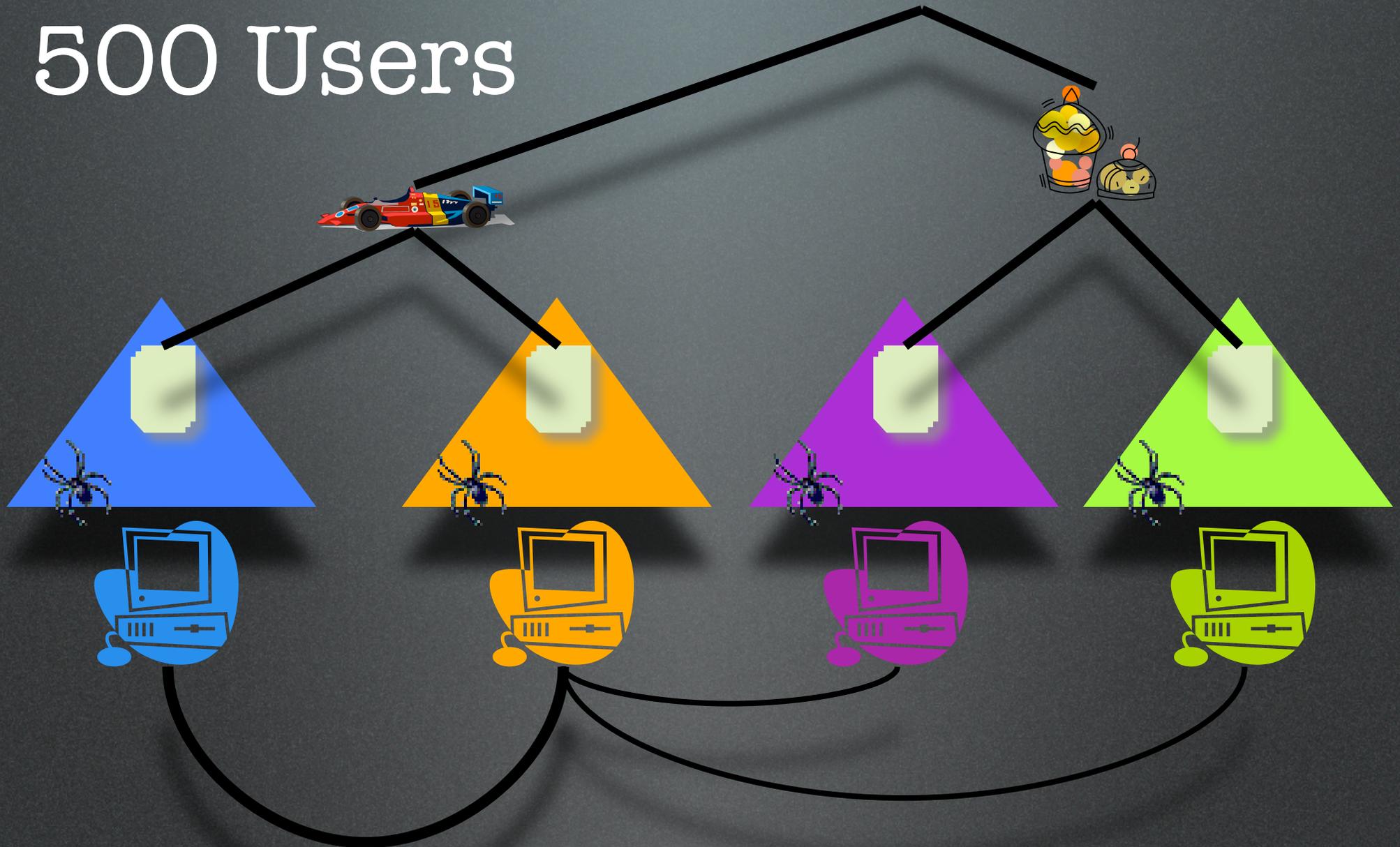
# Simulating 500 Users

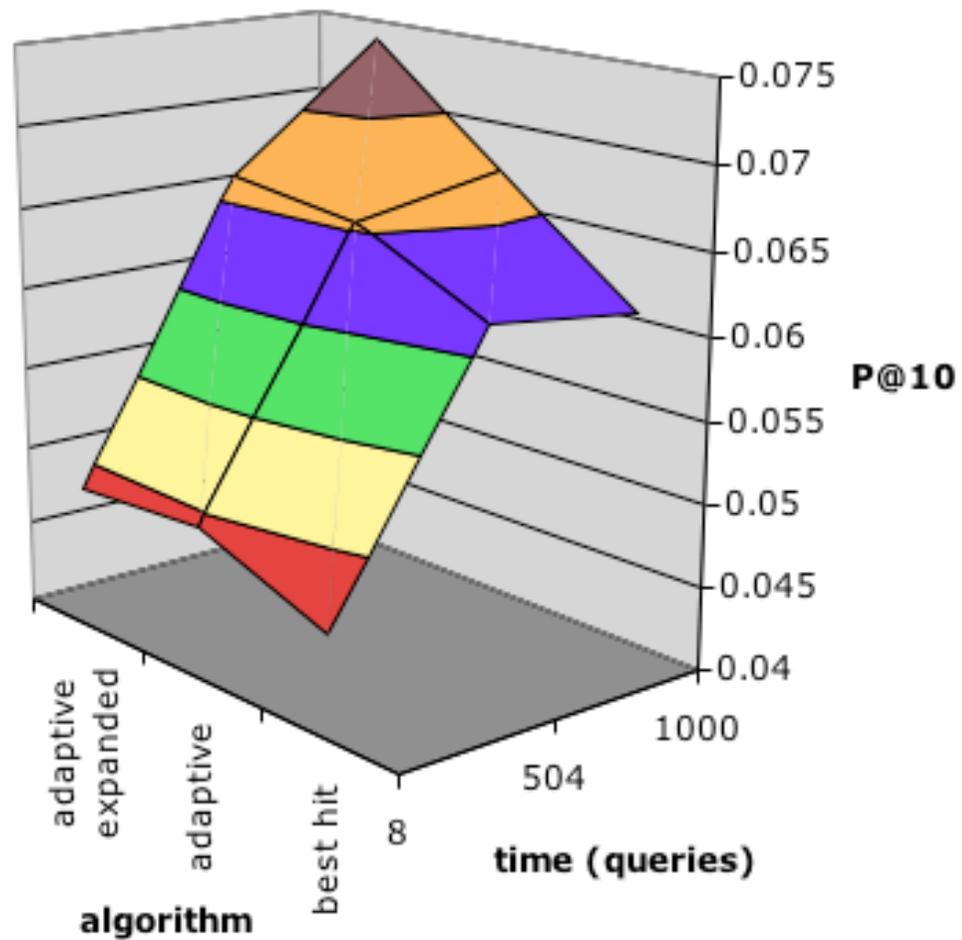
ODP (dmoz.org)



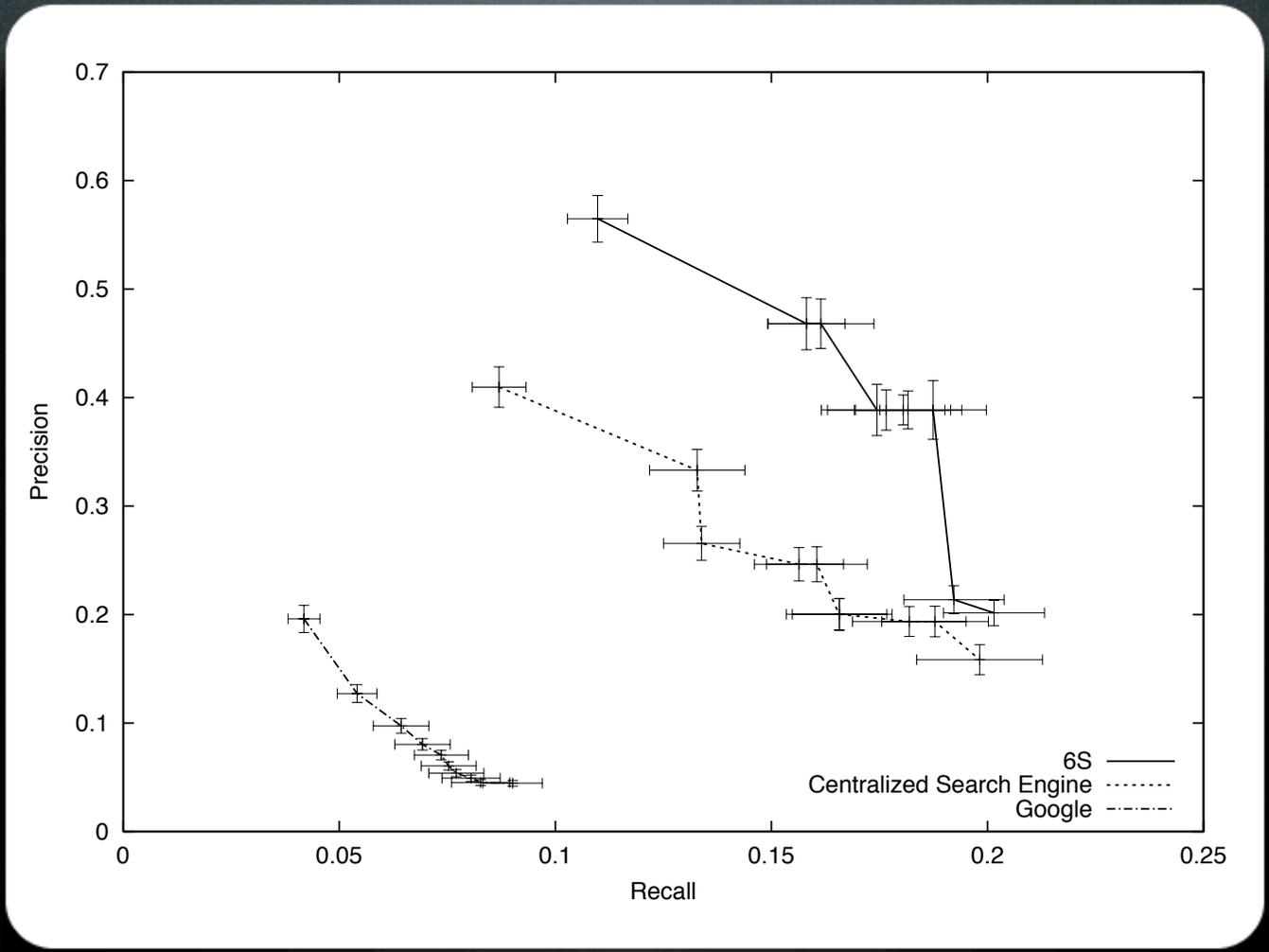
# Simulating 500 Users

ODP (dmoz.org)

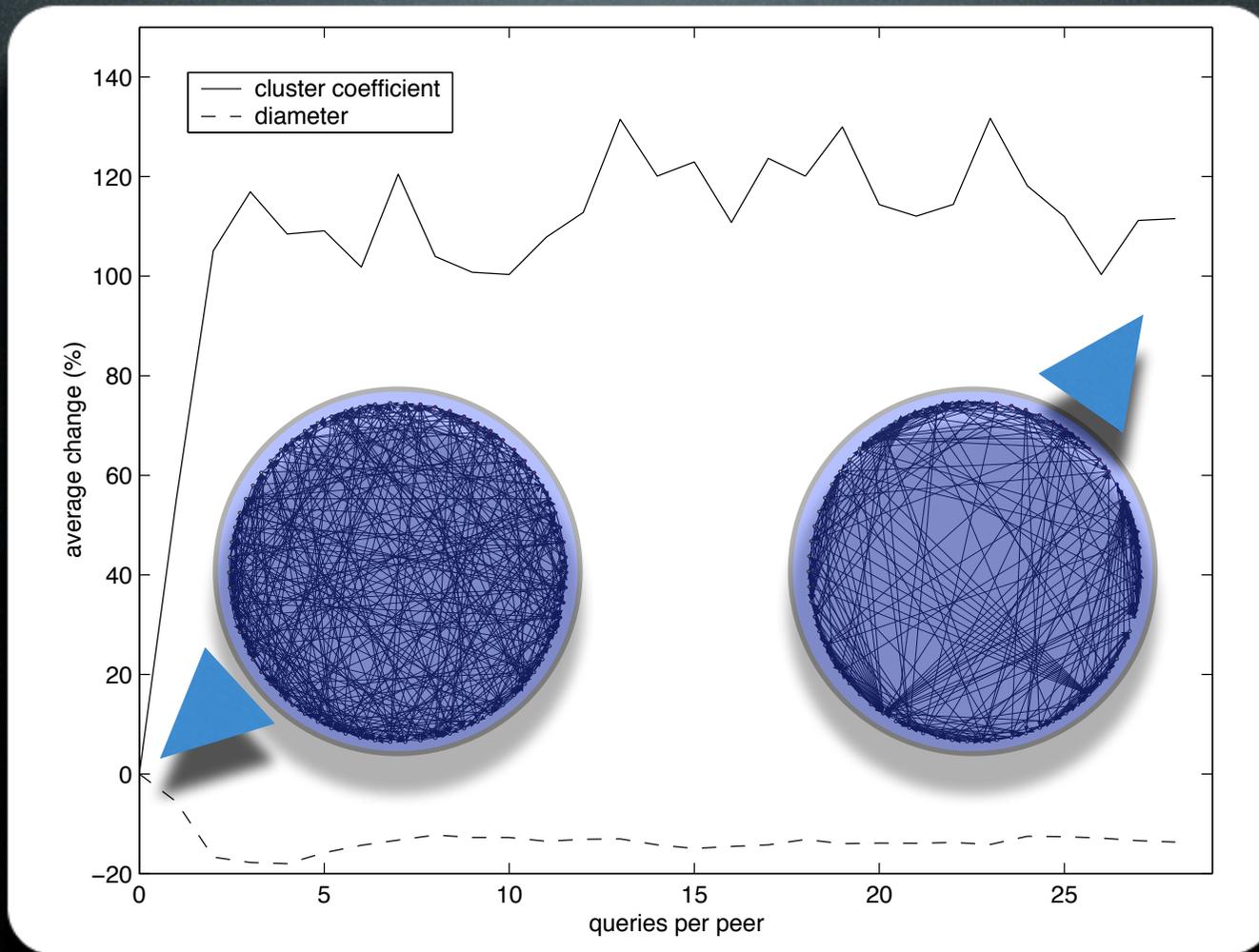




P@10

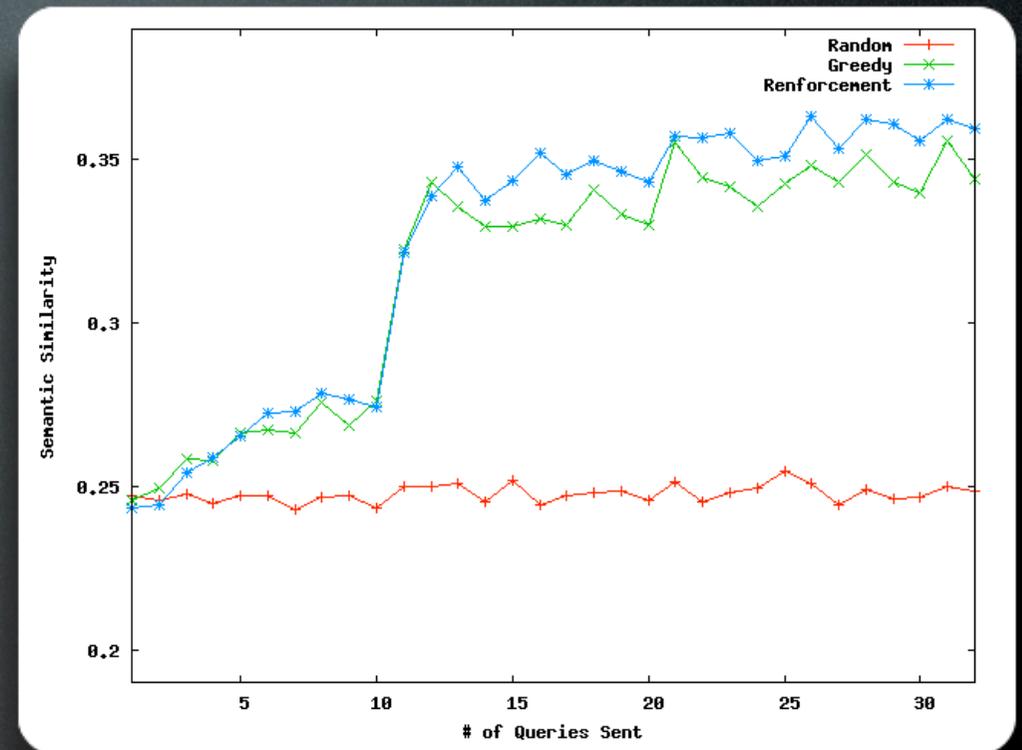


# Distributed vs Centralized

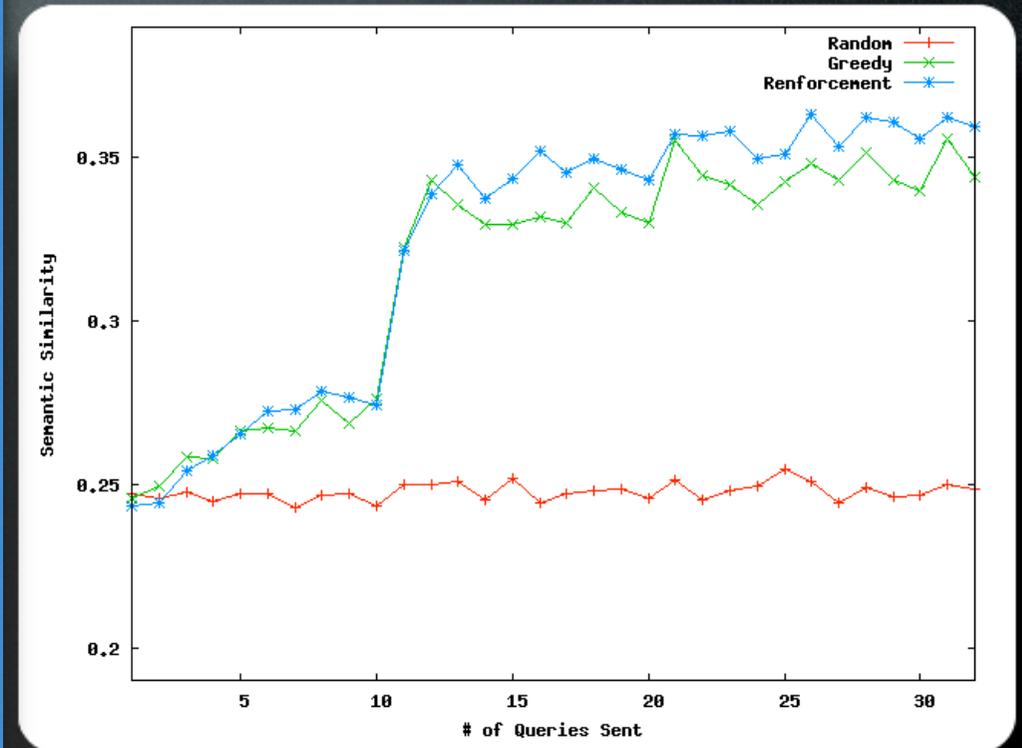


# Small-world

# Semantic Similarity

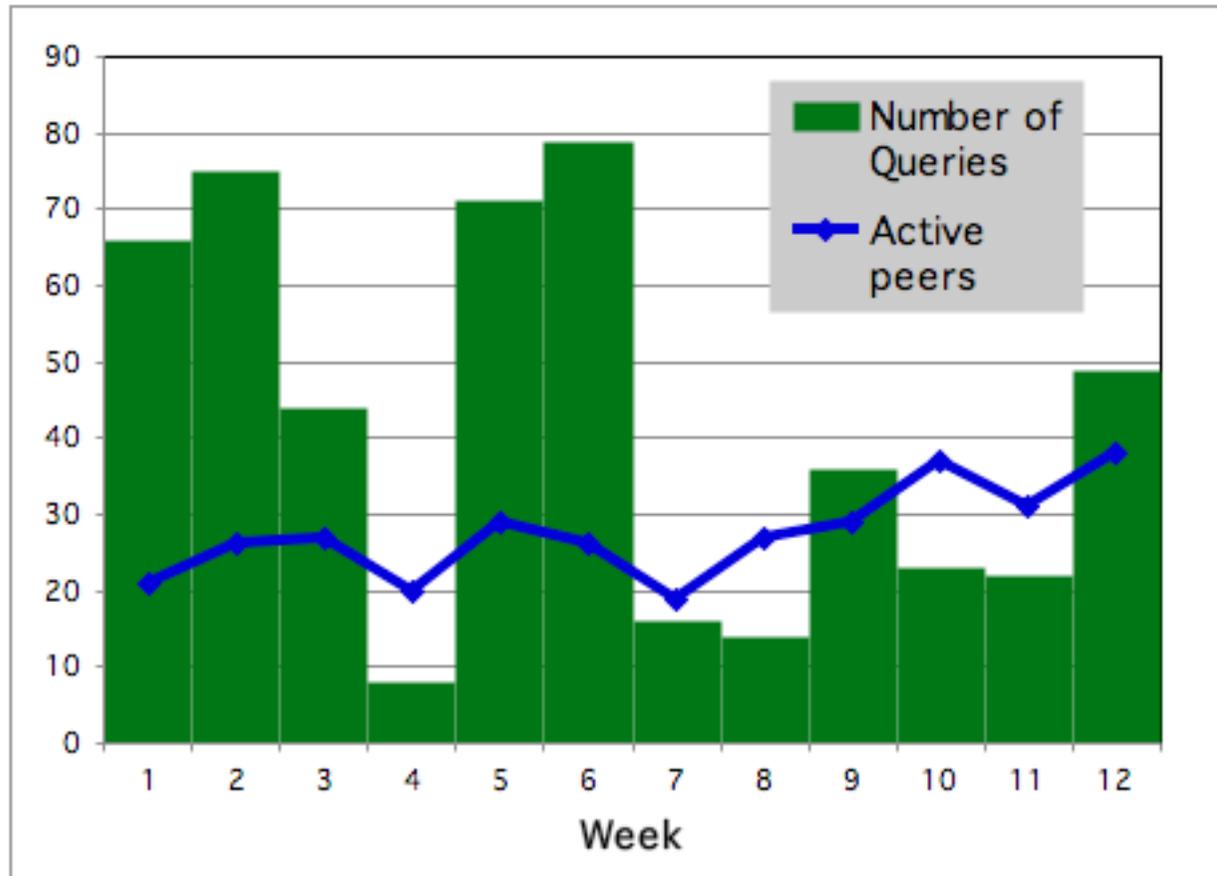


# Semantic Similarity

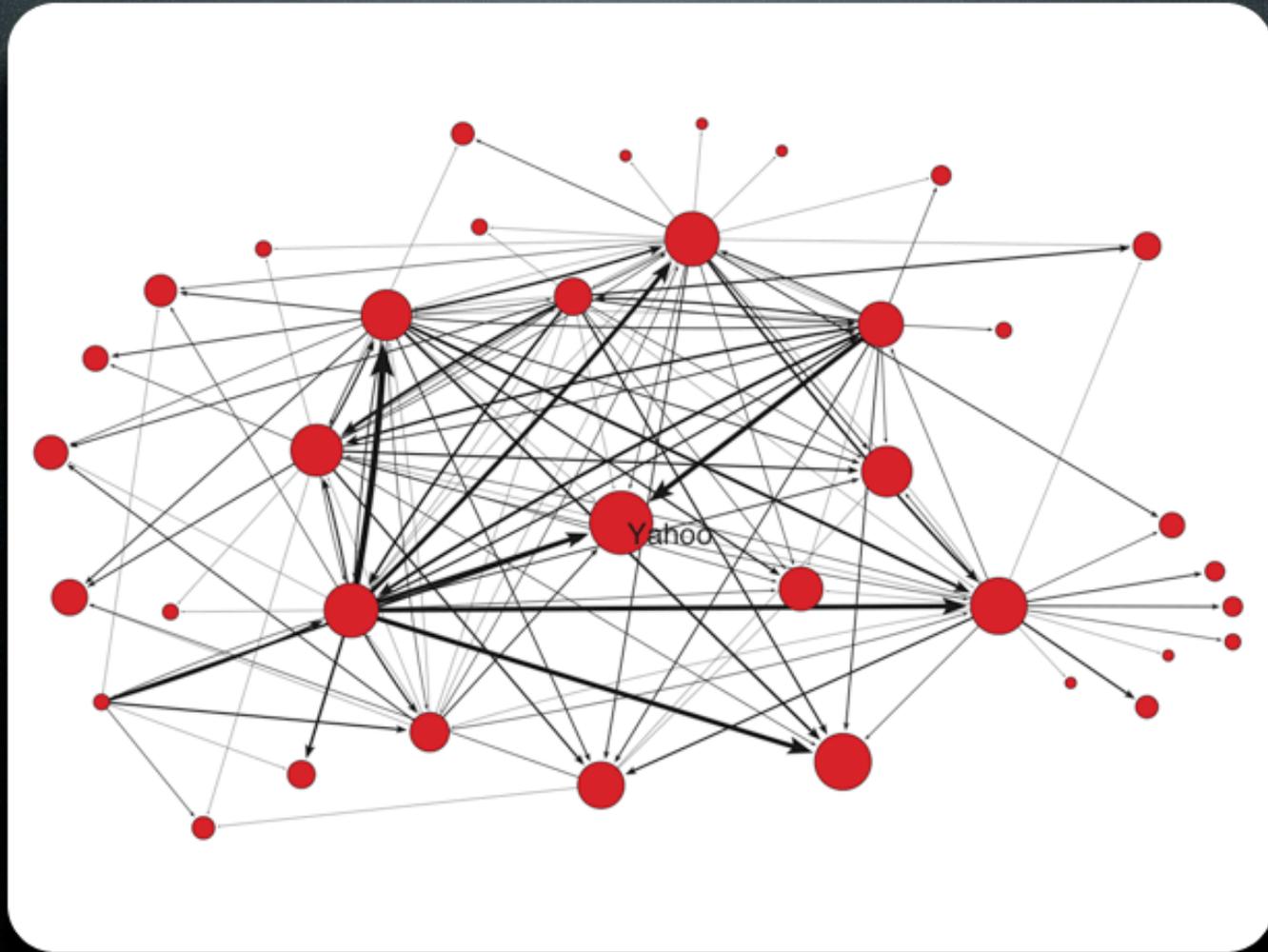


# Ongoing Work

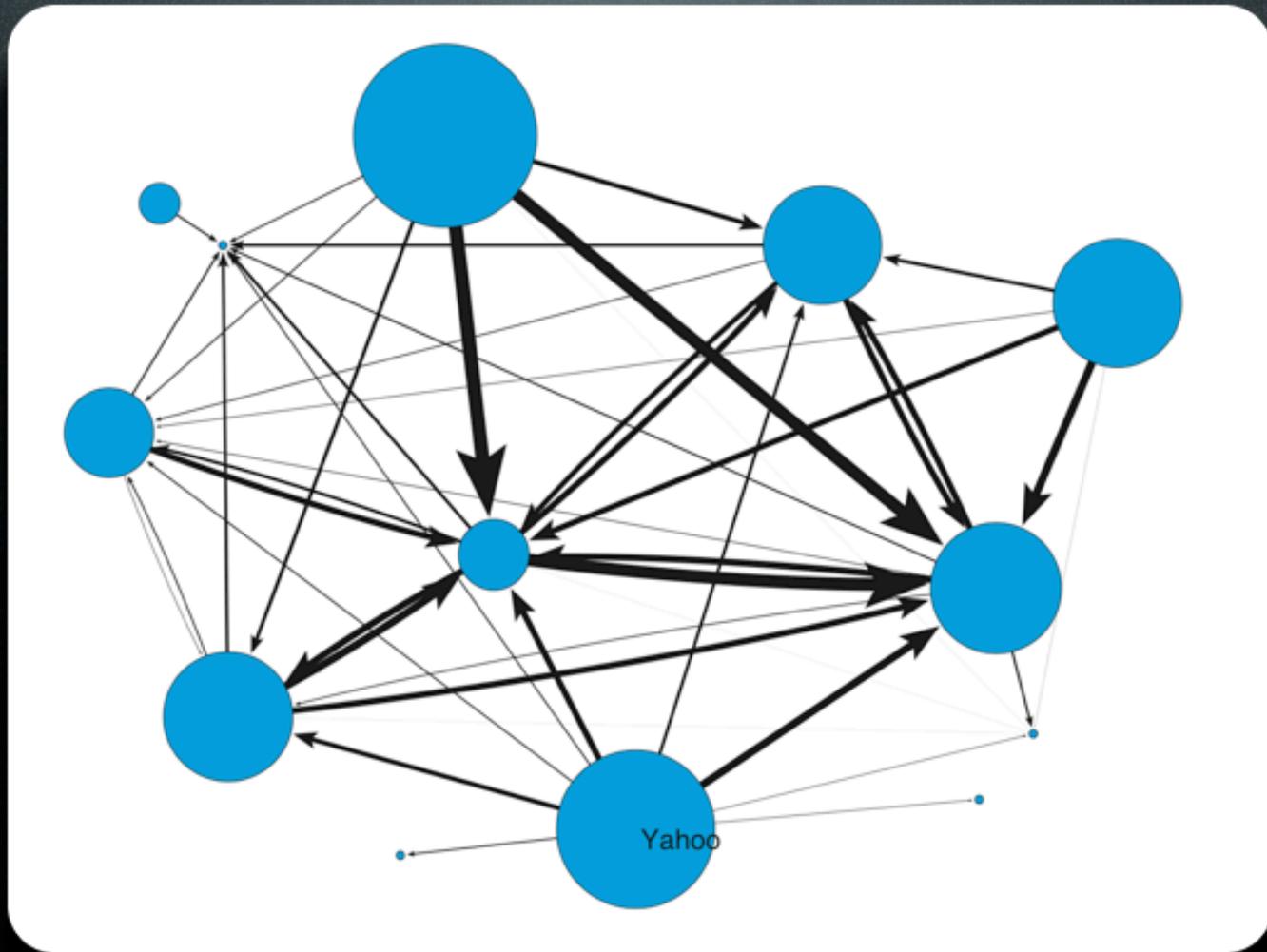
- Improve coverage/diversity in query routing algorithm
- Spam protection: trust/reputation subsystem
- User study with 6S application



User study



Query network

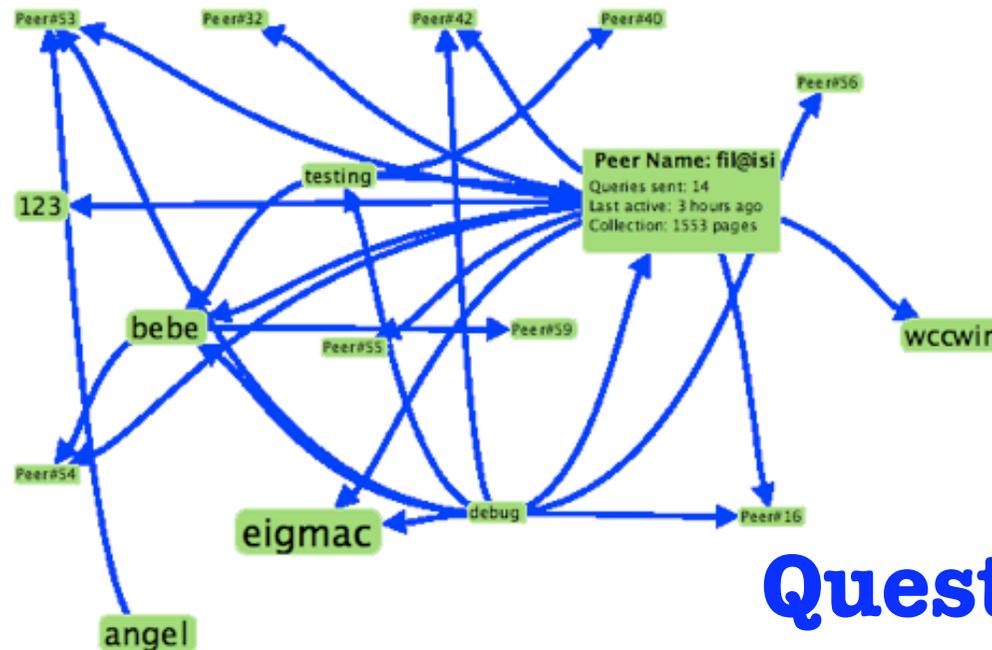


Result network

## 6S: P2P Web index collecting and sharing System

[Home](#) [Members](#) [Download](#) [Publications](#) [Getting Started](#) [Network Viewer](#) [Forum](#)

Mouse over a node to see peer information. Mouse over an edge to see queries. Click on the background for zooming to fit. Scroll or hold right mouse button for zooming. Drag the background for panning. The graph is generated by [Prefuse](#).



# Questions?

Indiana University Computer Science Department  
Comments: {rakavipa, lewu} at cs dot indiana dot edu.  
© Indiana University Page Last Updated: September 20, 2007

This material is based upon work supported by the National Science Foundation under award N. IIS-0133124 and IIS-0348940. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.



Thank you!  
Questions?

[informatics.indiana.edu/fil](http://informatics.indiana.edu/fil)

Indiana University School of  
**informatics**



Research supported by NSF  
CAREER Award IIS-0348940