Web mapping, modeling, mining, and mingling

Tutorial for the 2006 MITACS Winter School
Modelling and Mining of Networked Information Spaces

Filippo Menczer
Informatics & Computer Science
Indiana University, Bloomington

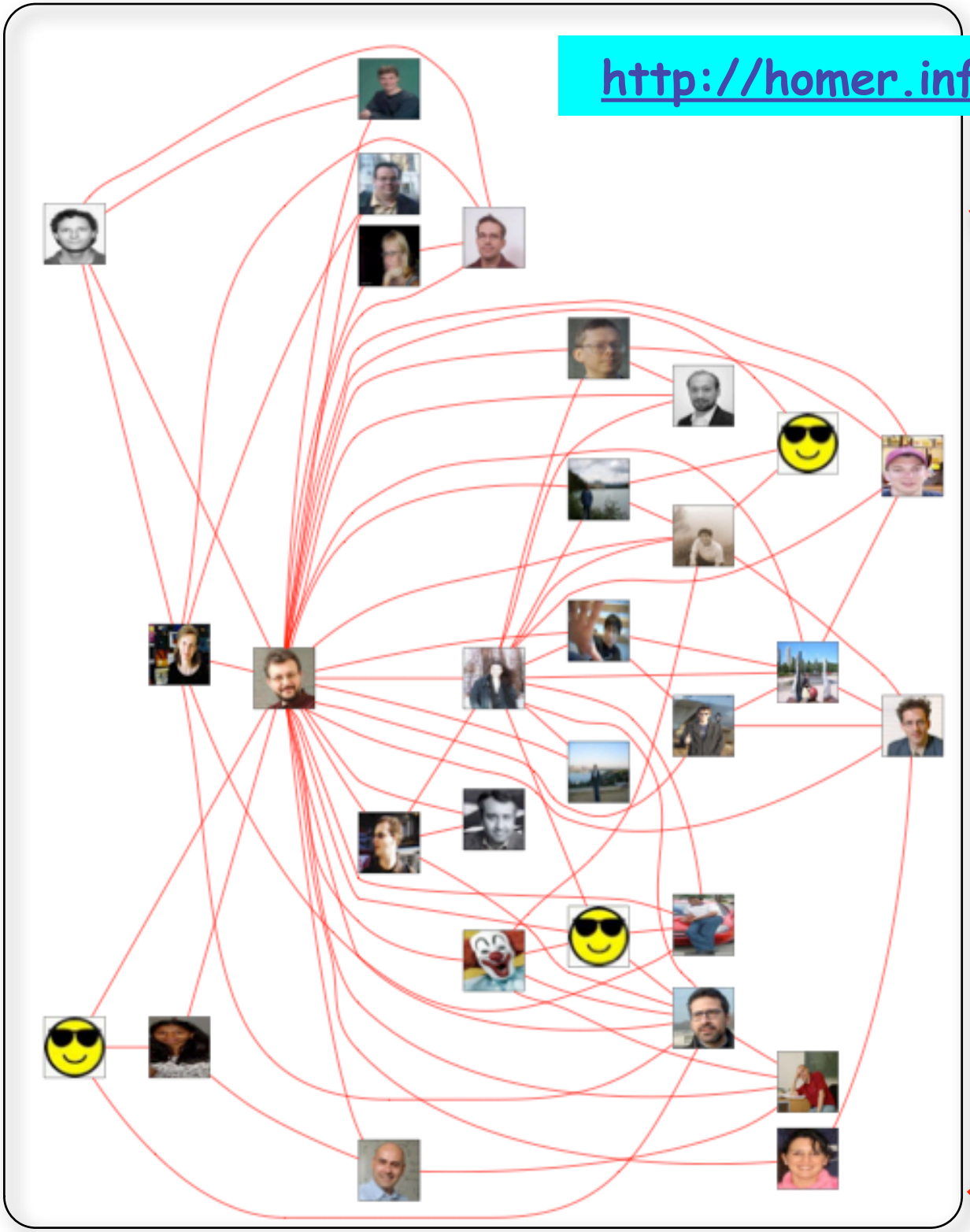
Indiana University Department of ComputerScience

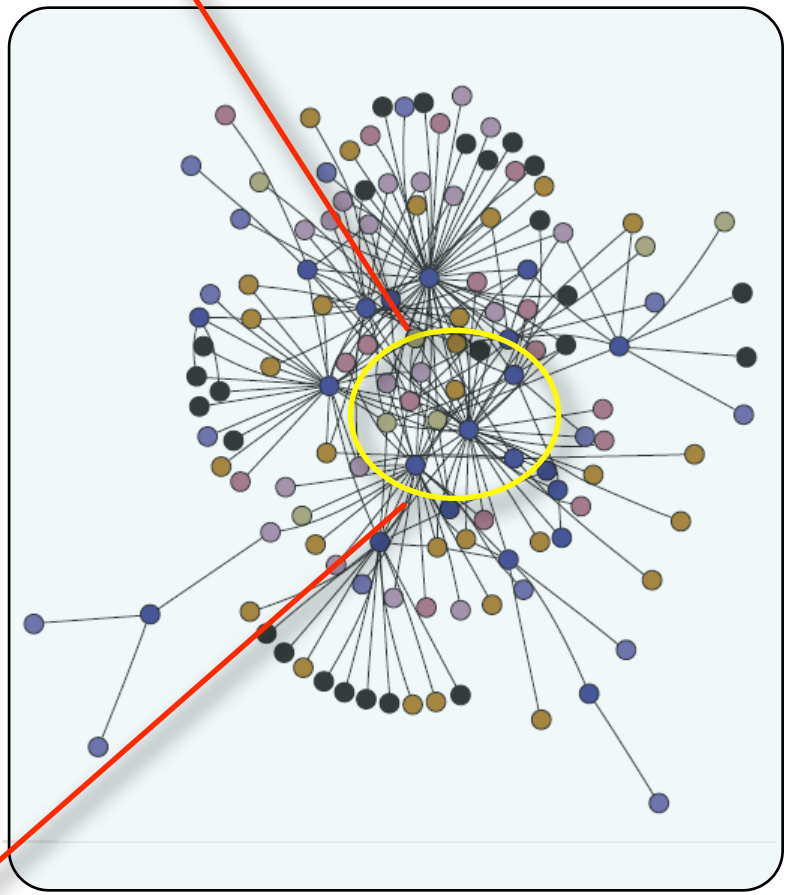
Indiana University School of informatics

http://homer.informatics.indiana.edu/~nan/

NaN:

Networks & agents Network



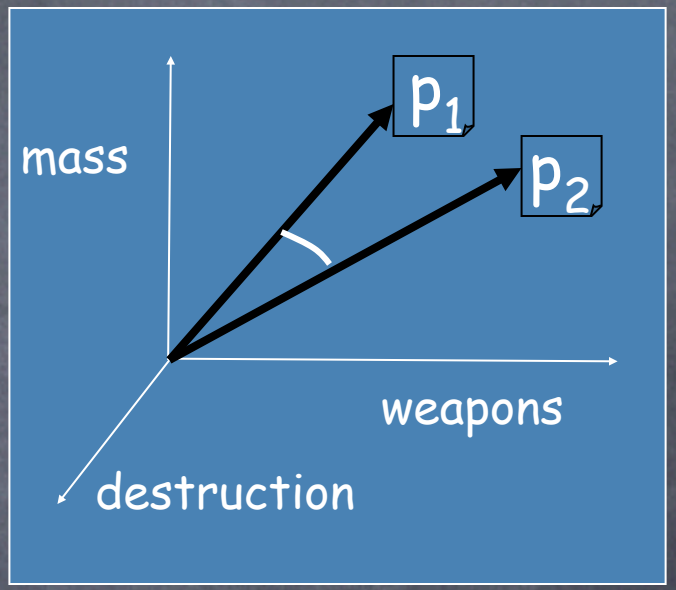


Outline

- Mapping
  - Topical locality
  - Content, link, and semantic topologies in the Web
- Modeling
  - How the Web evolves and why content matters
  - Consequences for navigation and crawling
- Mining
  - Topical Web crawlers
  - Adaptive, intelligent crawling techniques
- Mingling
  - Social Web search & recommendation
  - Distributed collaborative peer search

The Web as a text corpus

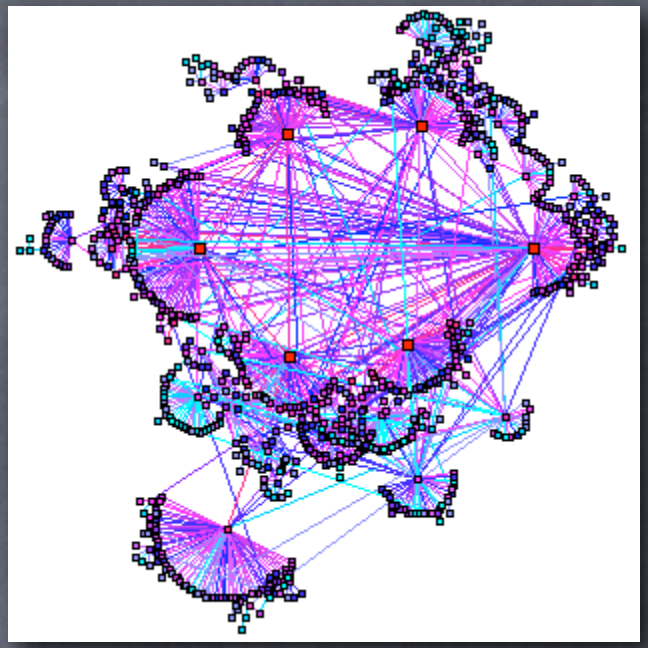
- Pages close in word vector space tend to be related
  - Cluster hypothesis (van Rijsbergen 1979)
  - The WebCrawler (Pinkerton 1994)
  - The whole first generation of search engines



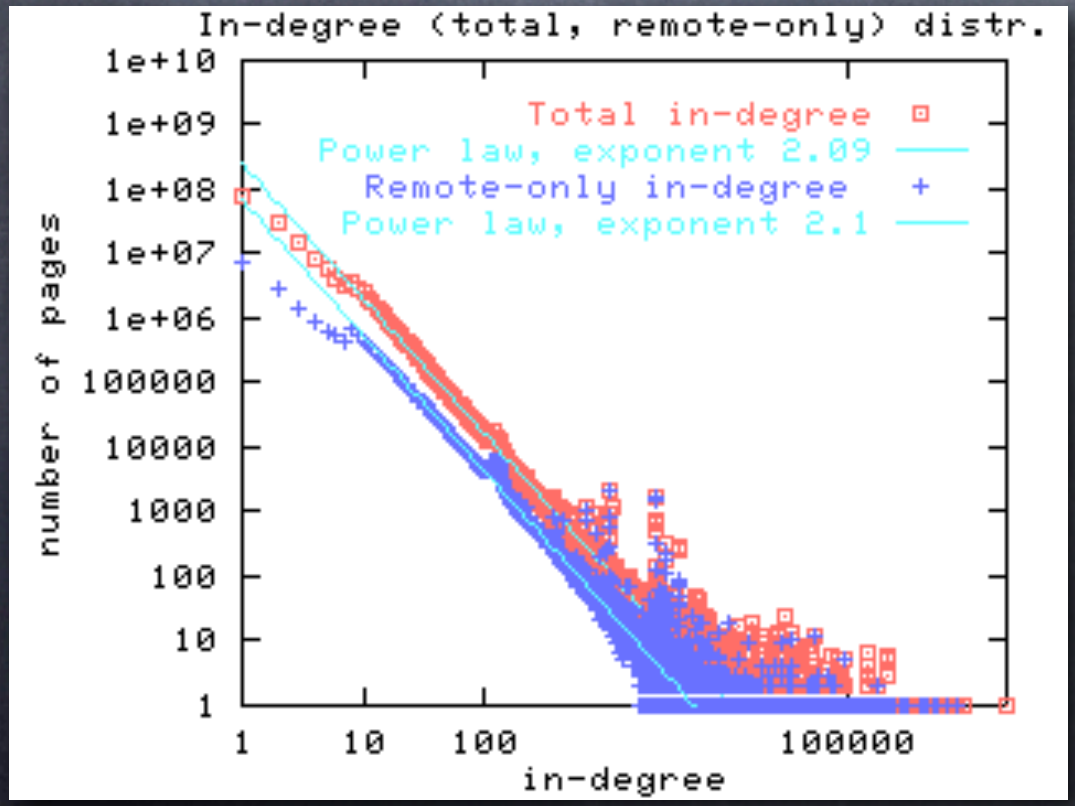
Enter the Web's link structure

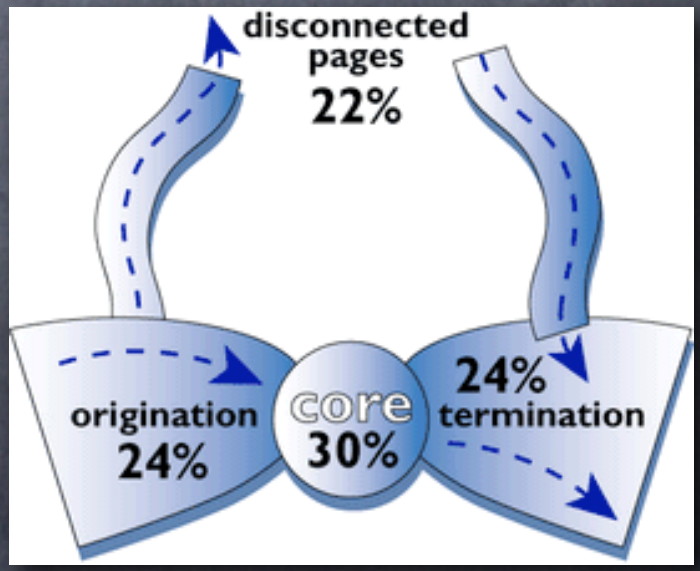
$$p(i) = \frac{\alpha}{N} + (1-\alpha) \sum_{j:j\to i} \frac{p(j)}{|\ell : j \to \ell|}$$

Brin & Page 1998



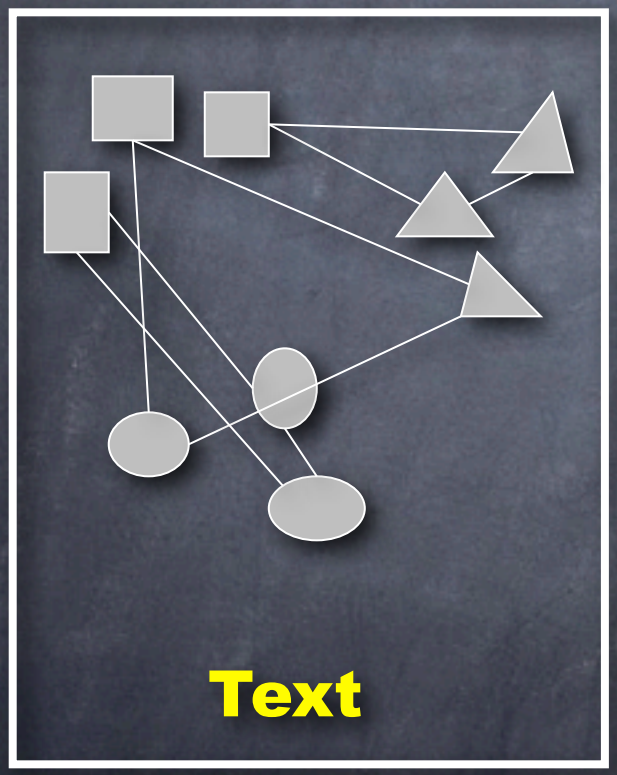
Barabasi & Albert 1999

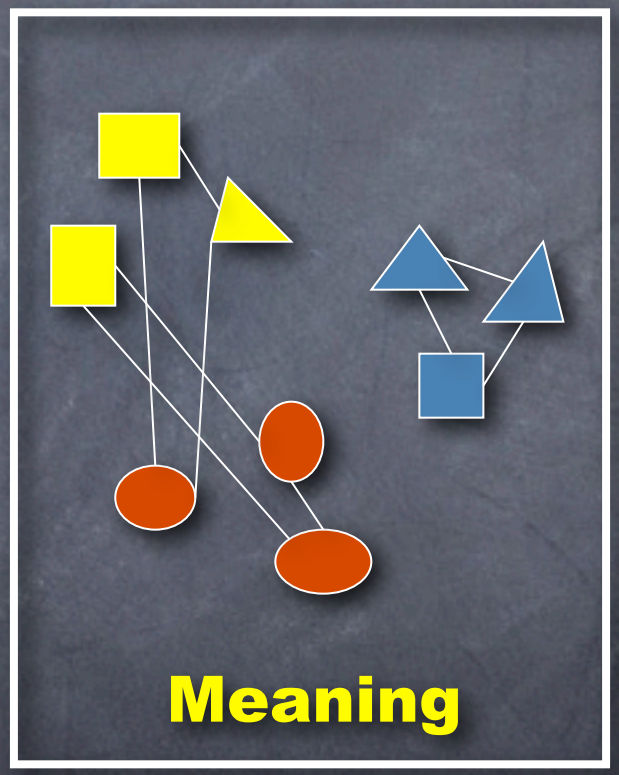


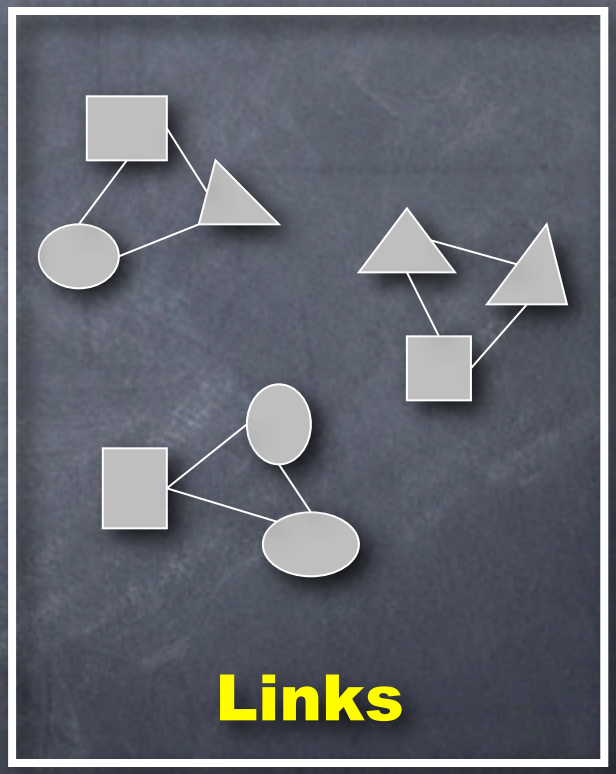


Broder & al. 2000

Three network topologies

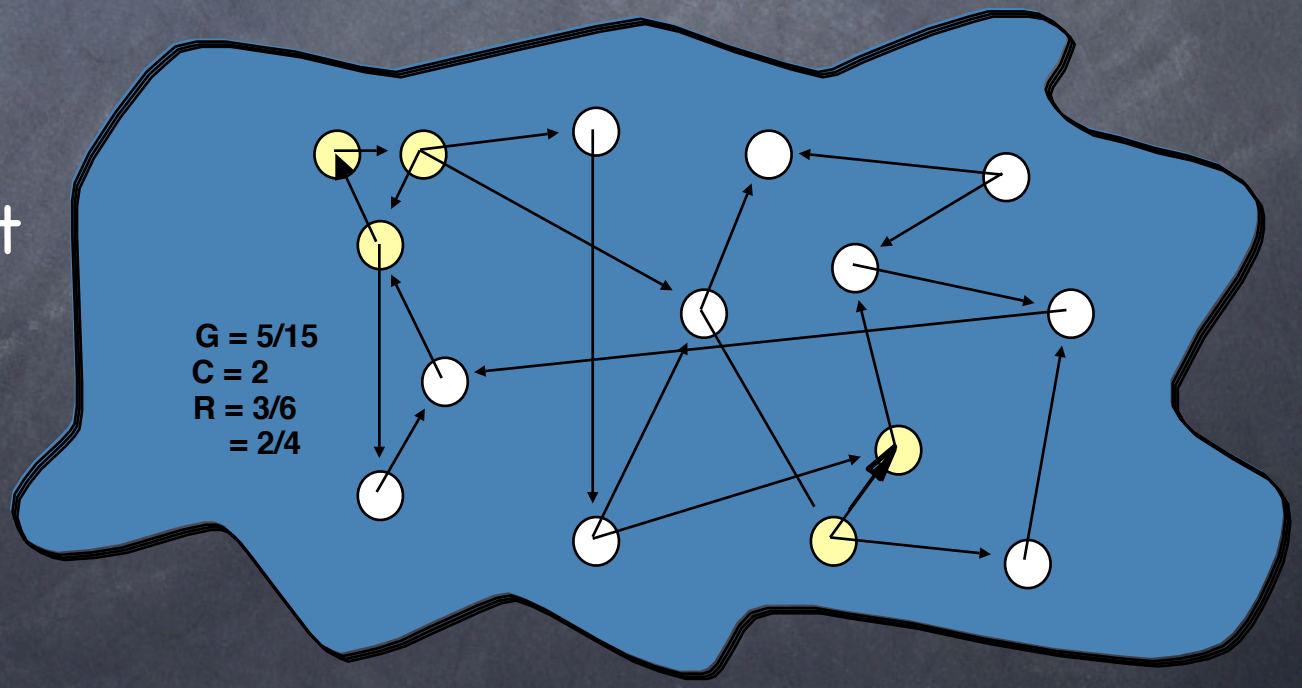






The "link-cluster" conjecture

- 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

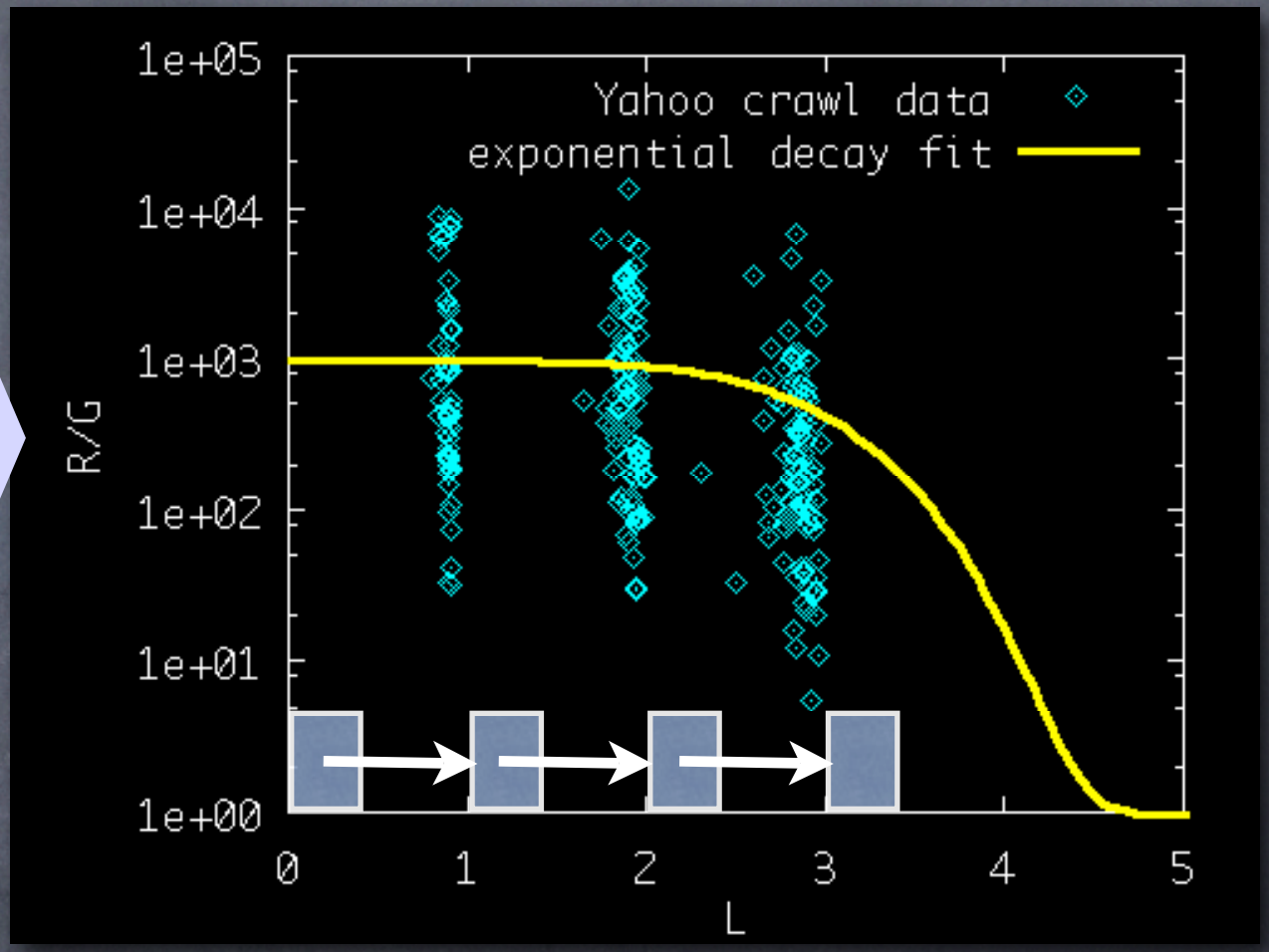
$$\frac{\eta^*}{G} - 1 = \frac{R - G}{1 - (R - G)}$$

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



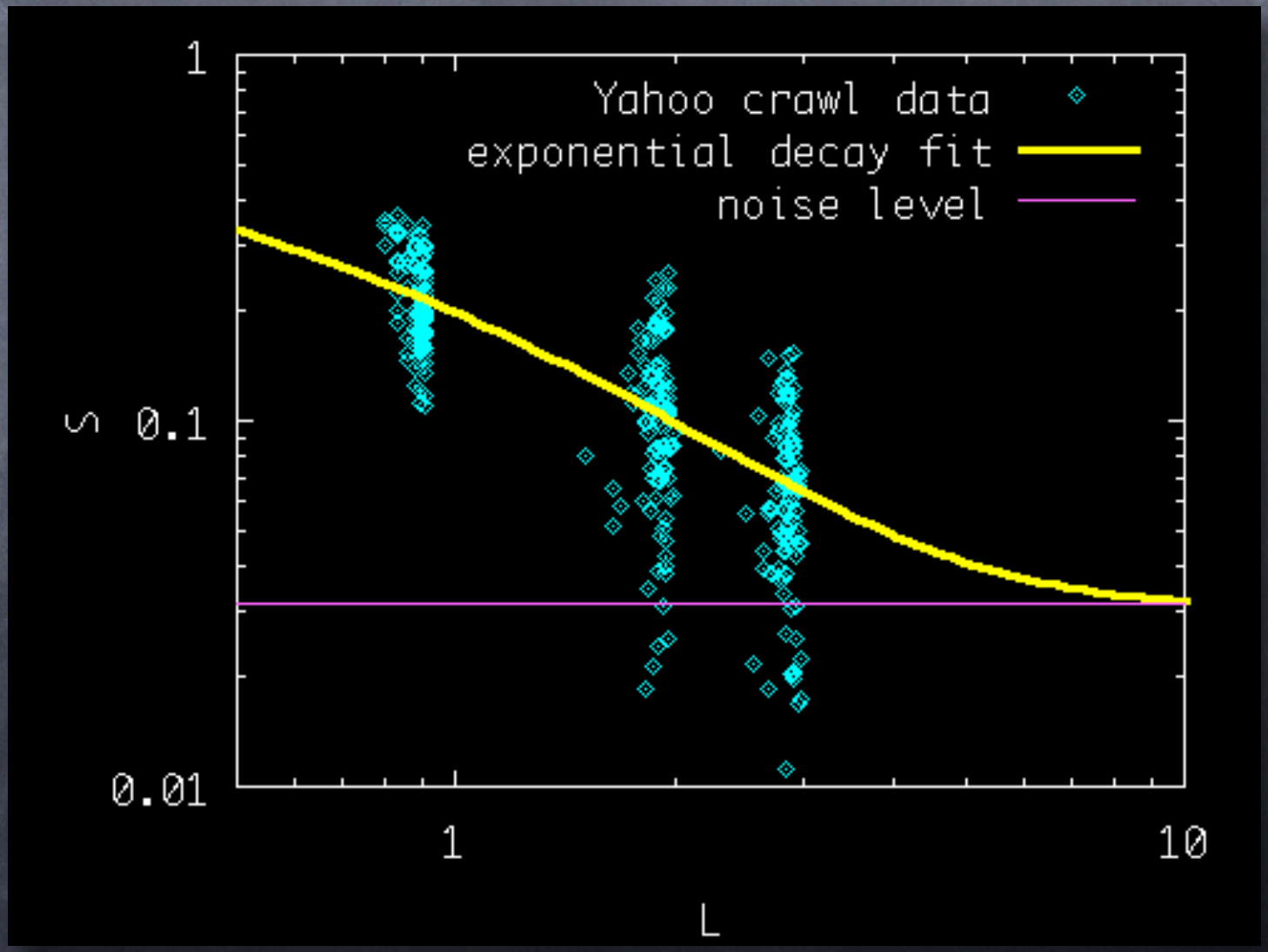
$$L(q,\delta) \equiv \frac{\sum_{\{p:\|path(q,p)\| \le \delta\}} \|path(q,p)\|}{\left|\{p : \|path(q,p)\| \le \delta\}\right|}$$

JASIST 2004

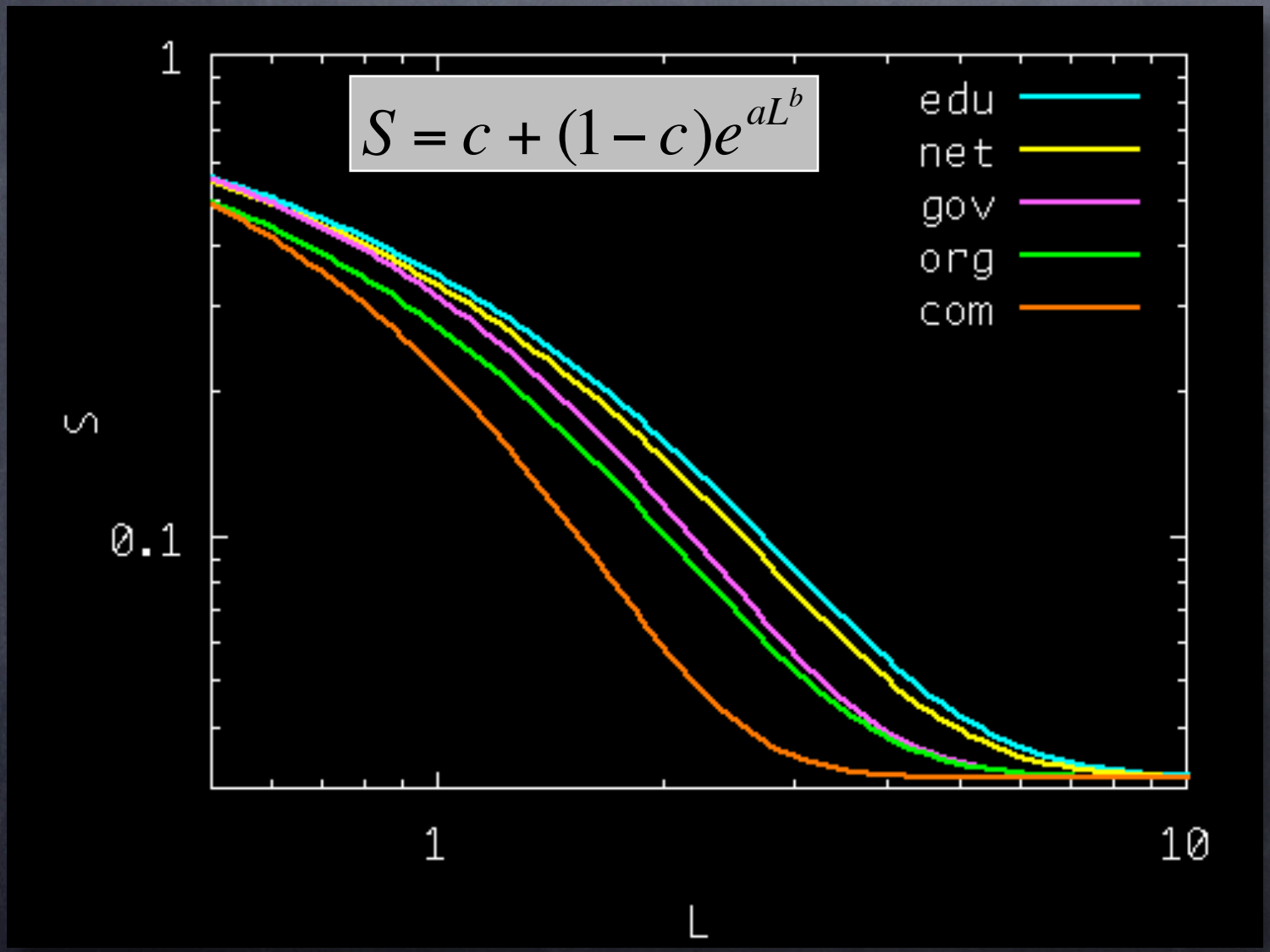
The "link-content" conjecture

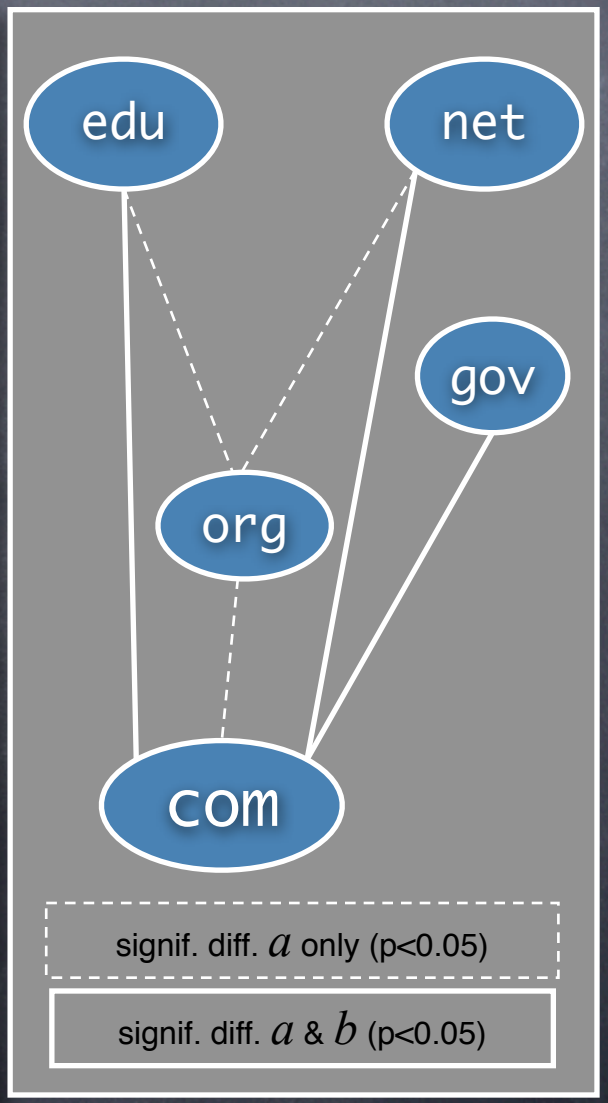
$$S(q,\delta) \equiv \frac{\sum_{\{p:\|path(q,p)\|\le\delta\}} sim(q,p)}{\left|\{p : \|path(q,p)\| \le \delta\}\right|}$$

- 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





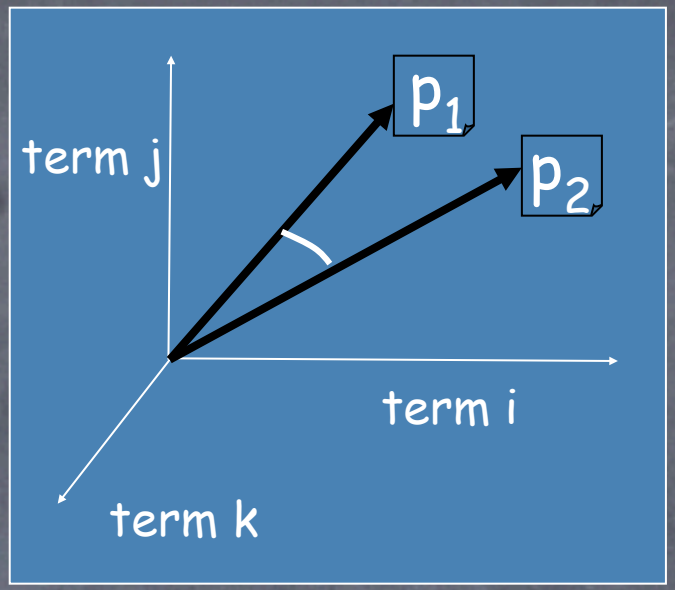
Discussion

Topic drift:
Myth or reality?

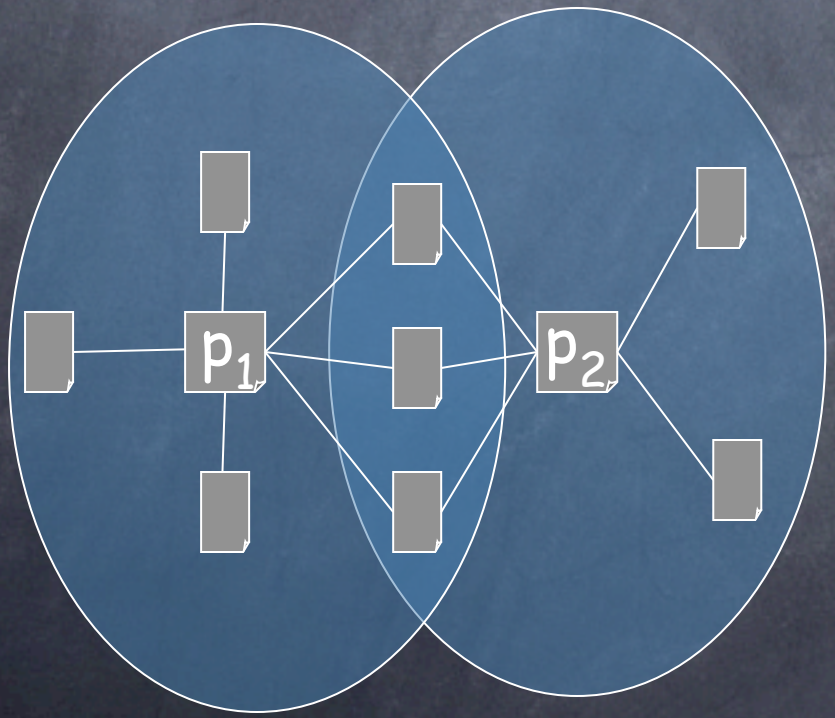
Mapping the relationship between links, content, and semantic topologies

- Given <u>any</u> 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!

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



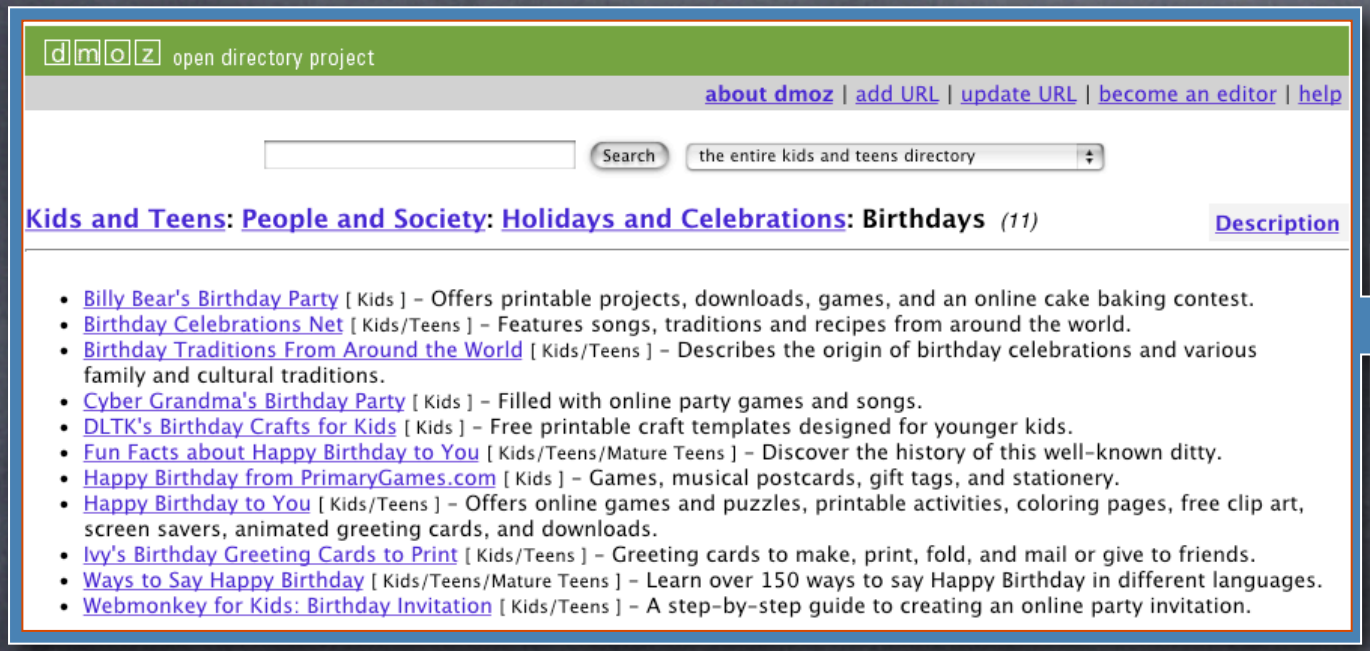
Content similarity

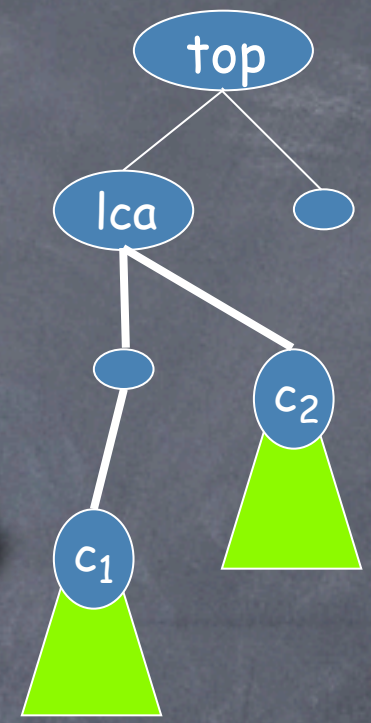


Link similarity

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

Semantic similarity



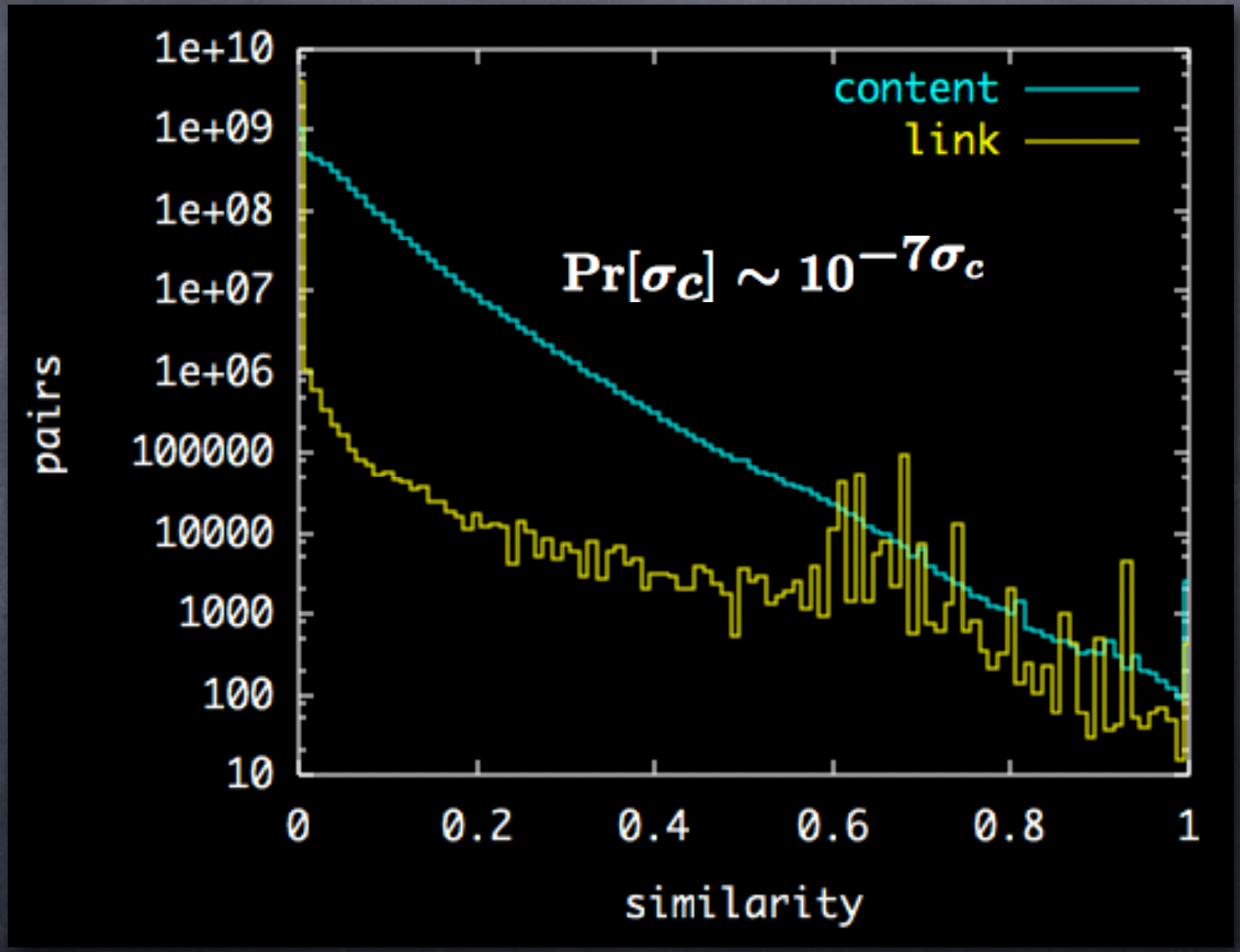


- Information-theoretic measure based on classification tree (Lin 1998)

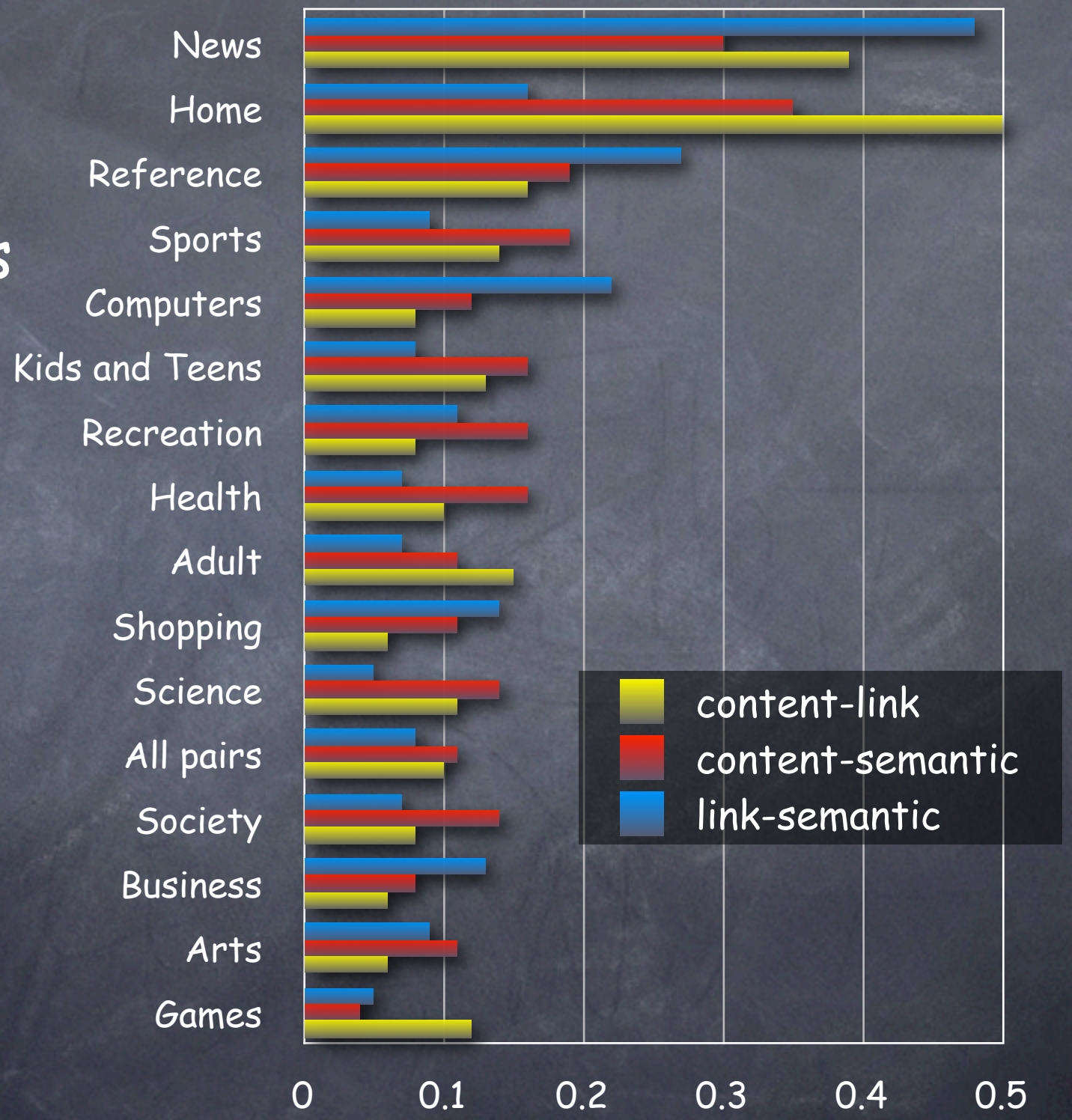
$$\sigma_s(c_1,c_2) = \frac{2\log \Pr[\mathrm{lca}(c_1,c_2)]}{\log \Pr[c_1] + \log \Pr[c_2]}$$

- Classic path distance in special case of balanced tree

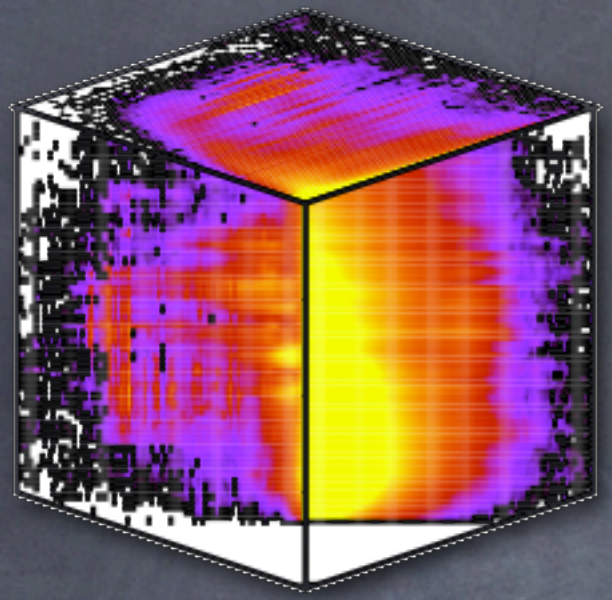
Individual metric distributions



Correlations between similarities



IEEE Internet Computing 2005



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

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

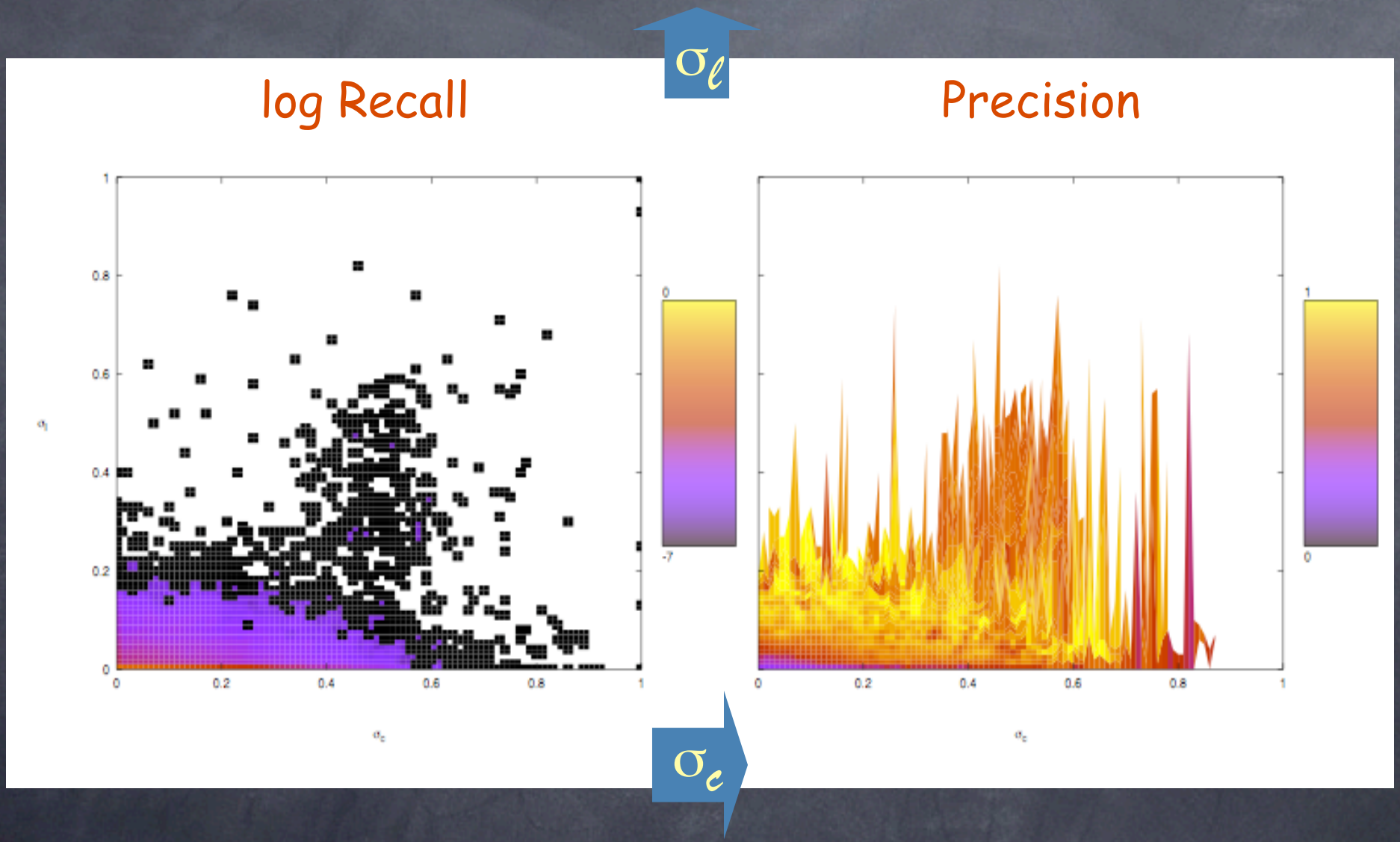
$$P(s_c, s_l) = \frac{\sum_{\{p,q:\sigma_c = s_c, \sigma_l = s_l\}} \sigma_s(p,q)}{\left|\{p,q : \sigma_c = s_c, \sigma_l = s_l\}\right|}$$

Averaging semantic similarity

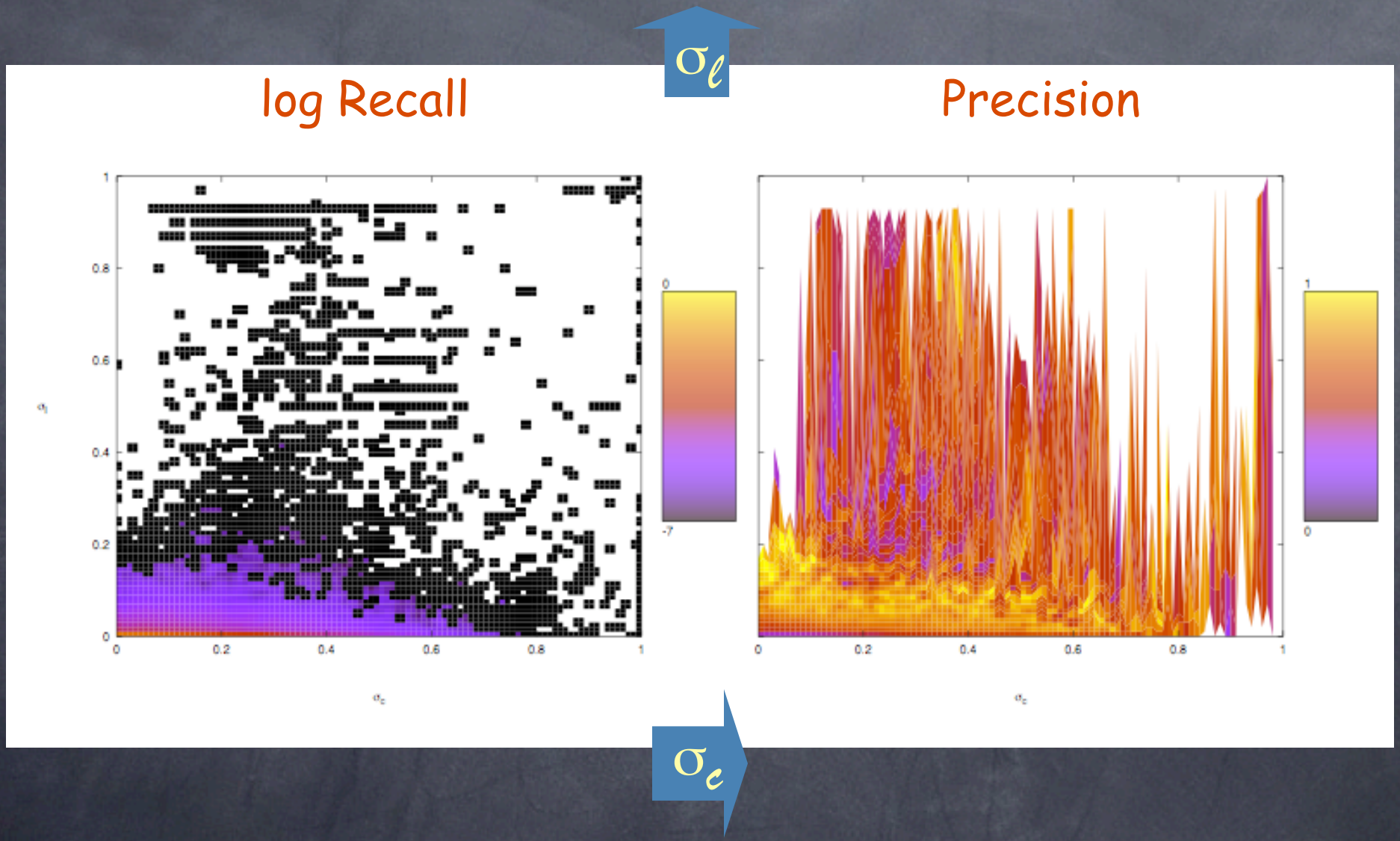
$$R(s_c, s_l) = \frac{\sum_{\{p,q:\sigma_c = s_c, \sigma_l = s_l\}} \sigma_s(p,q)}{\sum_{\{p,q\}} \sigma_s(p,q)}$$

Summing semantic similarity

Business


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

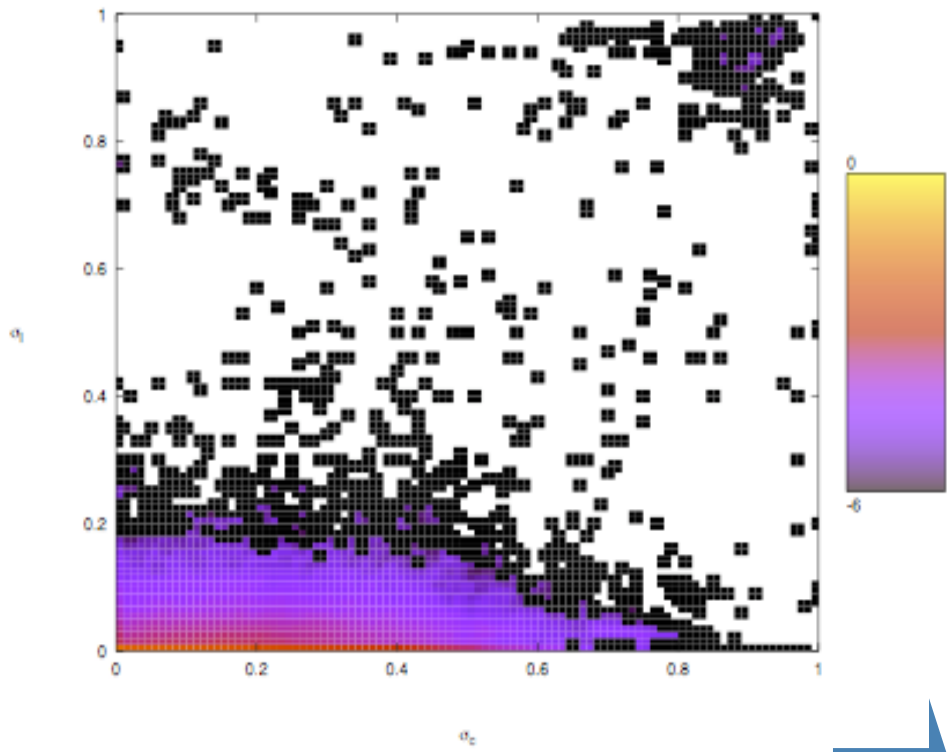
Science



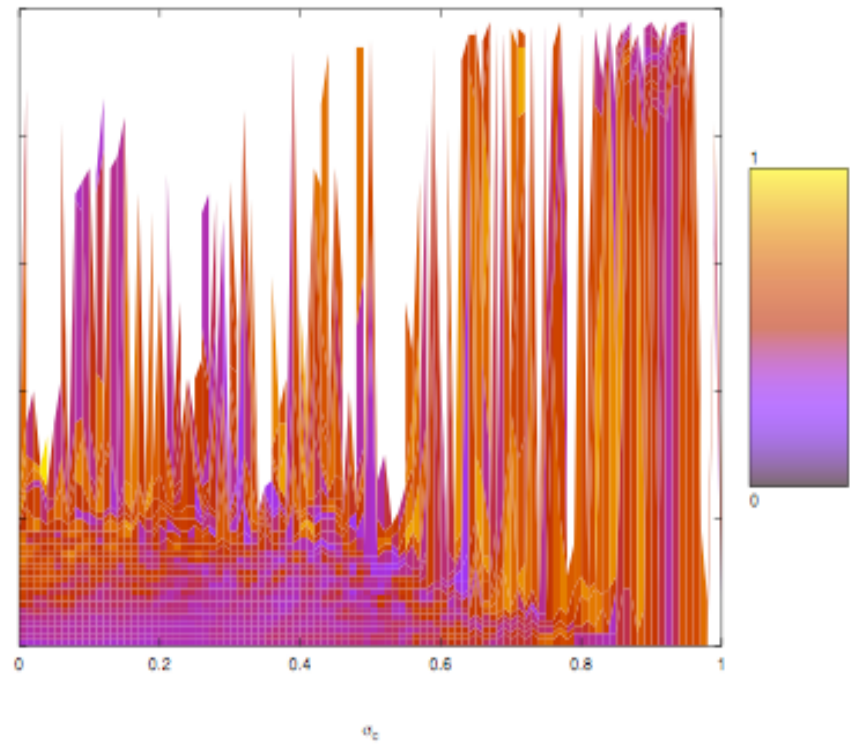
Adult

σ_ℓ

log Recall

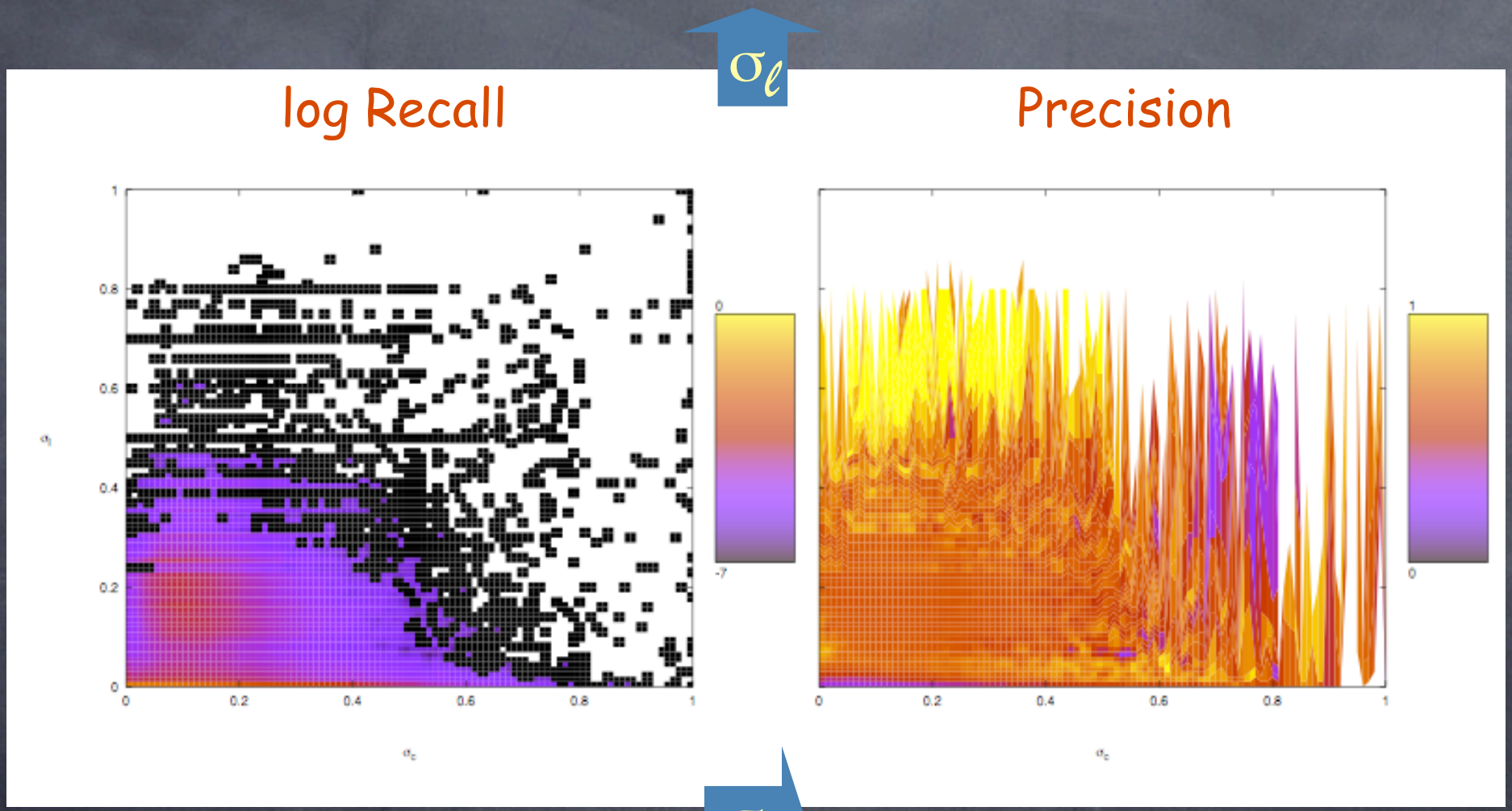


Precision

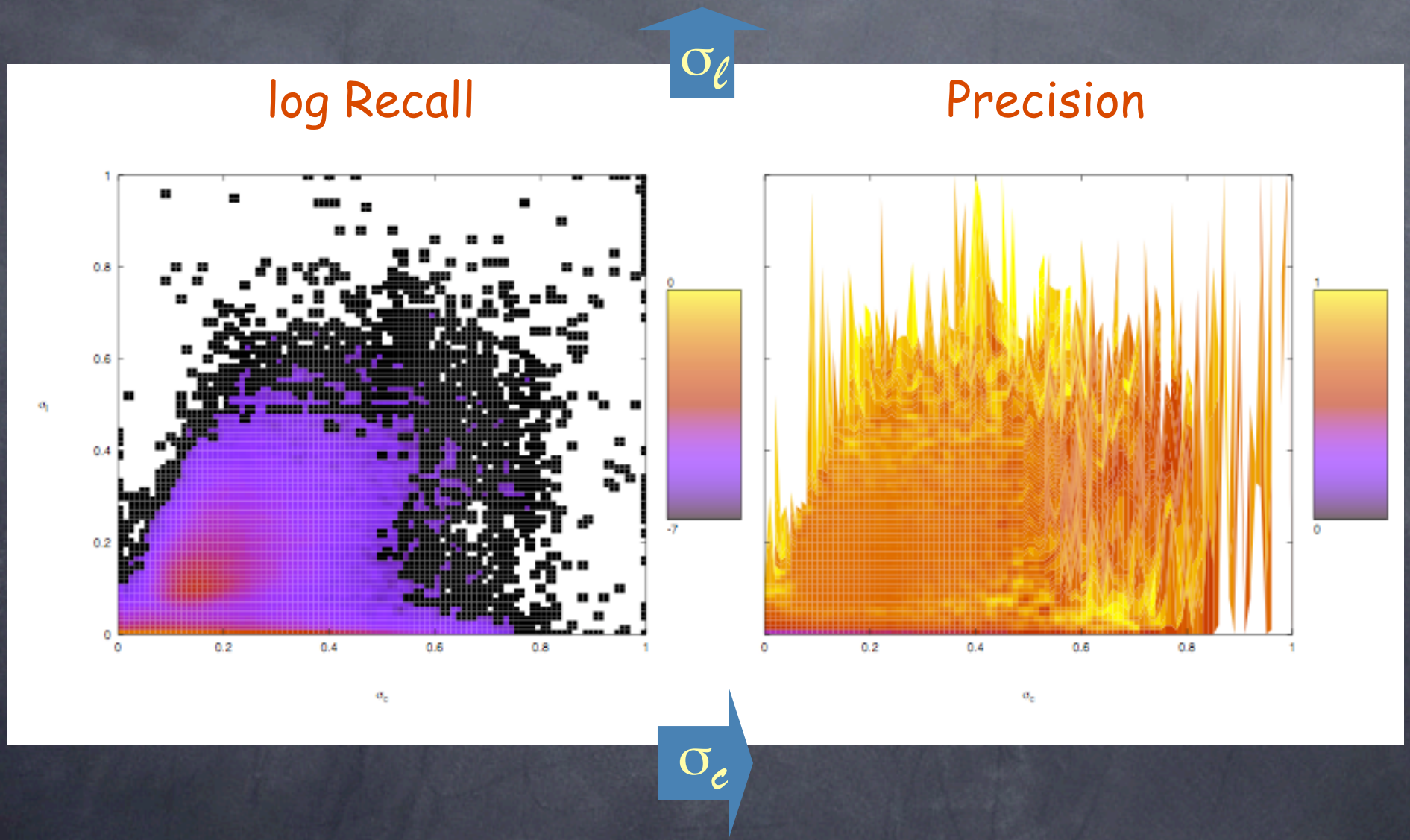


σ_c

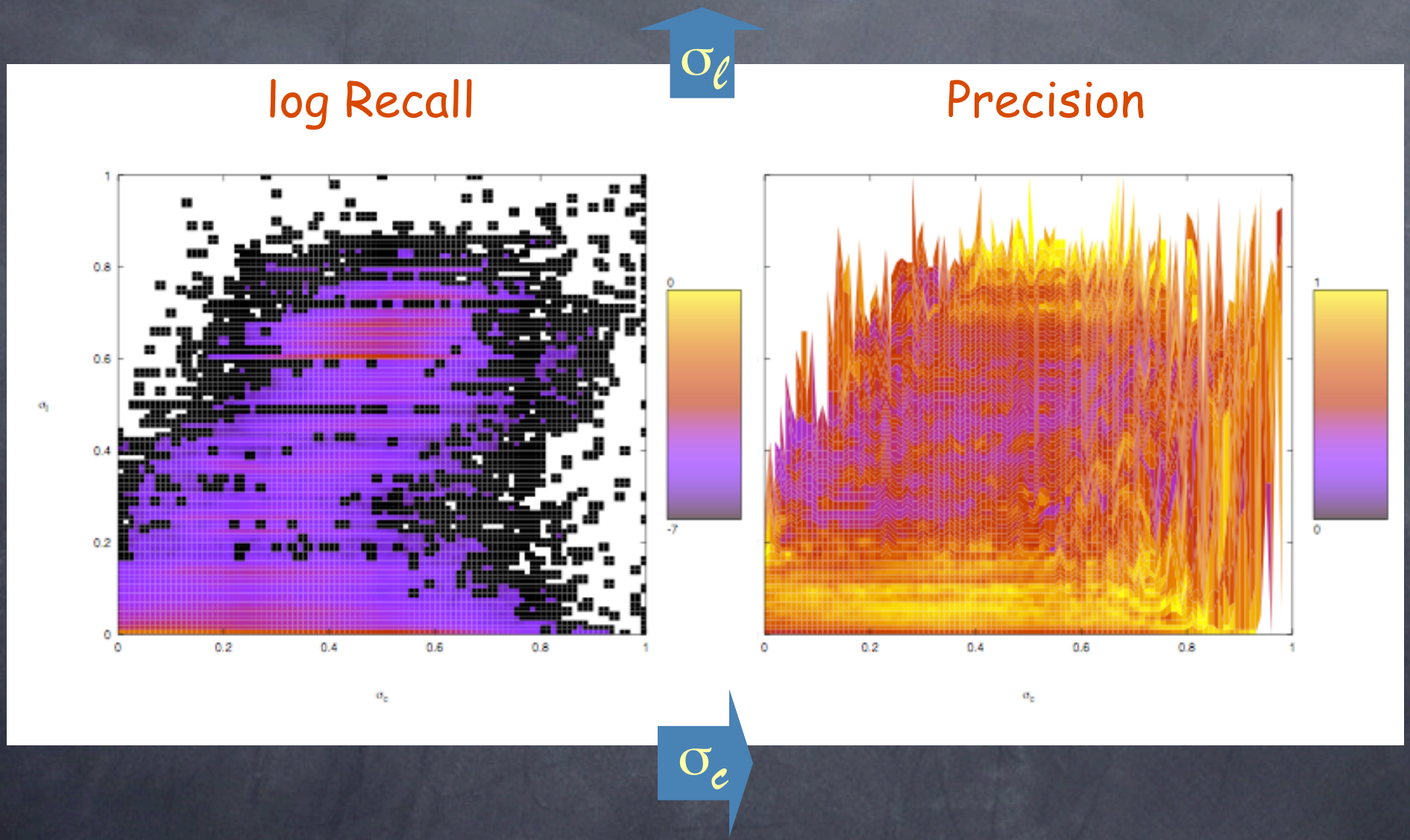
Computers


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

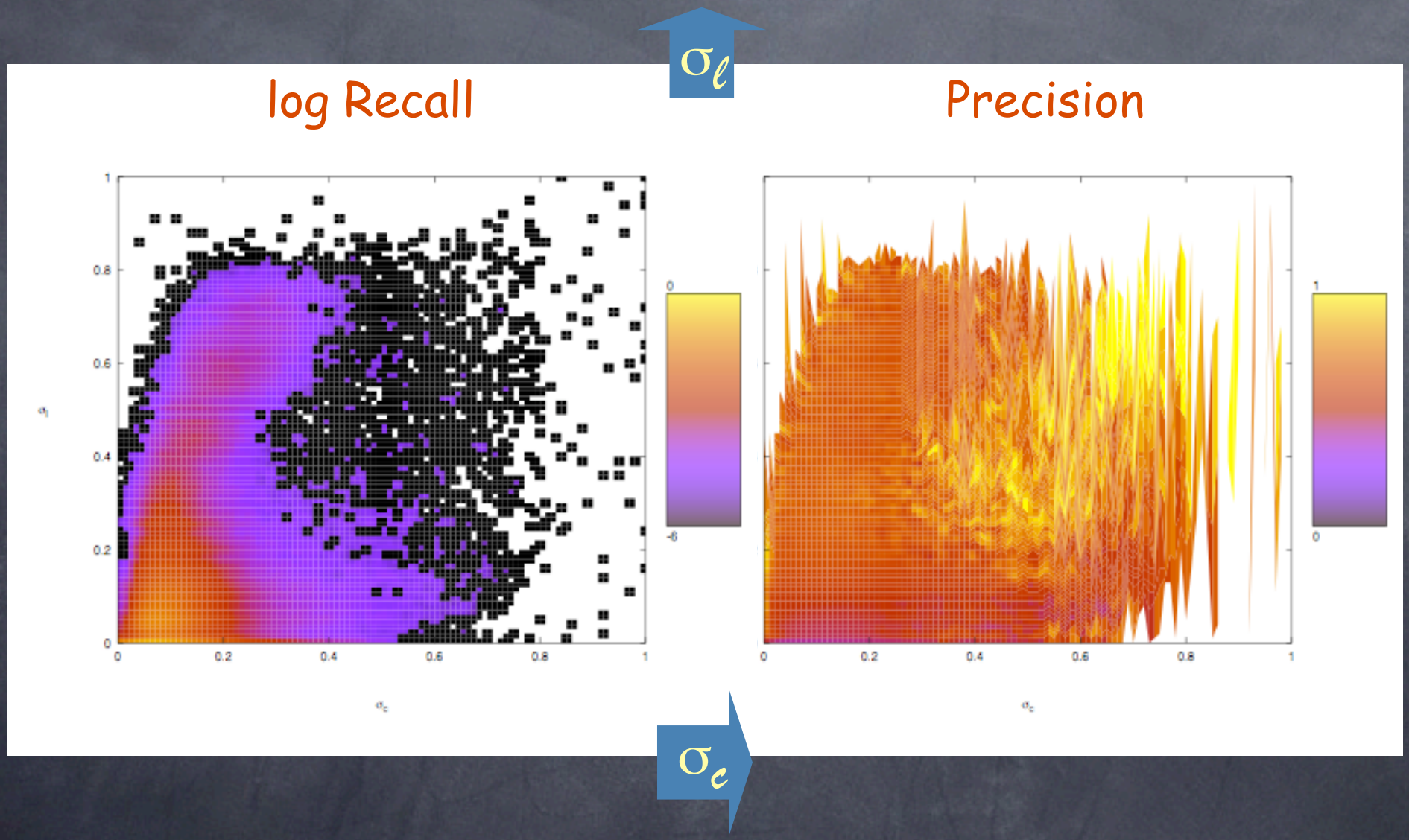
Reference


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

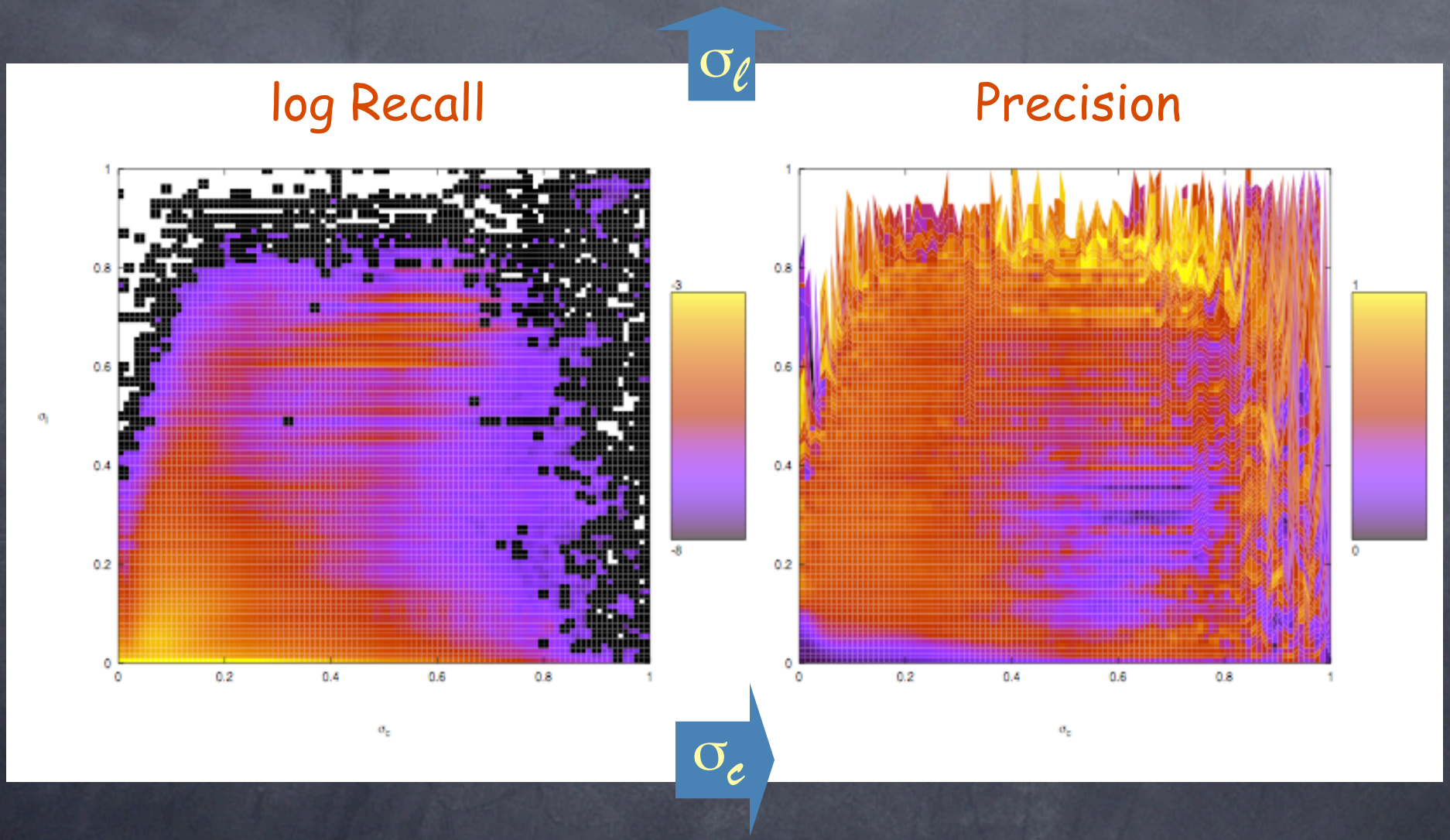
Home


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

News


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

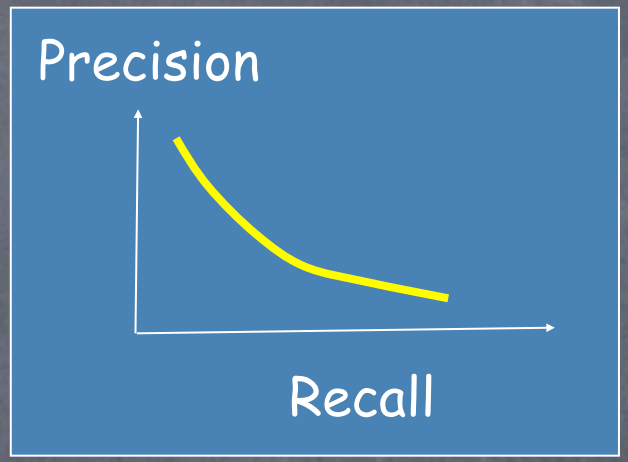
All pairs



Discussion: So what?

So what?

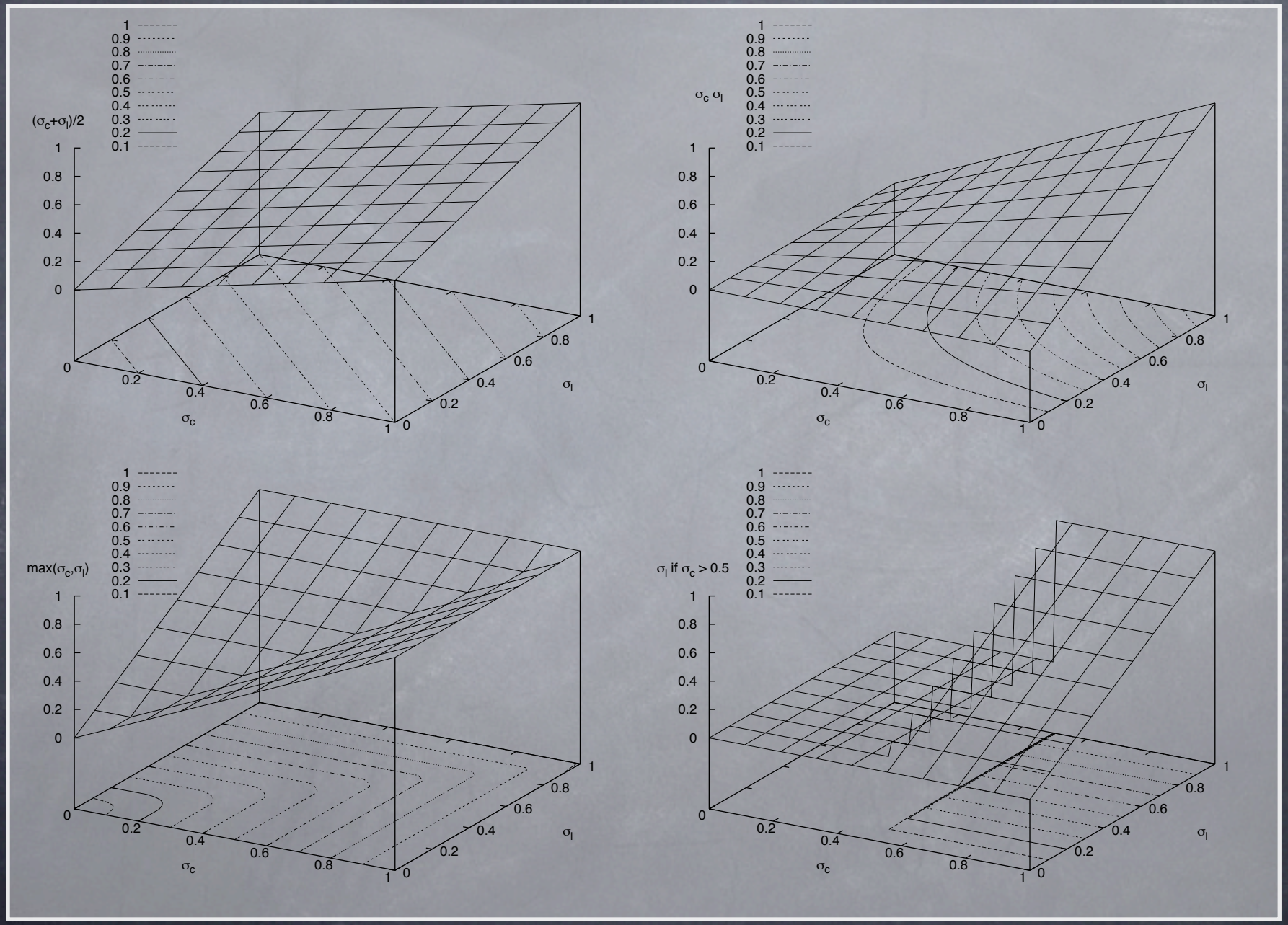
Approximate relevance by semantic similarity



$$P(f,\beta)=\frac{\sum_{p,q:f(\sigma_c(p,q),\sigma_l(p,q))\geq\beta}\sigma_s(p,q)}{|p,q:f(\sigma_c(p,q),\sigma_l(p,q))\geq\beta|}$$

$$R(f,\beta)=\frac{\sum_{p,q:f(\sigma_c(p,q),\sigma_l(p,q))\geq\beta}\sigma_s(p,q)}{\sum_{p,q}\sigma_s(p,q)}$$

Rank by combining content and link similarity



All pairs

$\sigma_l \cdot H(\sigma_c - 0.5)$

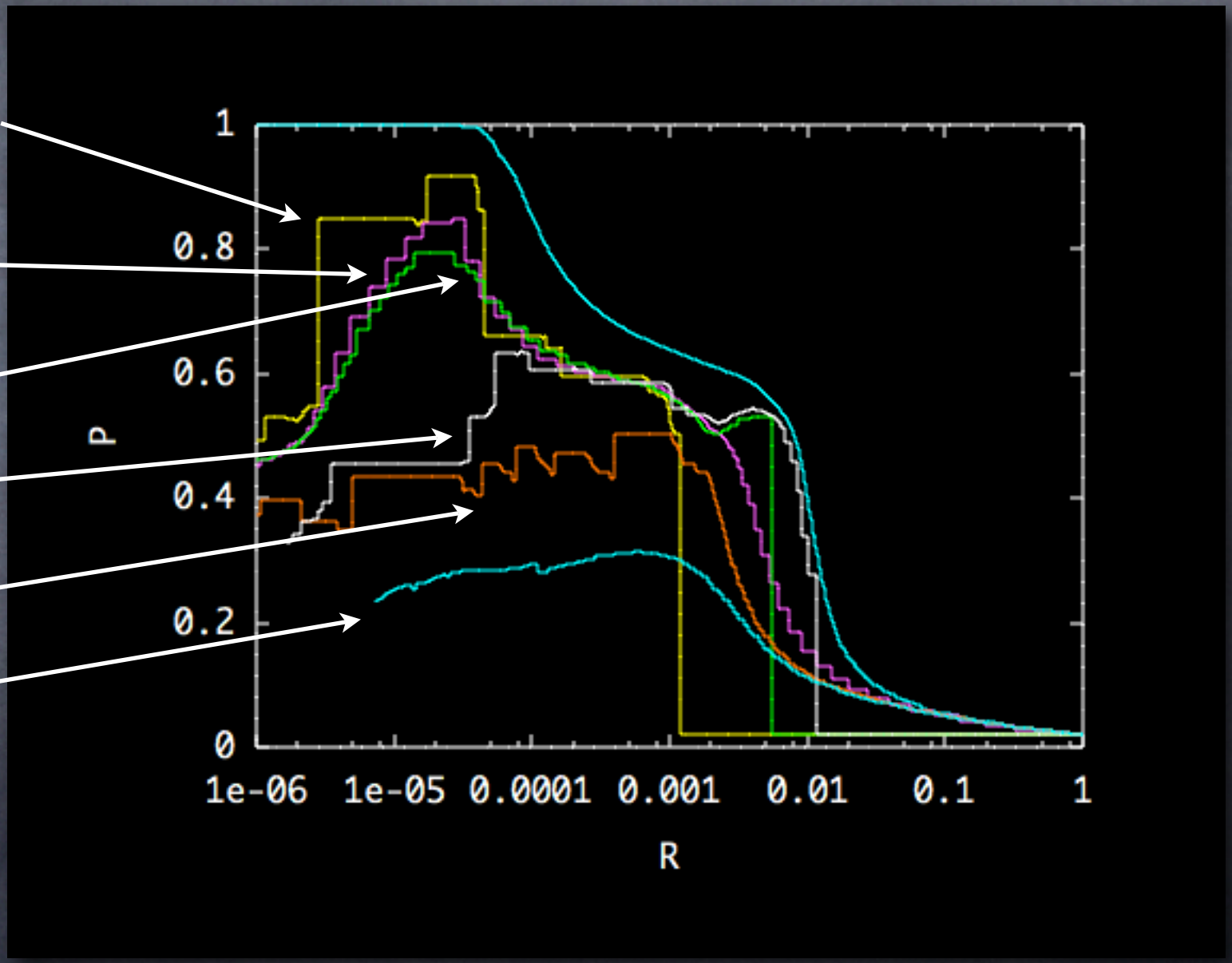
$\frac{2}{3}\sigma_l + \frac{1}{3}\sigma_c$

$\sigma_c \cdot \sigma_l$

σ_l

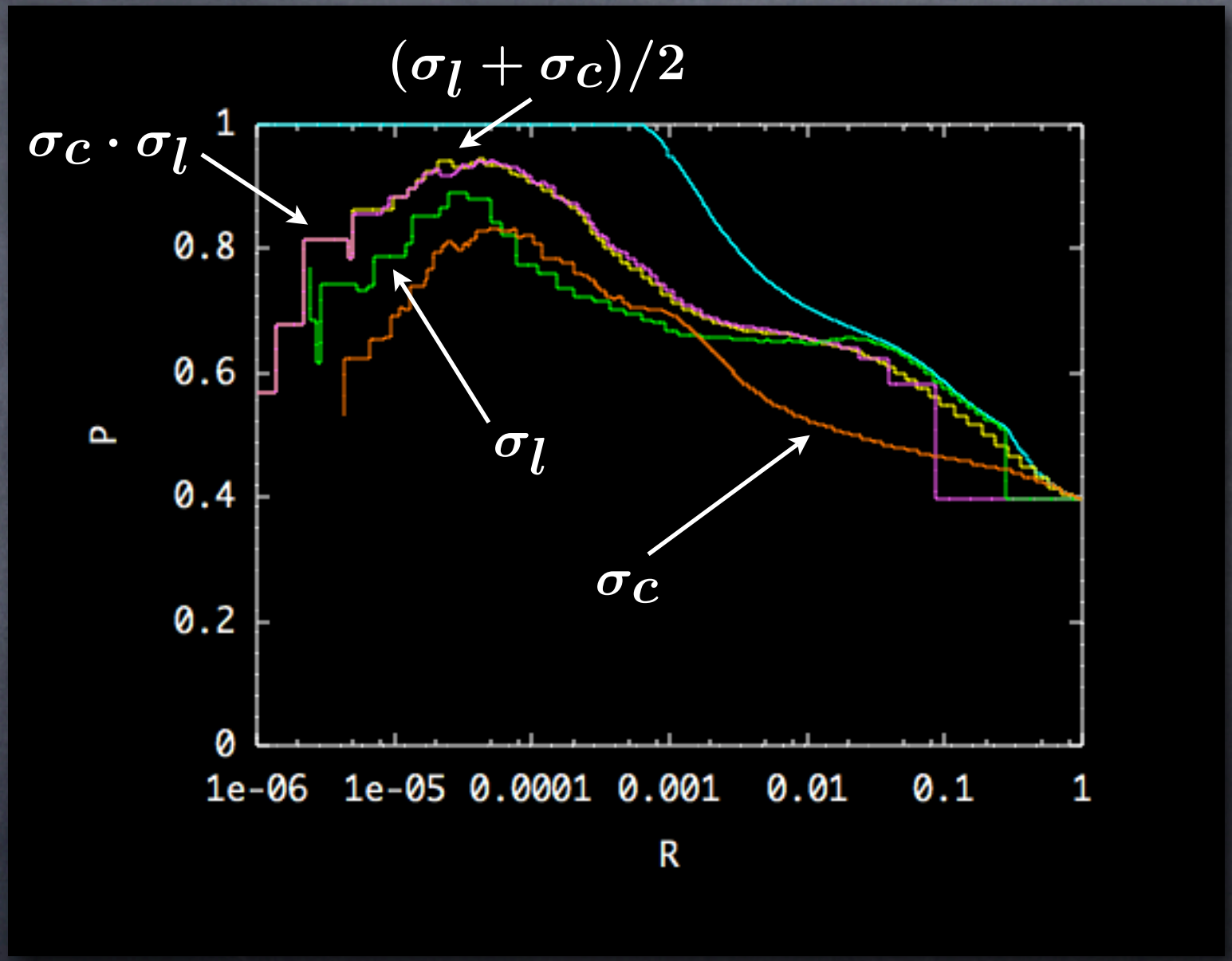
$\max(\sigma_c, \sigma_l)$

σ_c

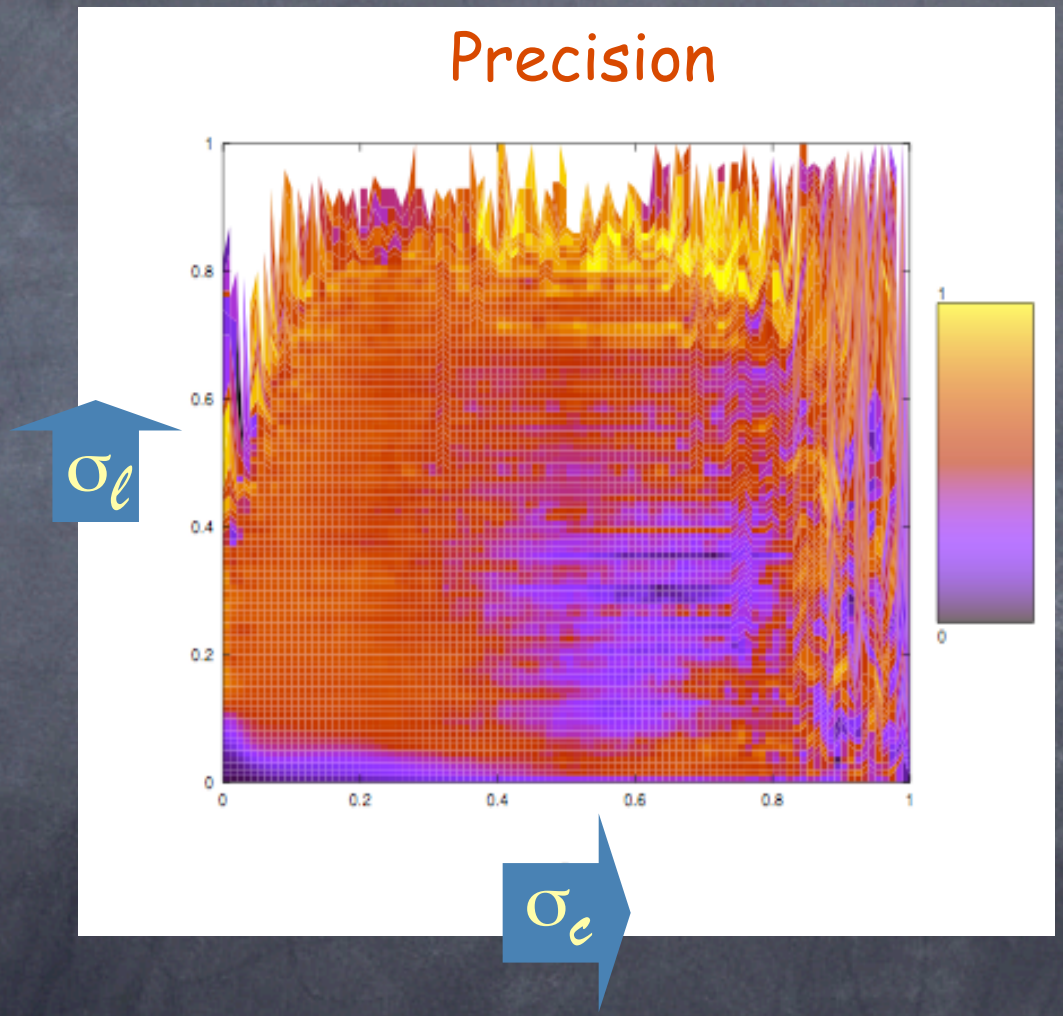


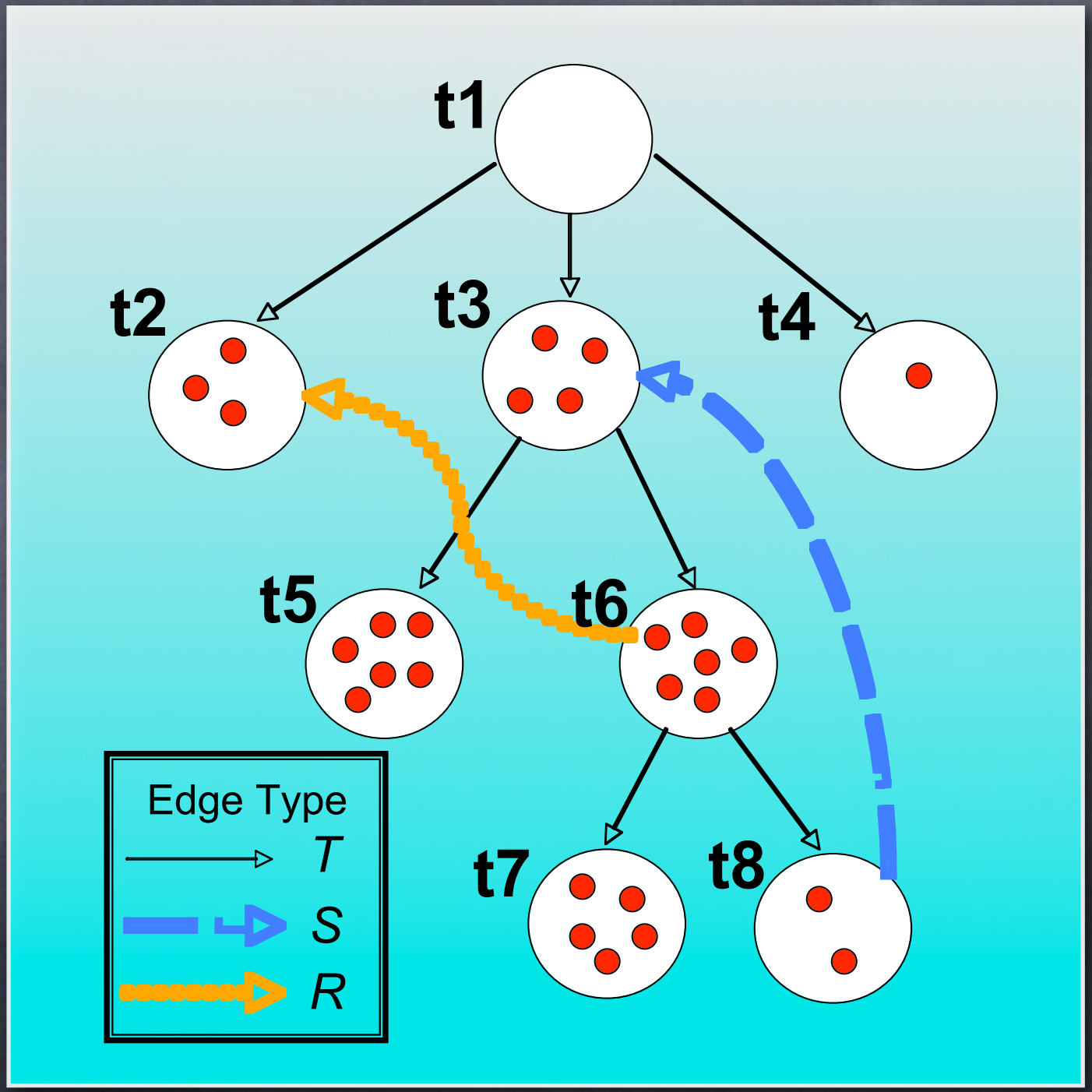
Proc. WWW 2004

News


$(\sigma_l + \sigma_c)/2$
$\sigma_c \cdot \sigma_l$
σ_l
σ_c
1
0.8
0.6
0.4
0.2
0
P
1e-06
1e-05
0.0001
0.001
0.01
0.1
1
R

What about the "hole"?




t1
t2
t3
t4
t5
t6
t7
t8
Edge Type
T
S
R

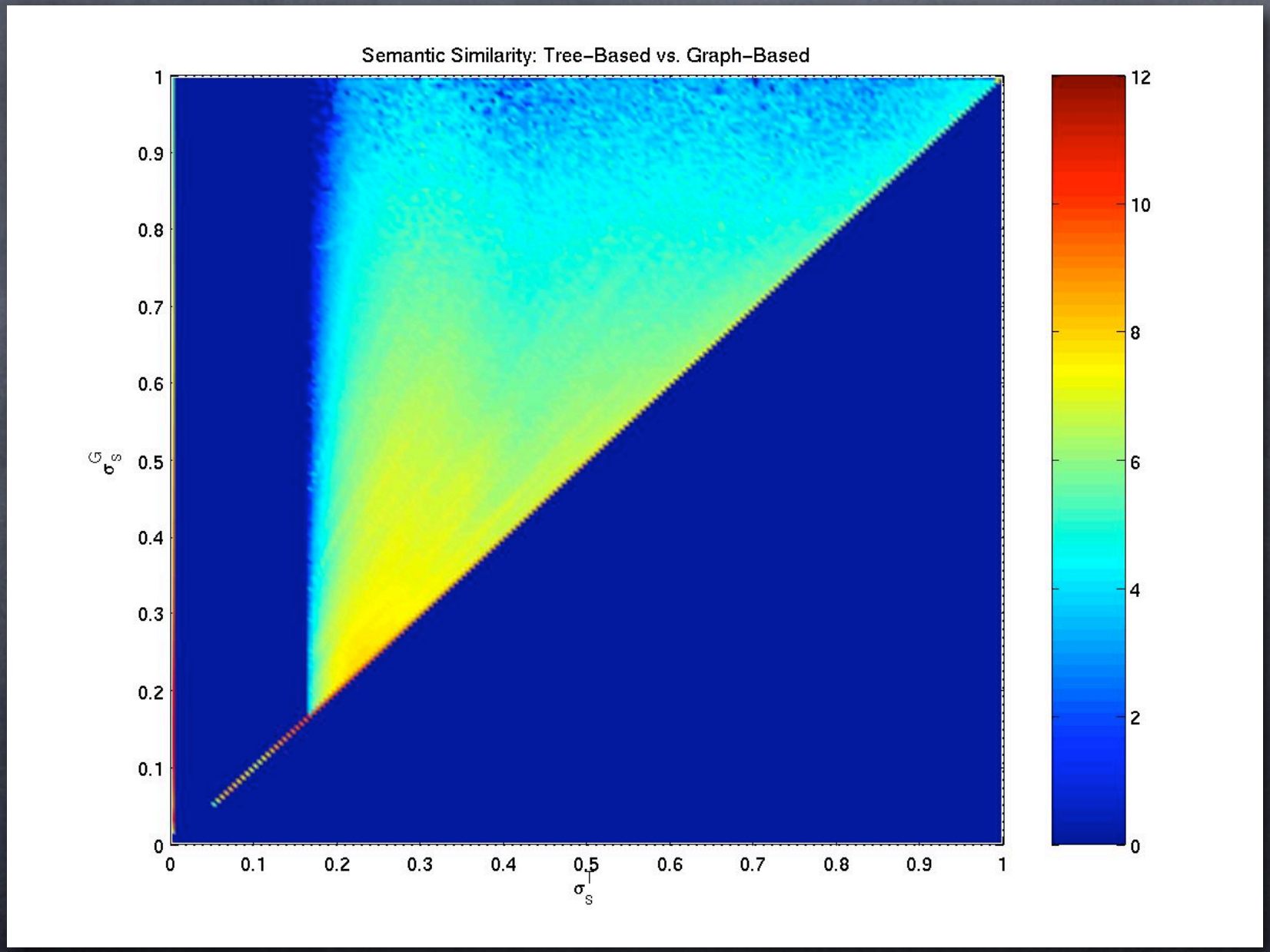
Better semantic similarity measure

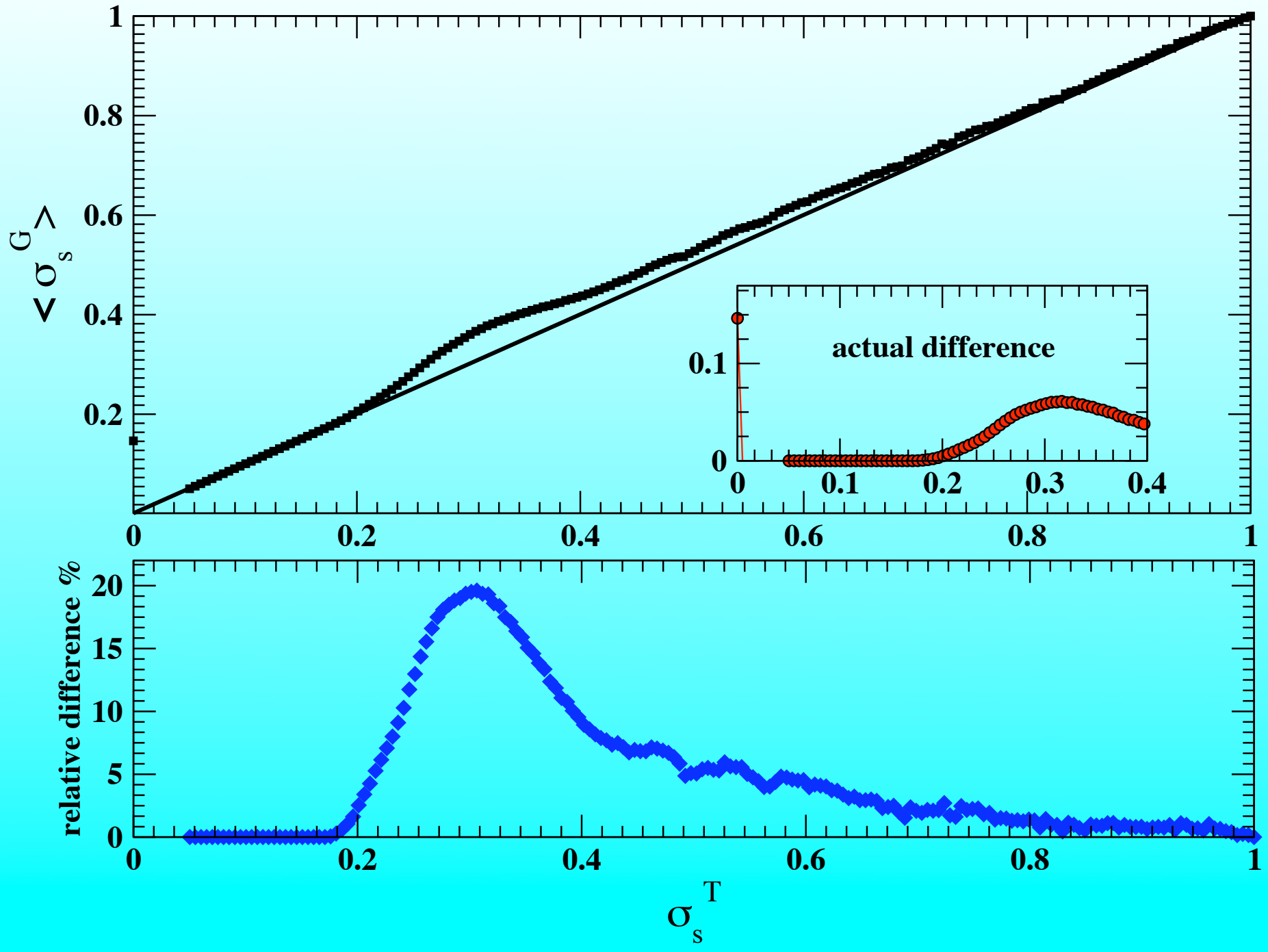
- Work w/Ana Maguitman & al.
  - Include cross-links (symbolic) and see-also links (related)
  - Transitive closure of topic graph
  - Compute entropy based on fuzzy membership matrix

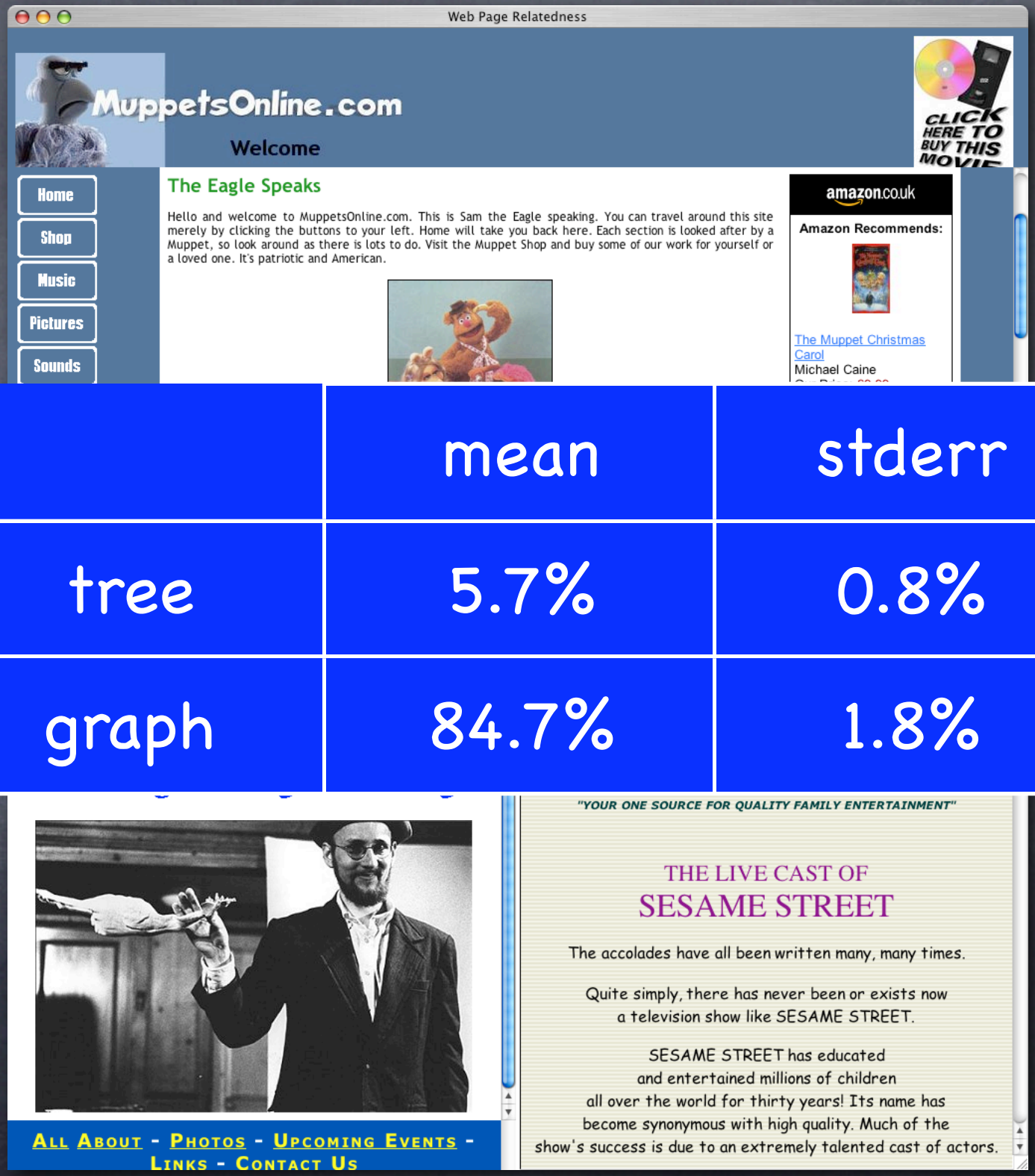
$$\mathbf{W} = \mathbf{T}^{+} \odot \mathbf{G} \odot \mathbf{T}^{+}$$

$$\sigma_s(t_1, t_2) = \max_k \frac{2 \cdot \min\left(\mathrm{W}_{k1}, \mathrm{W}_{k2}\right) \cdot \log \Pr[t_k]}{\log(\Pr[t_1|t_k] \cdot \Pr[t_k]) + \log(\Pr[t_2|t_k] \cdot \Pr[t_k])}$$

Differences

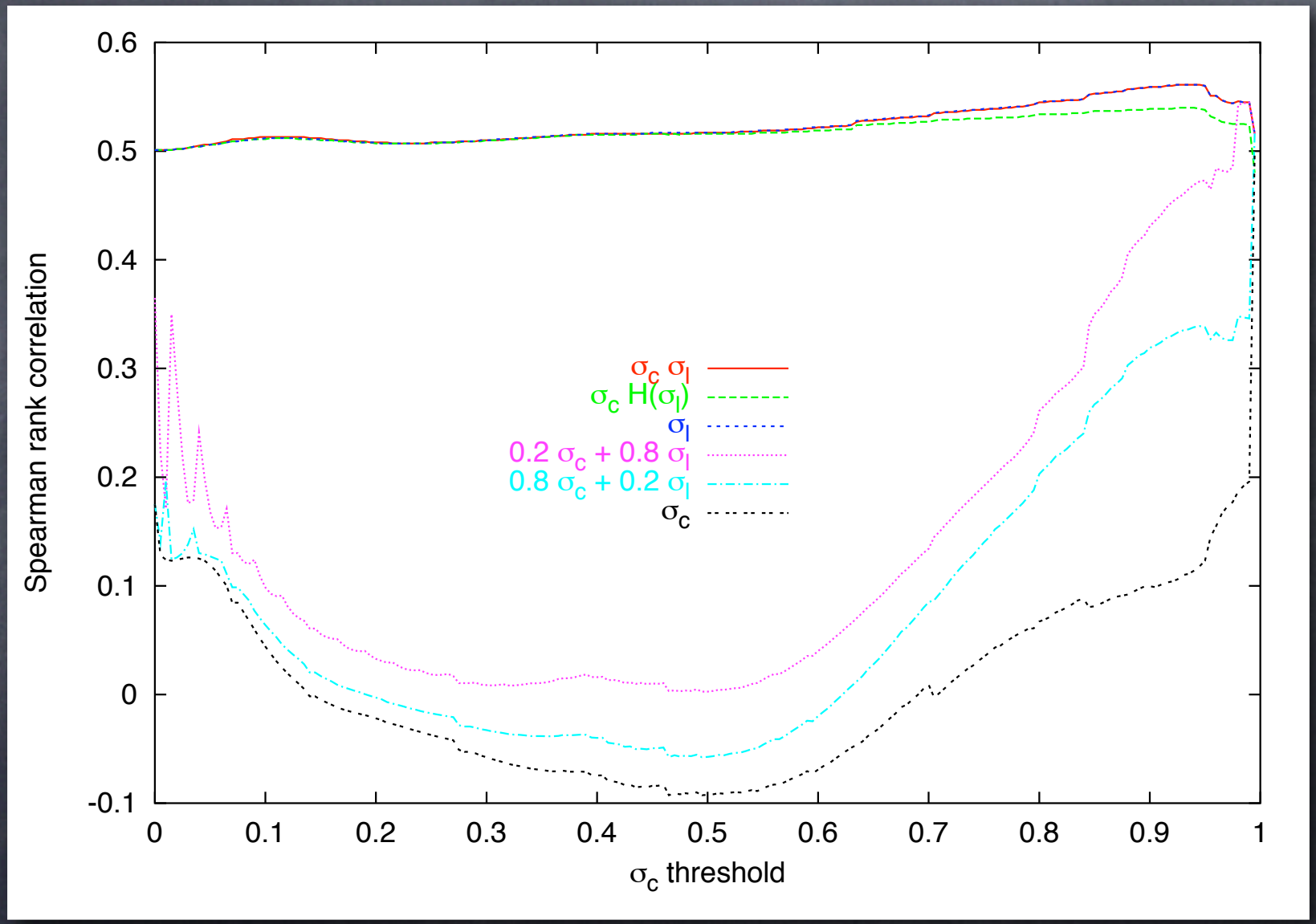



1
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1
$<\sigma_s^G>$
actual difference
0.1
0
0
0.1
0.2
0.3
0.4
relative difference %
20
15
10
5
0
0
0.2
0.4
0.6
0.8
1
σ_s^T



	mean	stderr
tree	5.7%	0.8%
graph	84.7%	1.8%

Combining content & links



Proc. WWW 2005, WWWJ 2006

Discussion: Is content really so bad? Why?

Outline

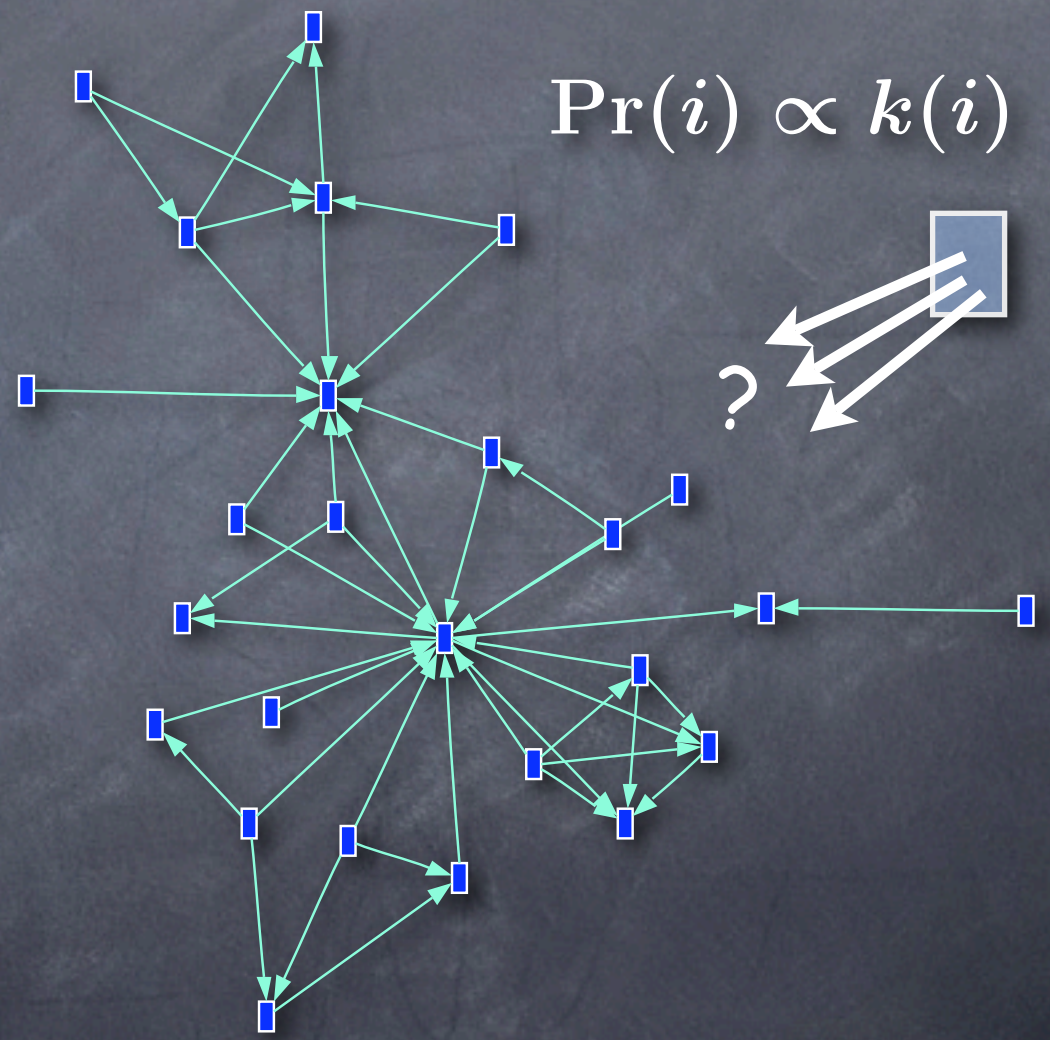
- ✓ Mapping
  - Topical locality
  - Content, link, and semantic topologies in the Web
- Modeling
  - How the Web evolves and why content matters
  - Consequences for navigation and crawling
- Mining
  - Topical Web crawlers
  - Adaptive, intelligent crawling techniques
- Mingling
  - Social Web search & recommendation
  - Distributed collaborative peer search

Preferential attachment

- "BA" model

  (Barabasi & Albert 1999, de Solla Price 1976)

  - At each step t add new page p
  - Create m new links from p to i (i<t)
- Rich-get-richer



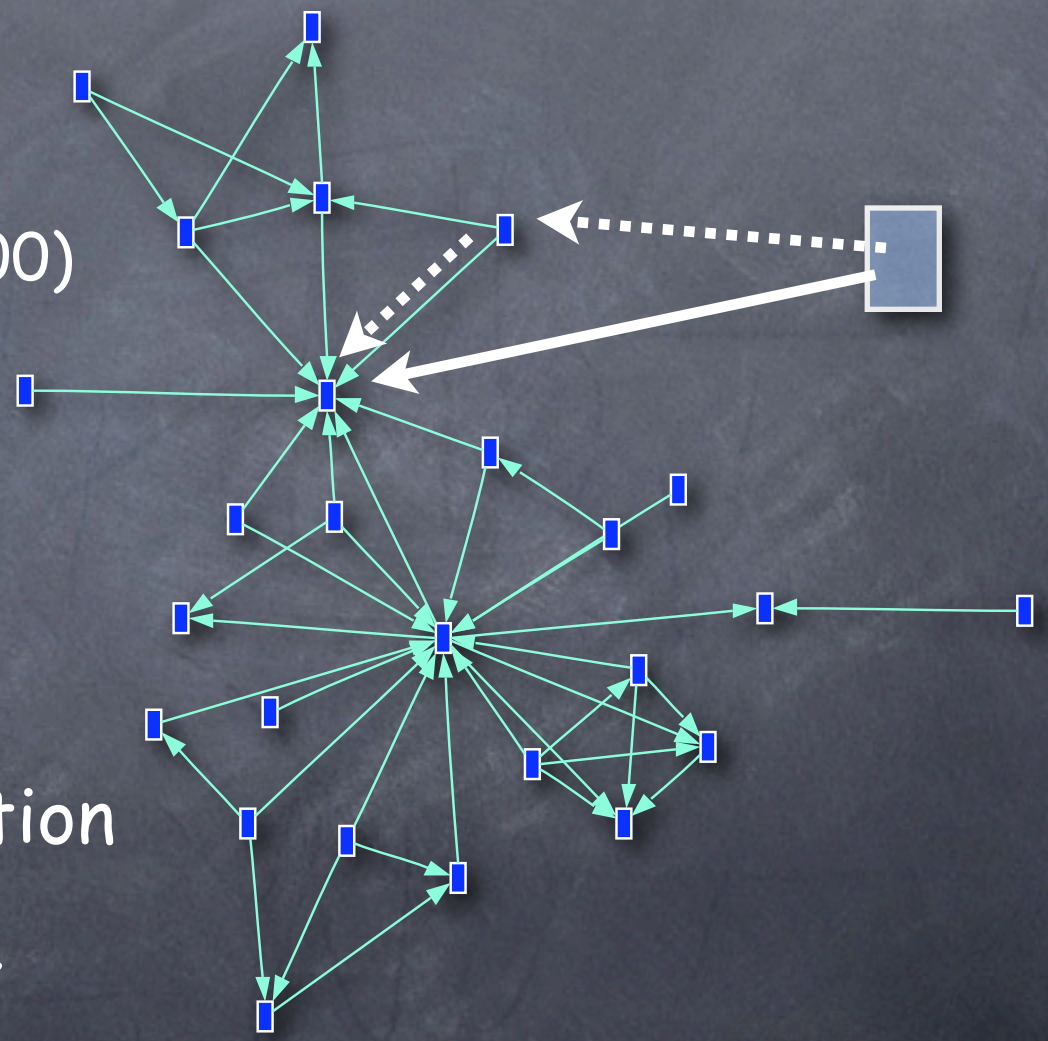
$$\Pr(i) = \frac{k(i)}{mt} \implies \Pr(k) \sim k^{-\gamma}$$

Other growth models

Web copying

(Kleinberg, Kumar & al 1999, 2000)

$$\Pr(i) \propto \Pr(j) \cdot \Pr(j \to i)$$



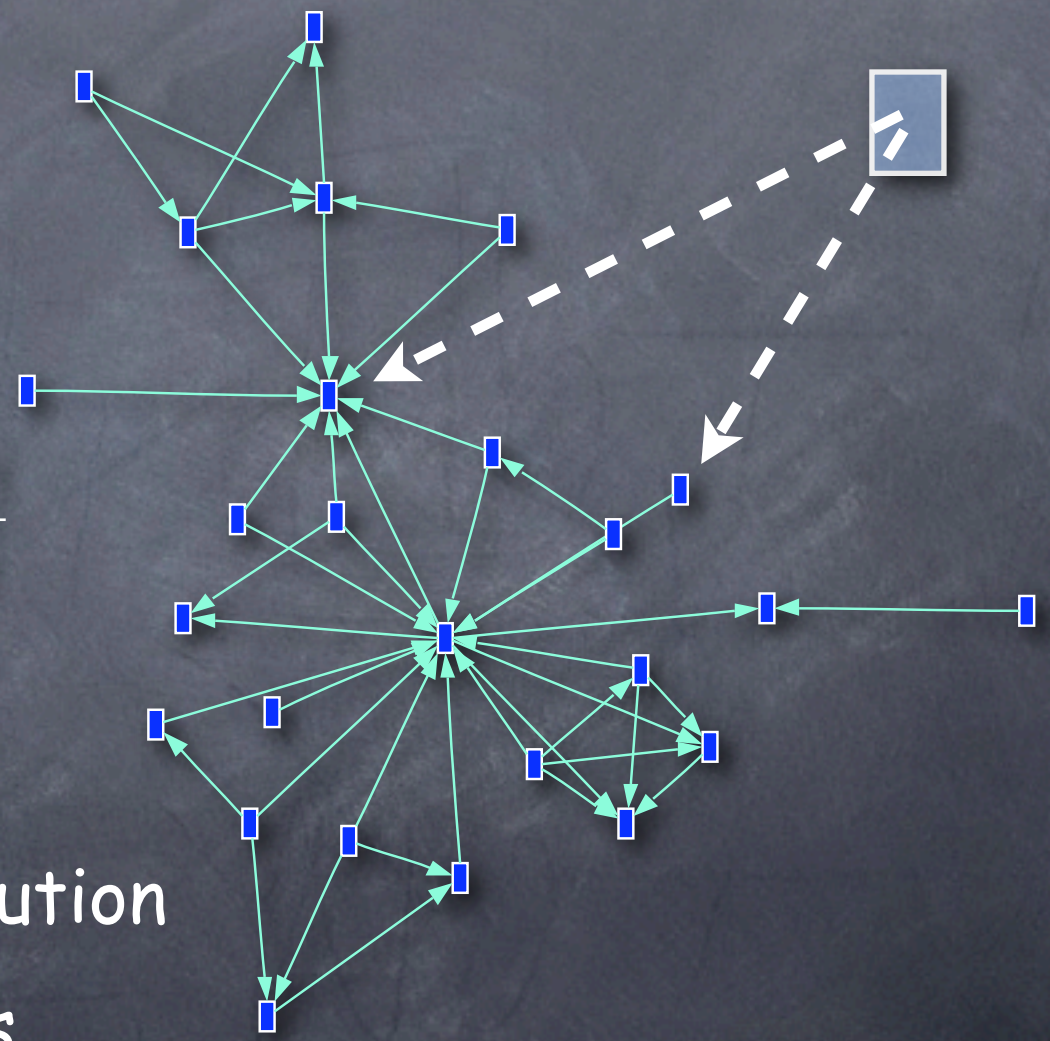
- same indegree distribution
- no need to know degree

Other growth models

Random mixture

(Pennock & al. 2002,
Cooper & Frieze 2001,
Dorogovtsev & al 2000)

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



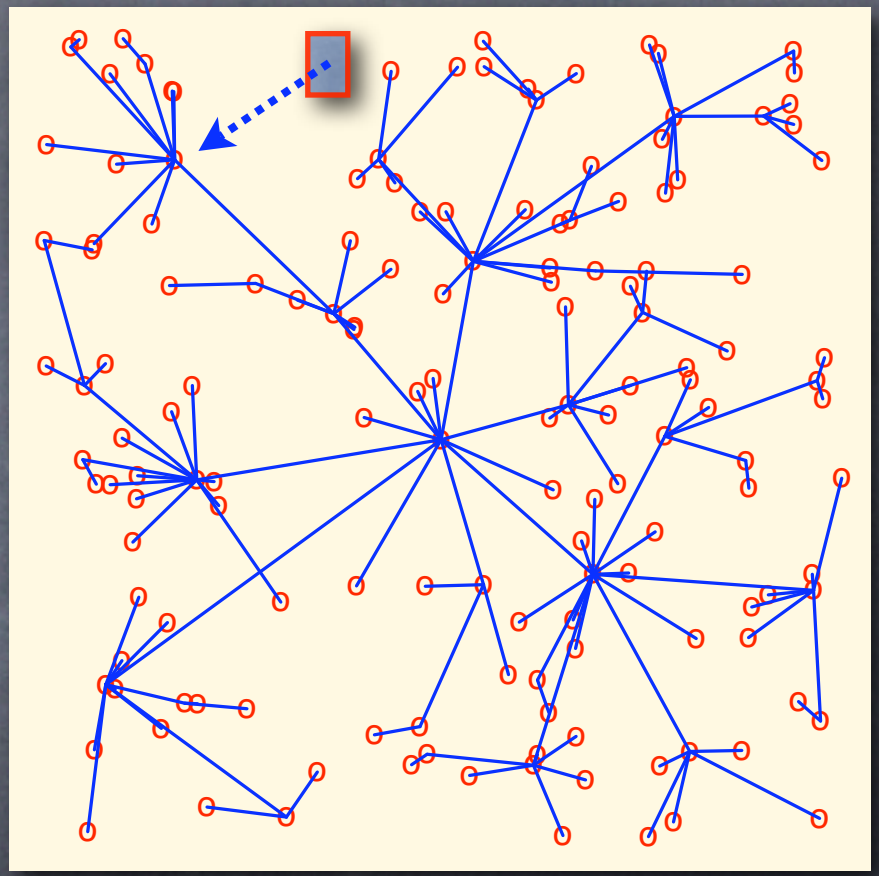
- winners don't take all
- general indegree distribution
- fits non-power-law cases

Other growth models

Mixture with Euclidean distance

(HOT: Fabrikant, Koutsoupias & Papadimitriou 2002)

$$i = \arg\min(\phi r_{it} + g_i)$$



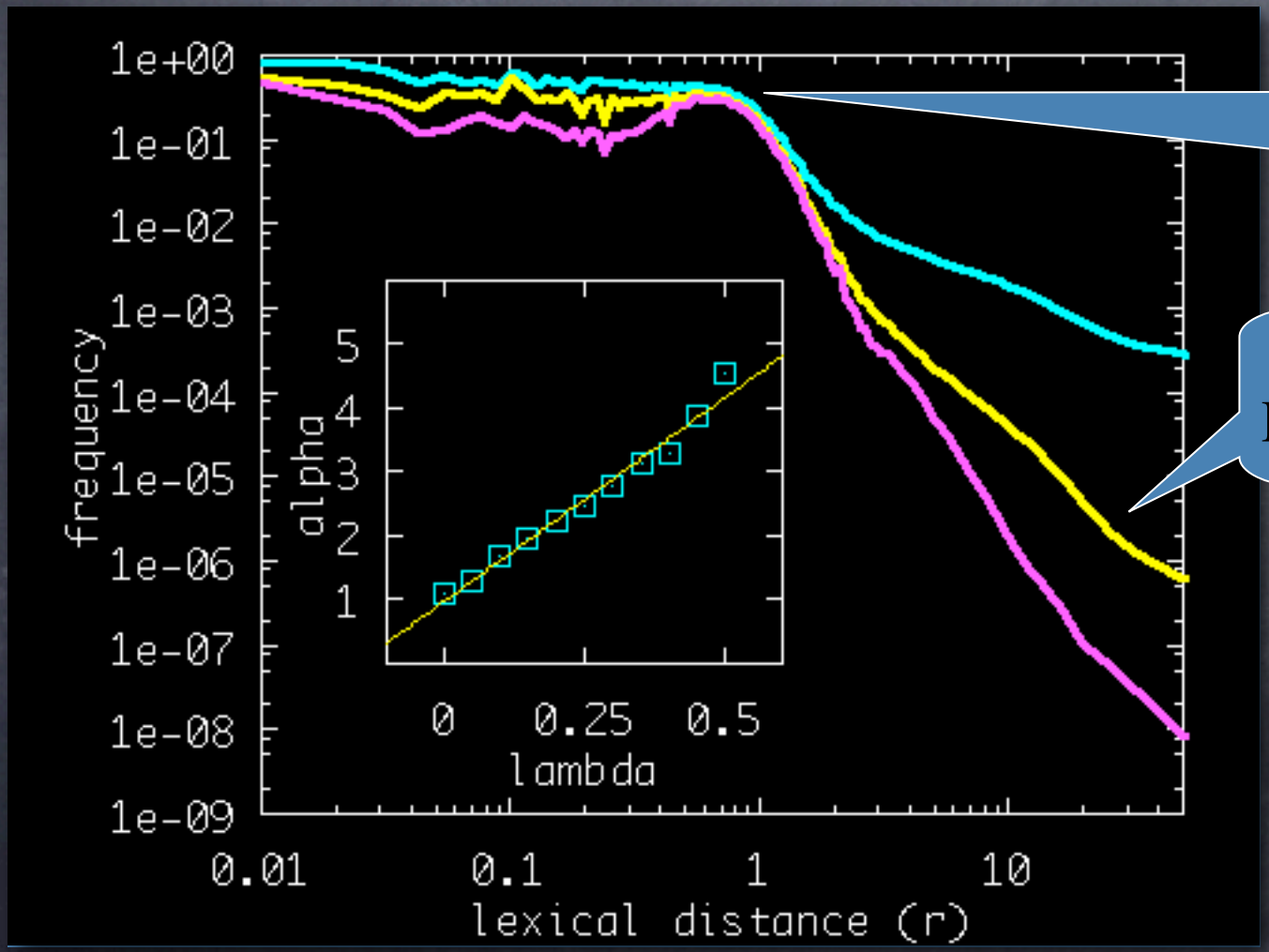
- tradeoff between centrality and geometric locality
- fits power-law in certain critical trade-off regimes

Discussion:
What about content?

Link probability vs lexical distance

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

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



Phase transition ρ^*

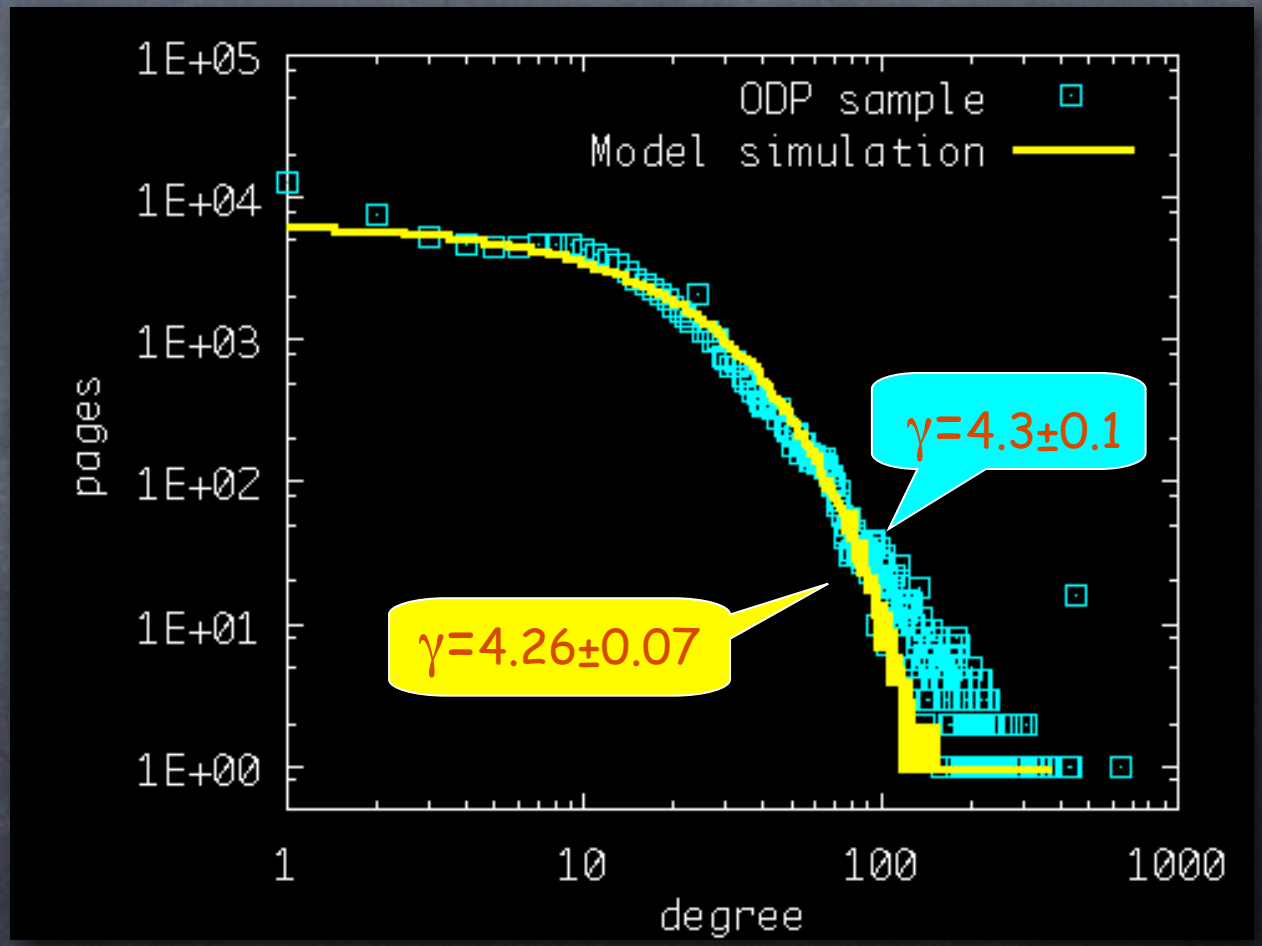
Power law tail $\Pr(\lambda \mid \rho) \sim \rho^{-\alpha(\lambda)}$

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

Local content-based growth model

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

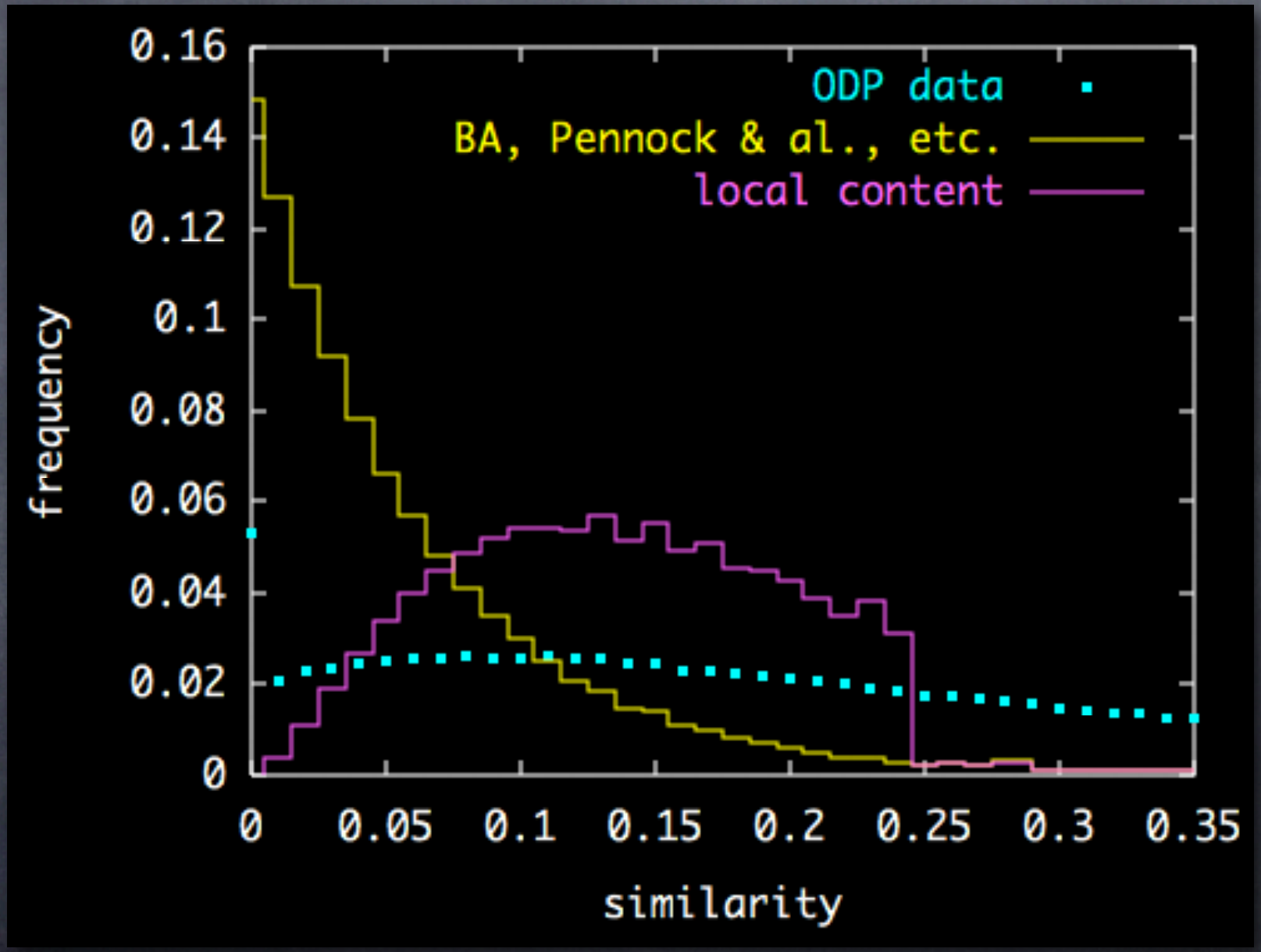
- Similar to preferential attachment (BA)
- Use degree info (popularity/ importance) only for nearby (similar/ related) pages



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 <u>linked</u> pairs

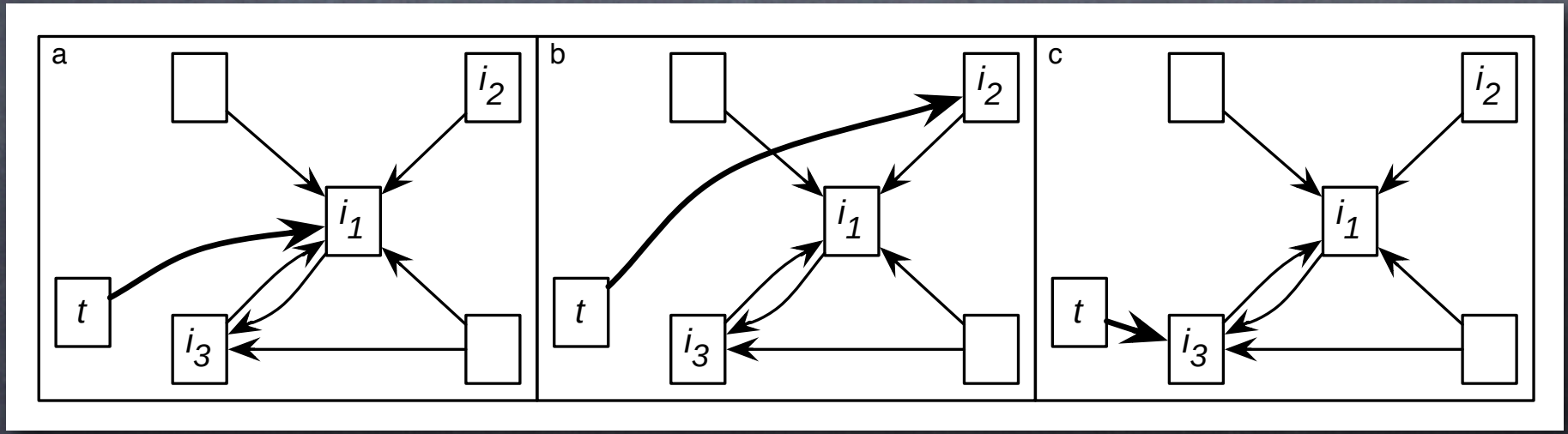
None of these models is right!



Back to the mixture model

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

degree-uniform mixture



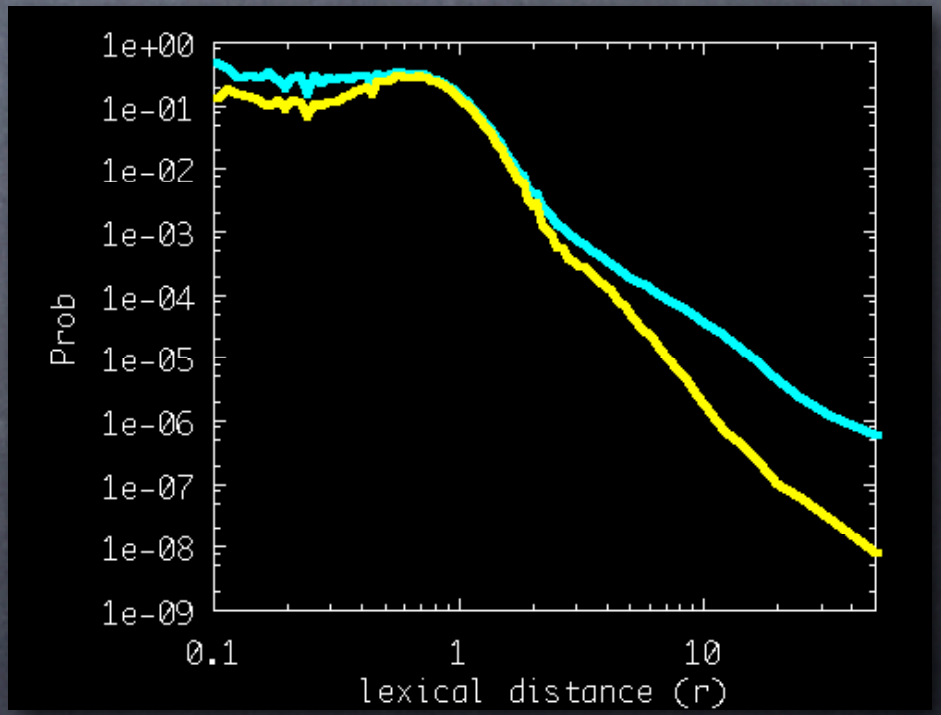
Bias choice by content similarity instead of uniform distribution

Degree-similarity mixture model

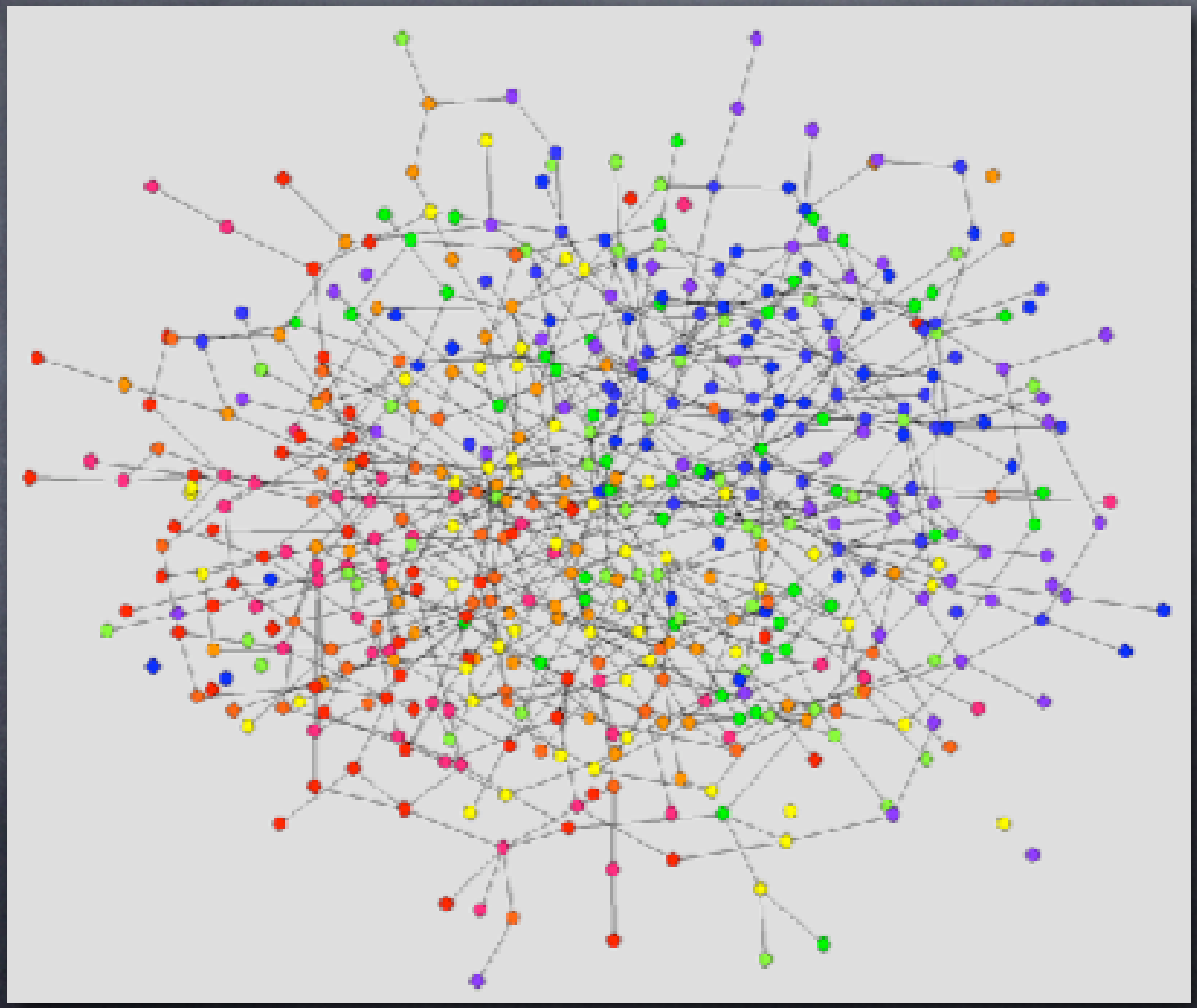
$$\Pr(i) \propto \psi \cdot \hat{\Pr}(i) + (1-\psi) \cdot \frac{k(i)}{mt}$$

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

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

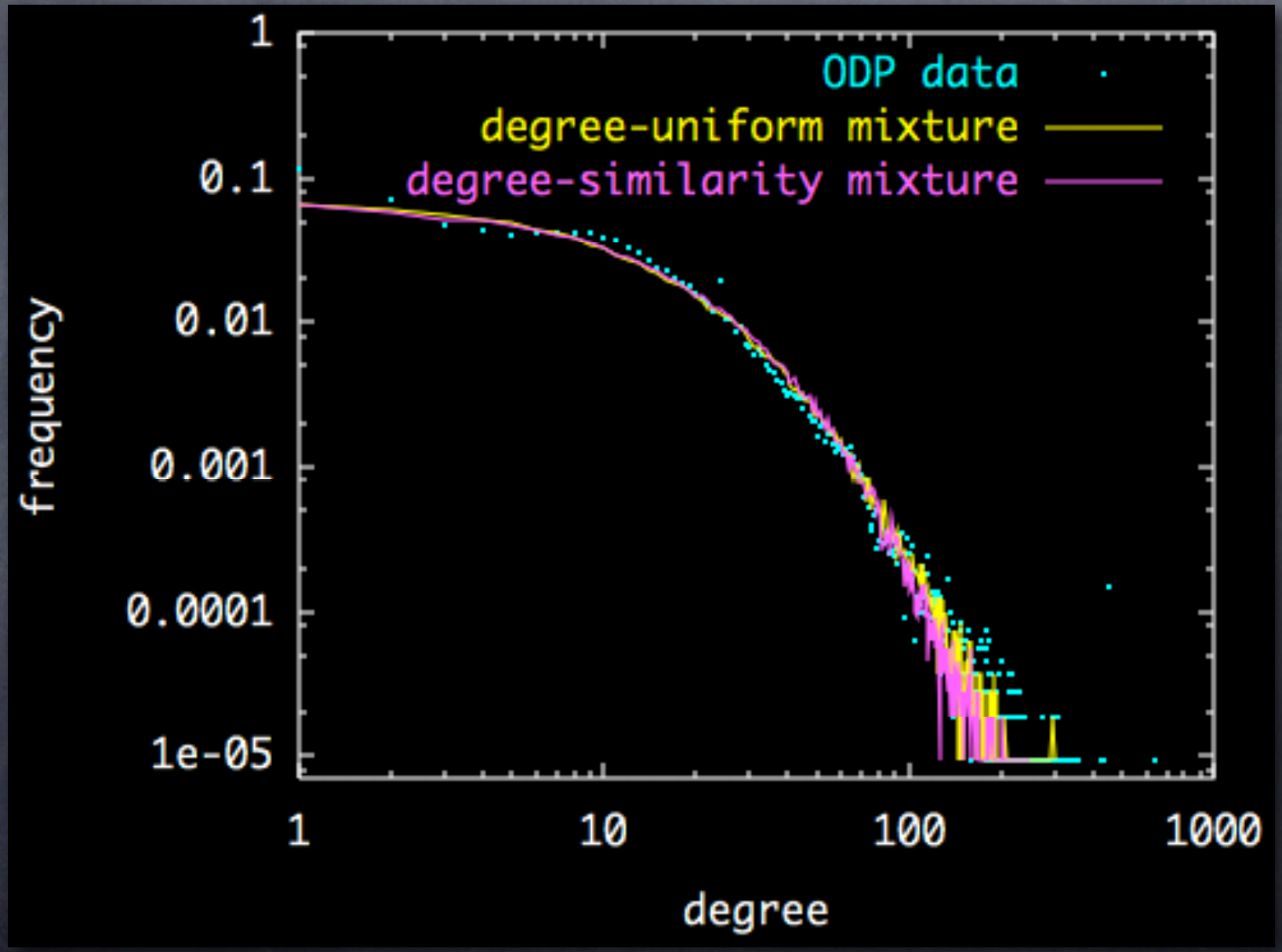


Build it...

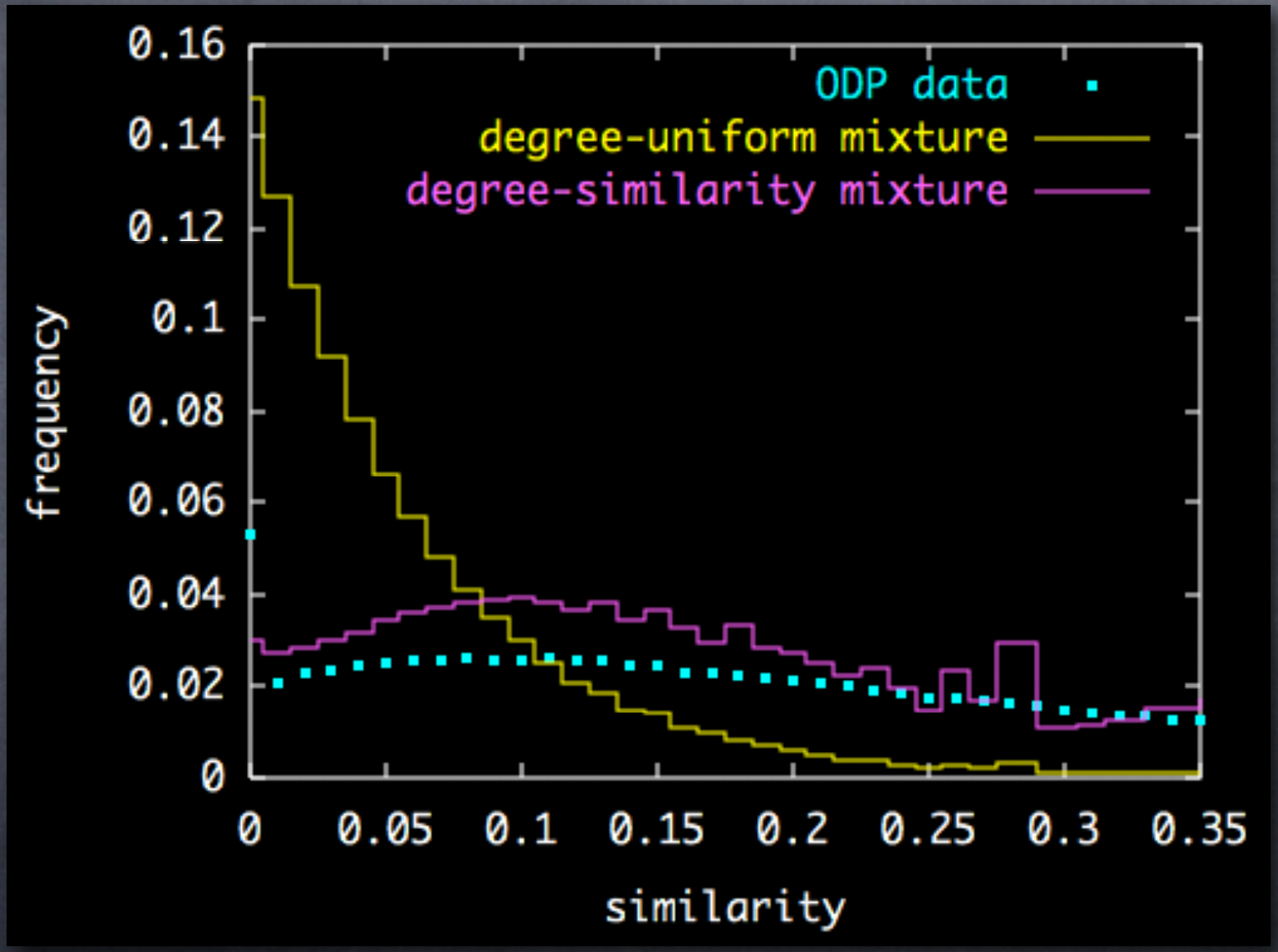


(M.M.)

Both mixture models get the degree distribution right...



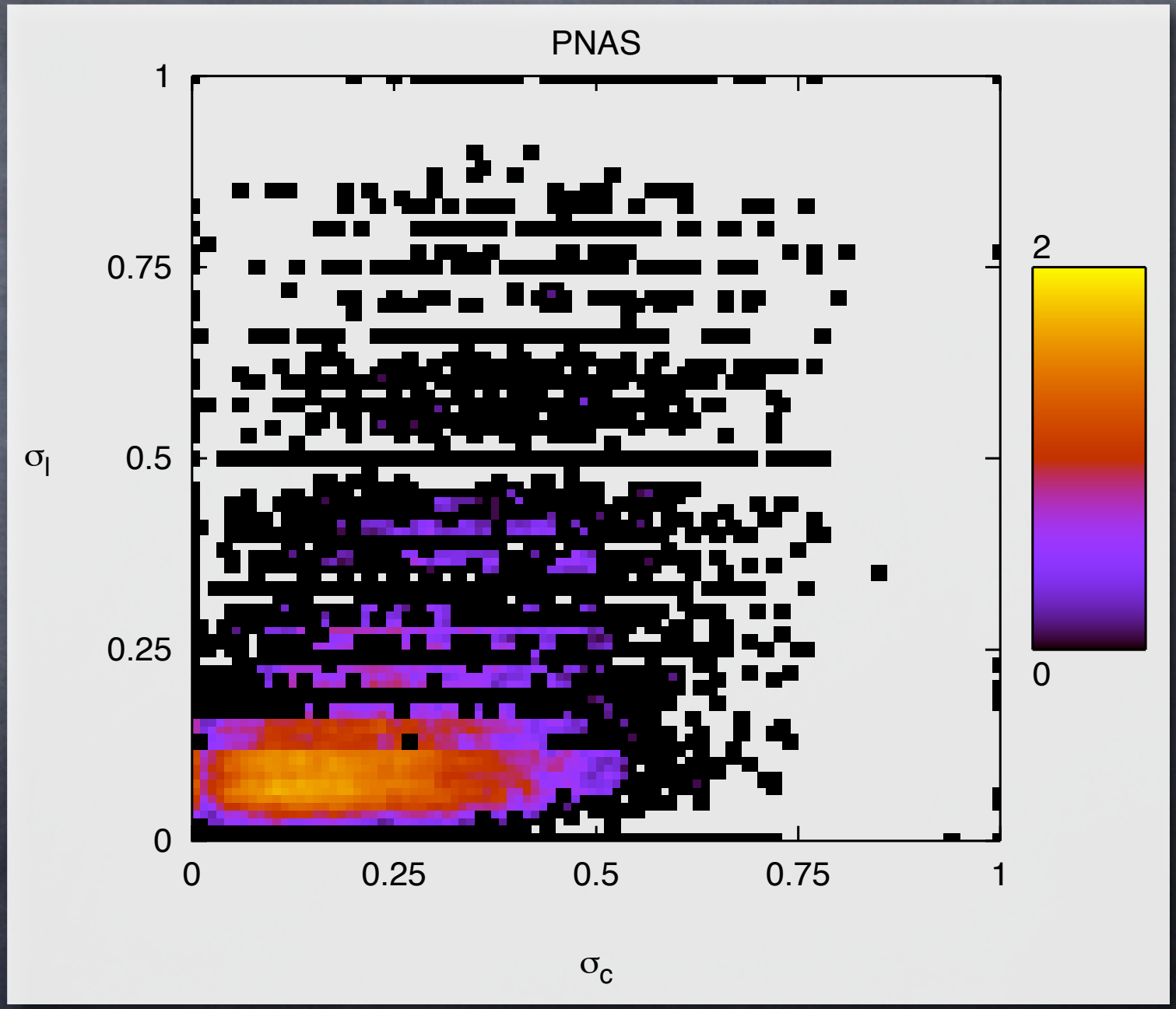
...but the degree-similarity mixture model predicts the similarity distribution better



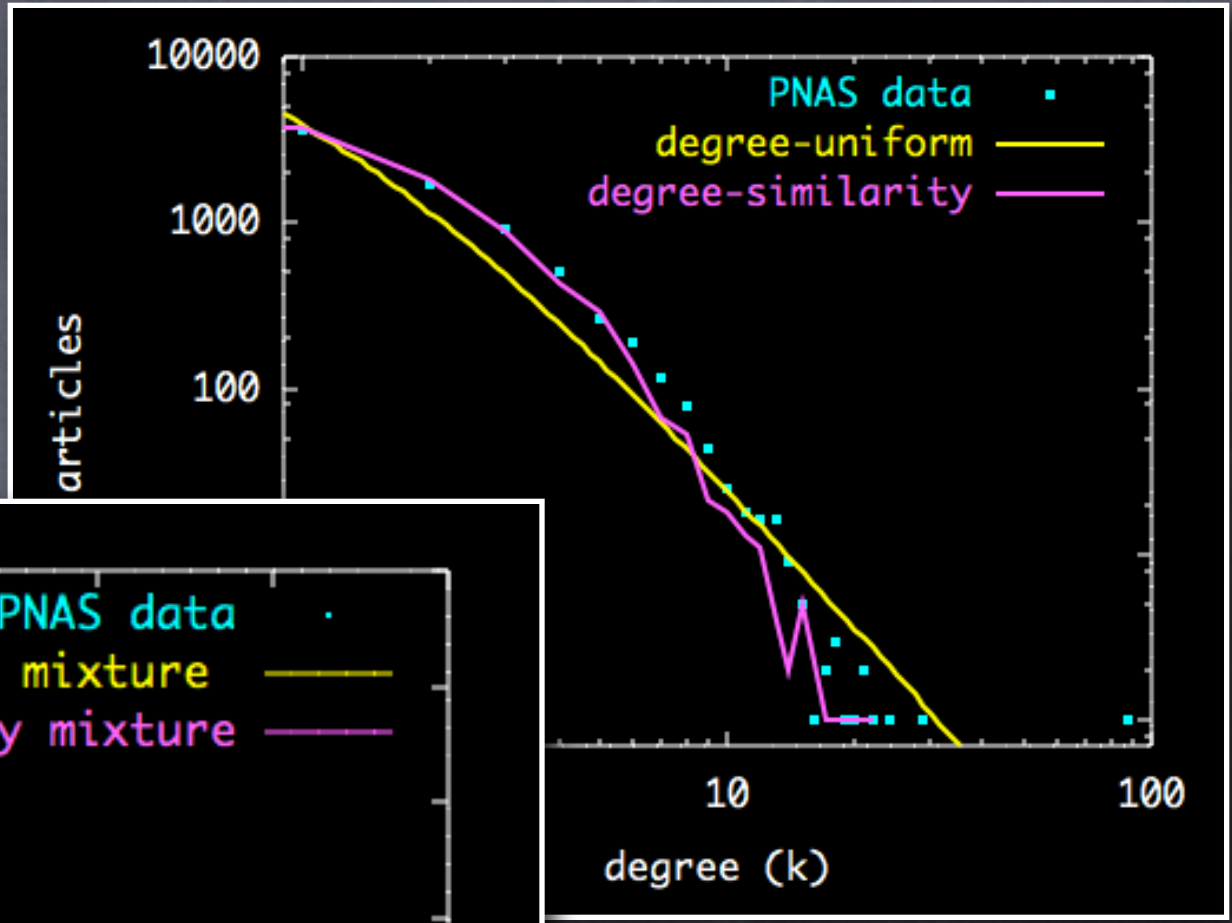
Proc. Natl. Acad. Sci. USA 101: 5261-5265, 2004

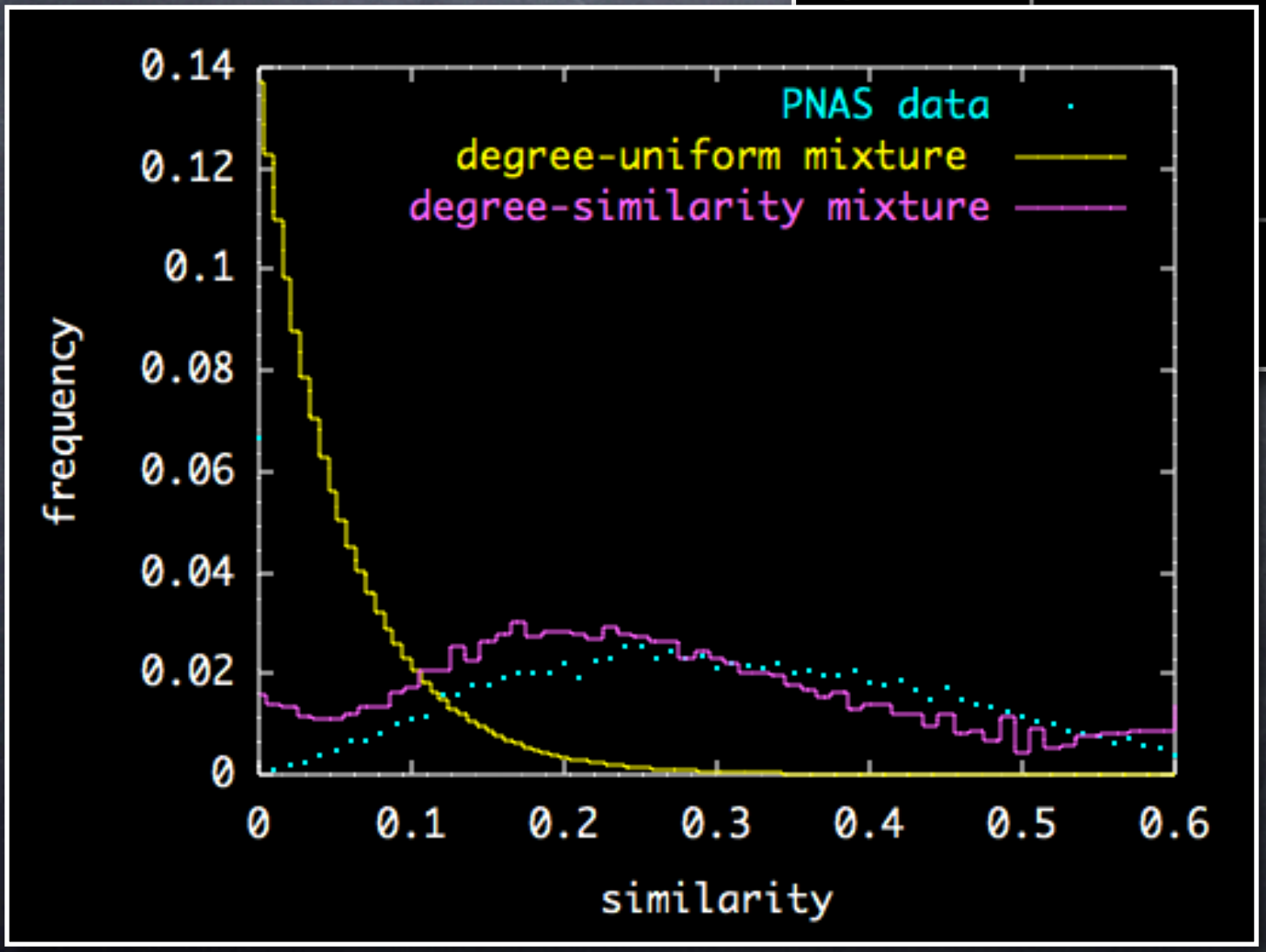
Citation networks

15,785 articles published in PNAS between 1997 and 2002



Citation networks





Discussion

- What questions remain open?
- What other factors should we consider?

Open Questions

- Understand distribution of content similarity across <u>all</u> pairs of pages
  - Using TF: exponential $\Pr(\sigma) \sim \alpha^{-\beta\sigma}$
  - Using TF-IDF: power law $\Pr(\sigma) \sim \alpha\sigma^{-\beta}$
- Growth model to explain co-evolution of both link topology and content similarity
- The role of search engines (keynote)

Efficient crawling algorithms?

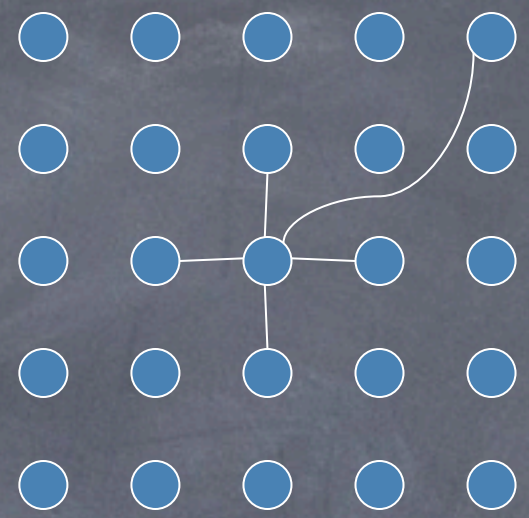
Theory: since the Web is a small world network, or has a scale free degree distribution, short paths exist between any two pages:

~ 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



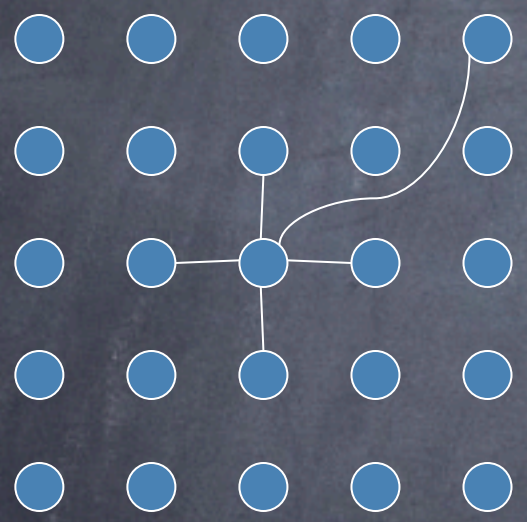
- Geographical networks (Kleinberg 2000)
  - Local links to all lattice neighbors
  - Long-range link probability distribution: power law $Pr \sim r^{-\alpha}$
    - r: lattice (Manhattan) distance
    - α: constant clustering exponent

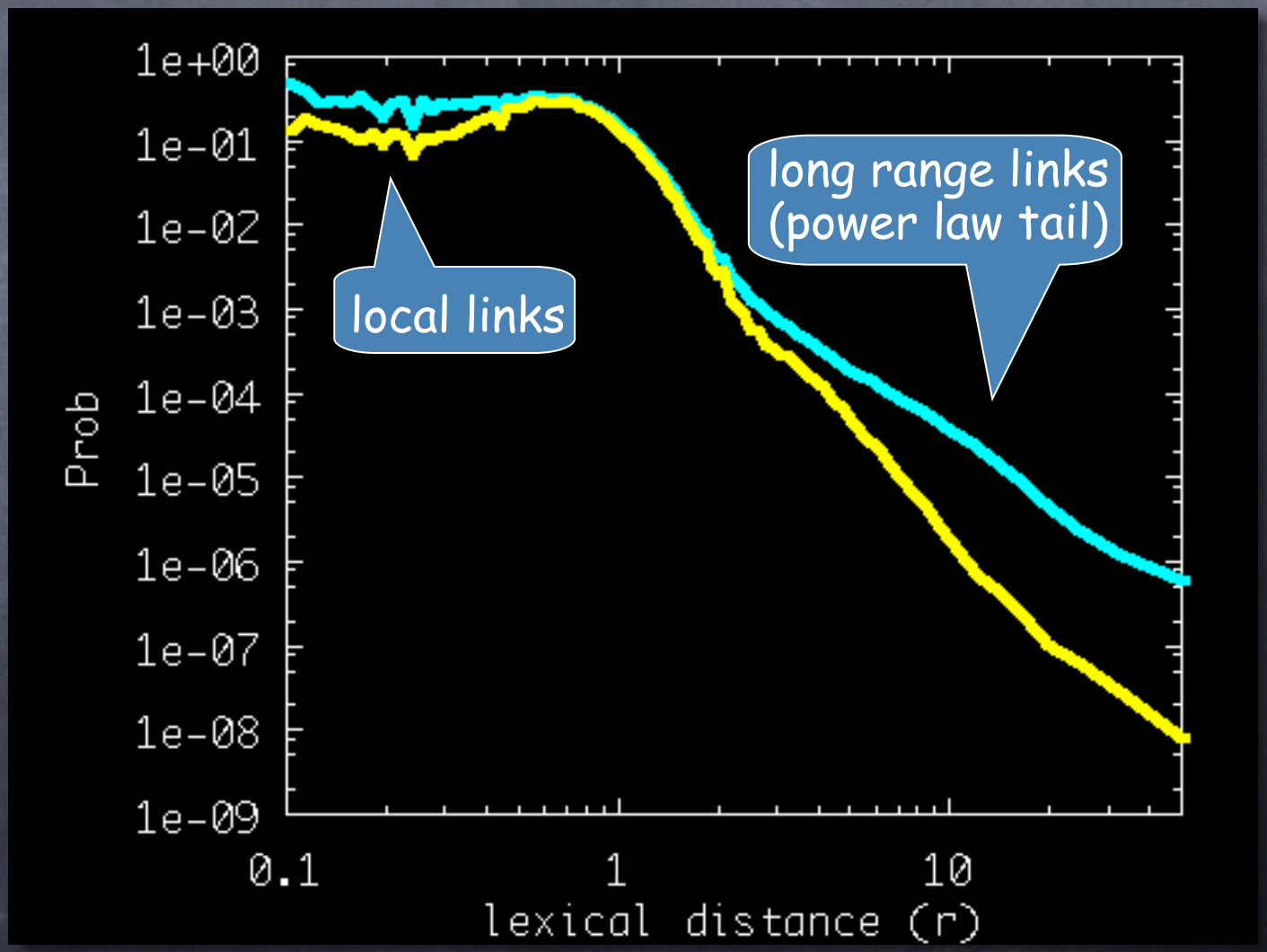
$$t \sim \log^2 N \Leftrightarrow \alpha = D$$

Is the Web a geographical network?

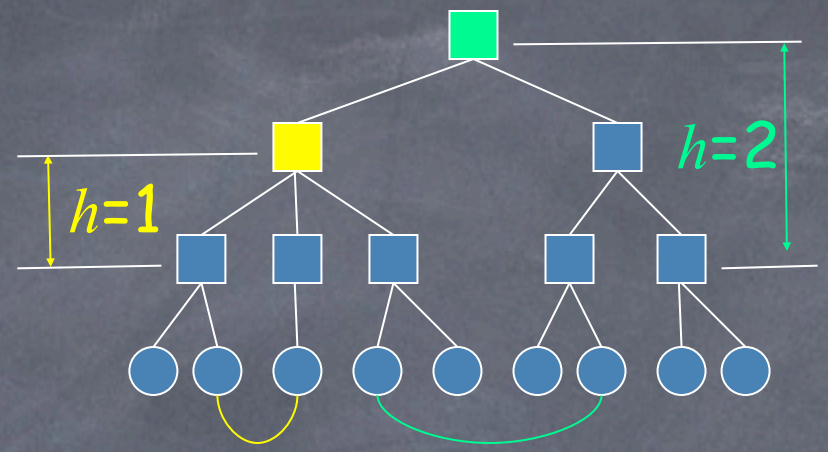
Replace lattice distance by lexical distance

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





Exception # 2



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

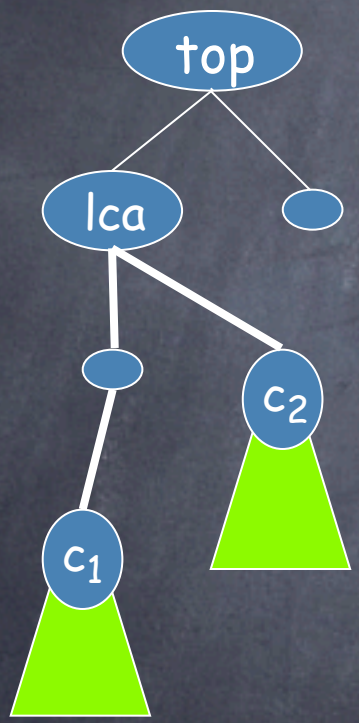
    $Pr \sim e^{-h}$
    - h: tree distance (height of lowest common ancestor)

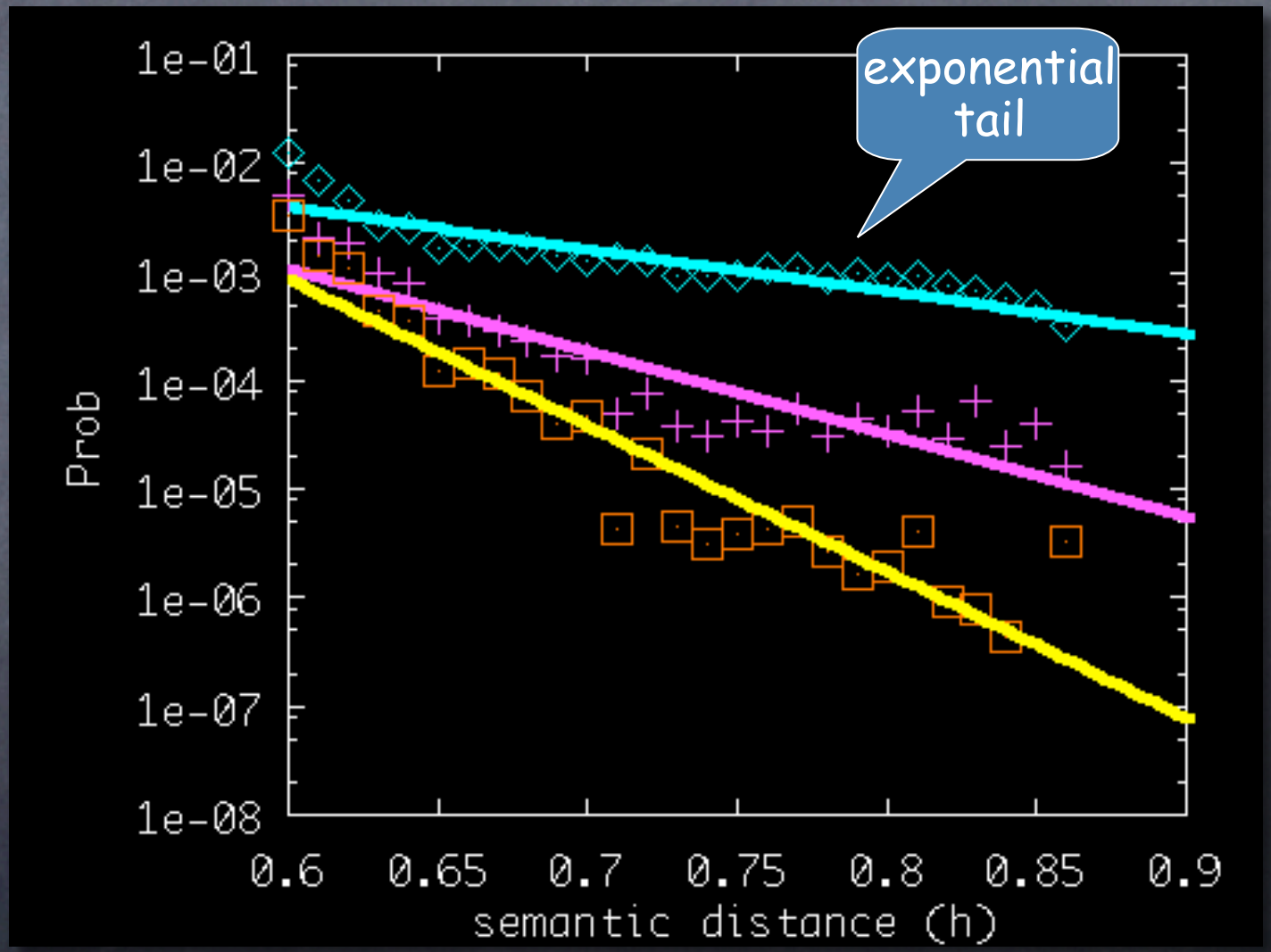
$$t \sim \log^{\varepsilon} N, \varepsilon \geq 1$$

Is the Web a hierarchical network?

Replace tree distance by semantic distance

$h = 1 - \sigma_s$





Discussion: Does this stuff really have any applications?

Outline

- ✓ Mapping
  - Topical locality
  - Content, link, and semantic topologies in the Web
- ✓ Modeling
  - How the Web evolves and why content matters
  - Consequences for navigation and crawling
- Mining
  - Topical Web crawlers
  - Adaptive, intelligent crawling techniques
- Mingling
  - Social Web search & recommendation
  - 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

- Seminal work
  - Cho & Garcia-Molina @ Stanford
  - Chakrabarti & al @ IBM Almaden / IIT
  - Amento & al @ ATT Shannon
  - Ben-Shaul & al @ IBM Haifa
  - many others…

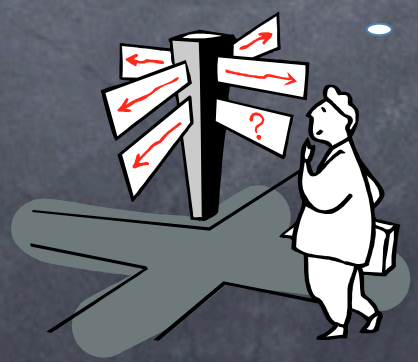
Topical crawlers

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.

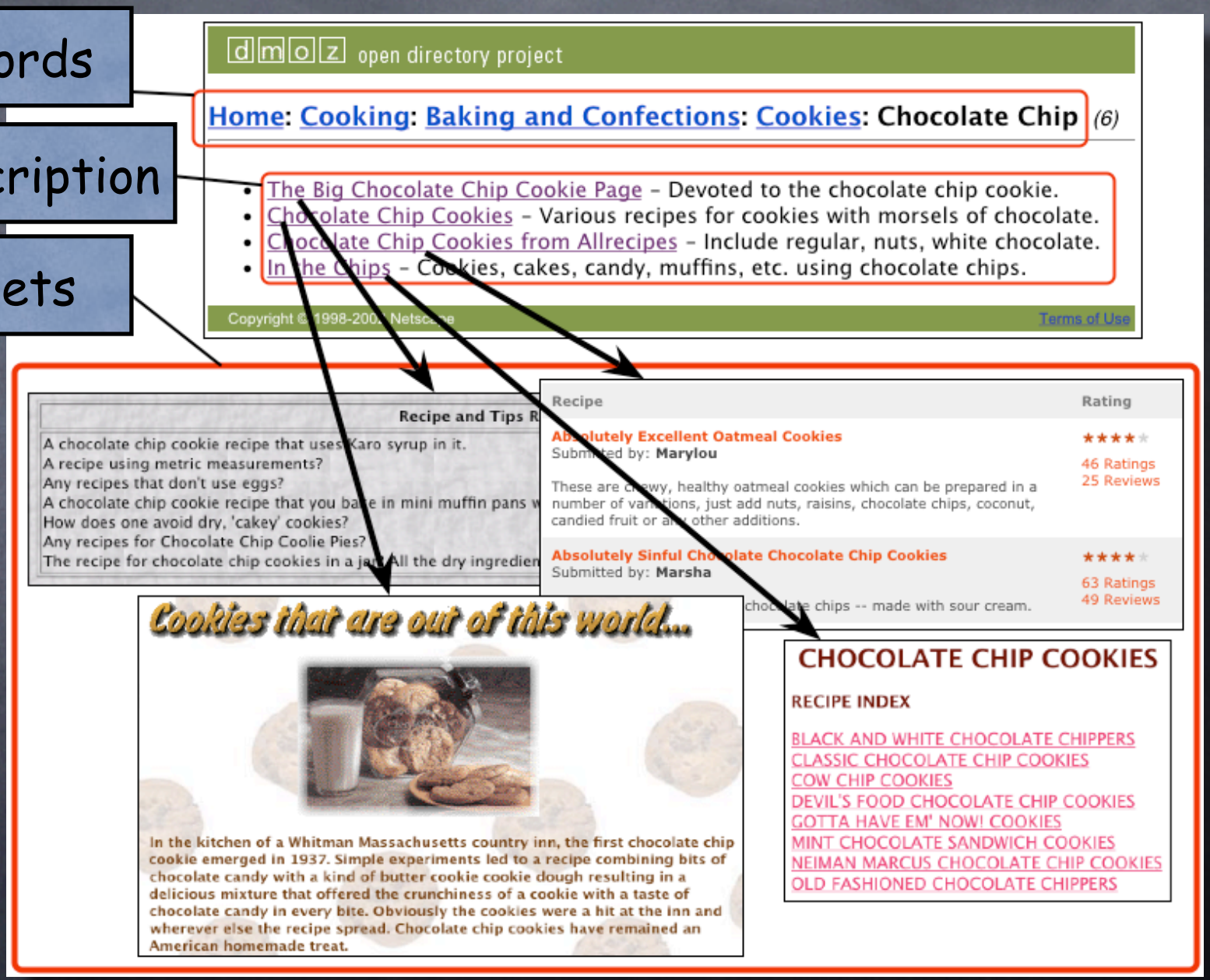
spears



Evaluating topical crawlers

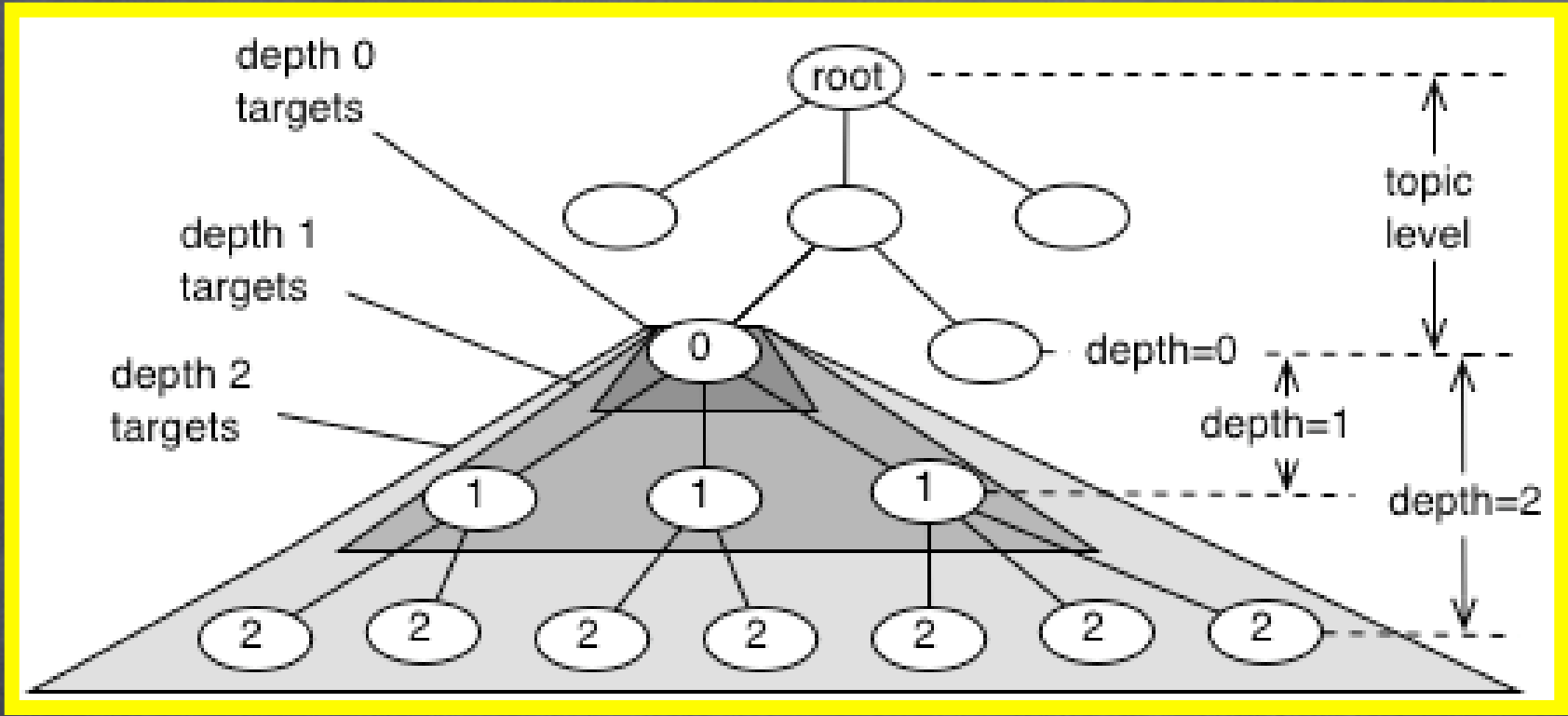
- Goal: build "better" crawlers to support applications
- Build an unbiased evaluation framework
  - Define common tasks of measurable difficulty
  - Identify topics, relevant targets
  - Identify appropriate performance measures
    - Effectiveness: quality of crawler pages, order, etc.
    - Efficiency: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

Evaluating topical crawlers: Topics



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

Topics and Targets



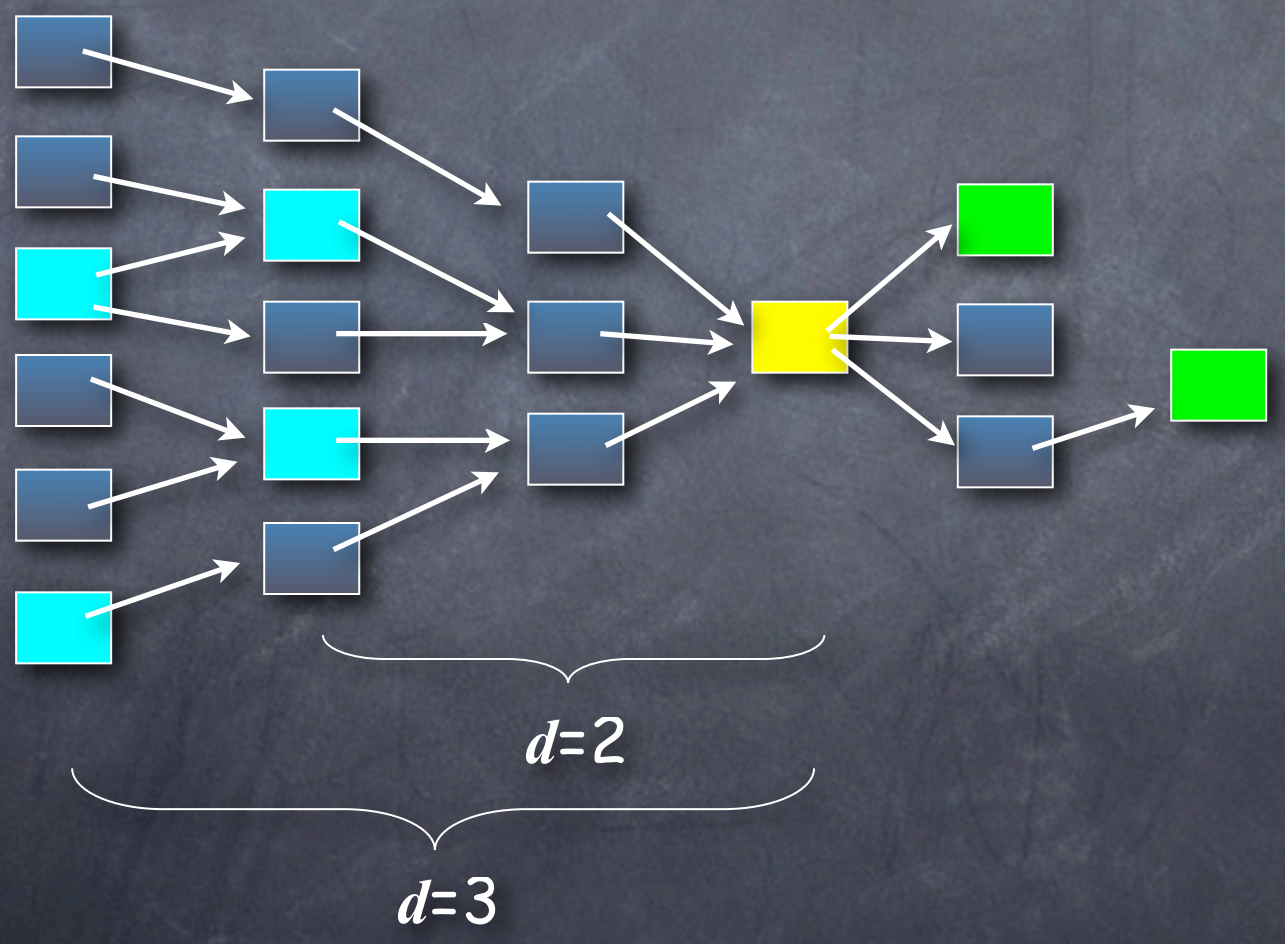
topic level ~ specificity

depth ~ generality

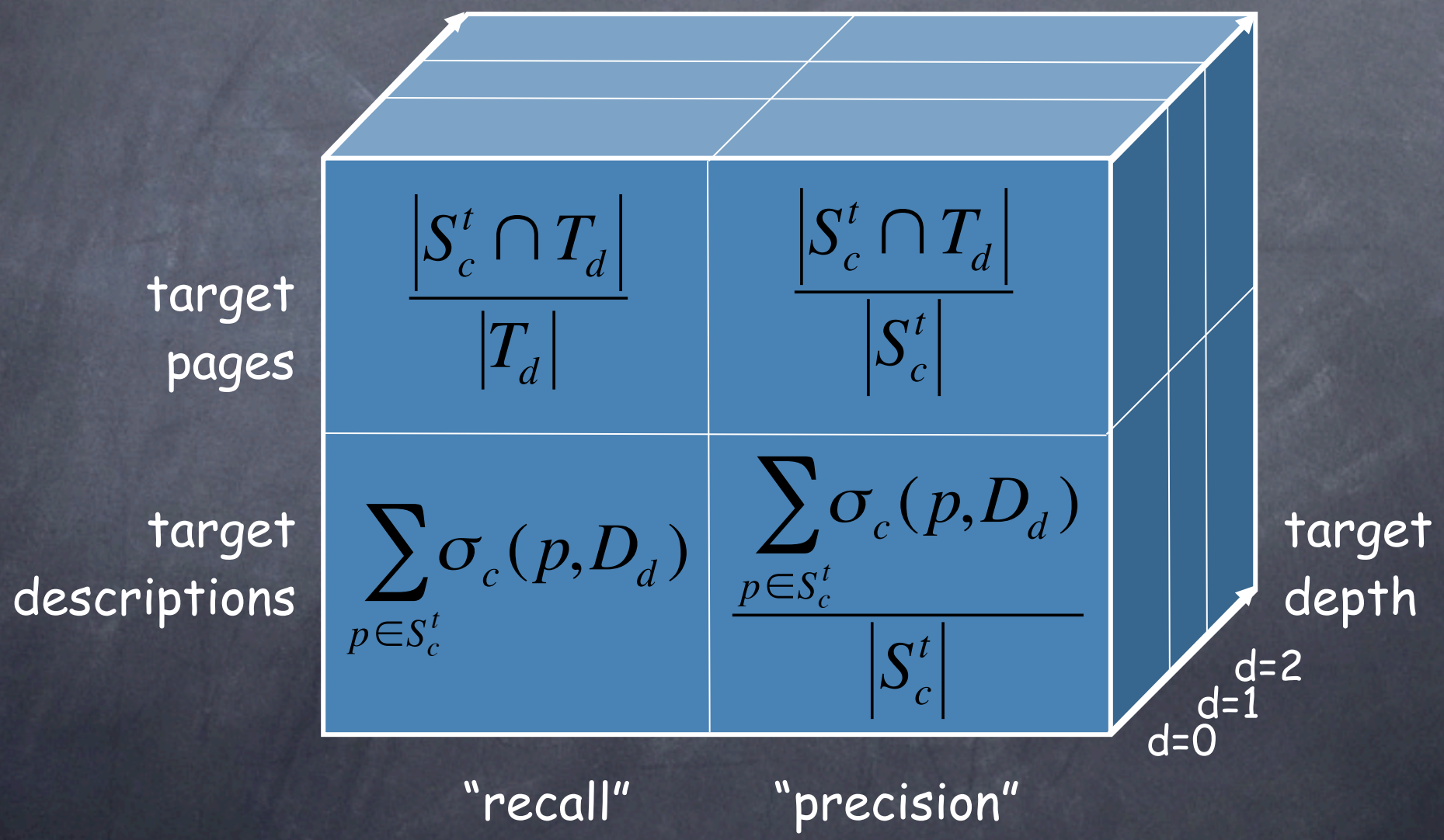
Evaluating topical crawlers: Tasks

Start from seeds, find targets

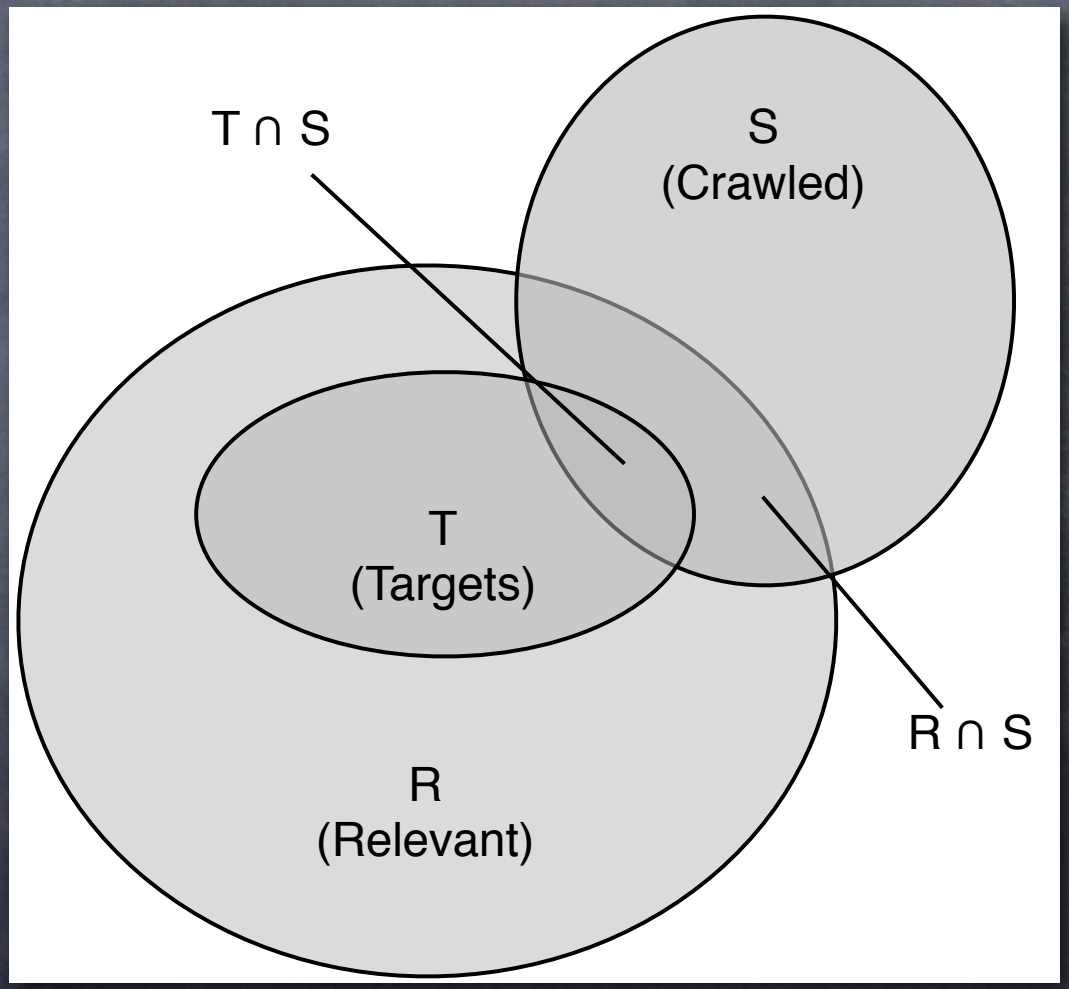
and/or pages similar to target descriptions



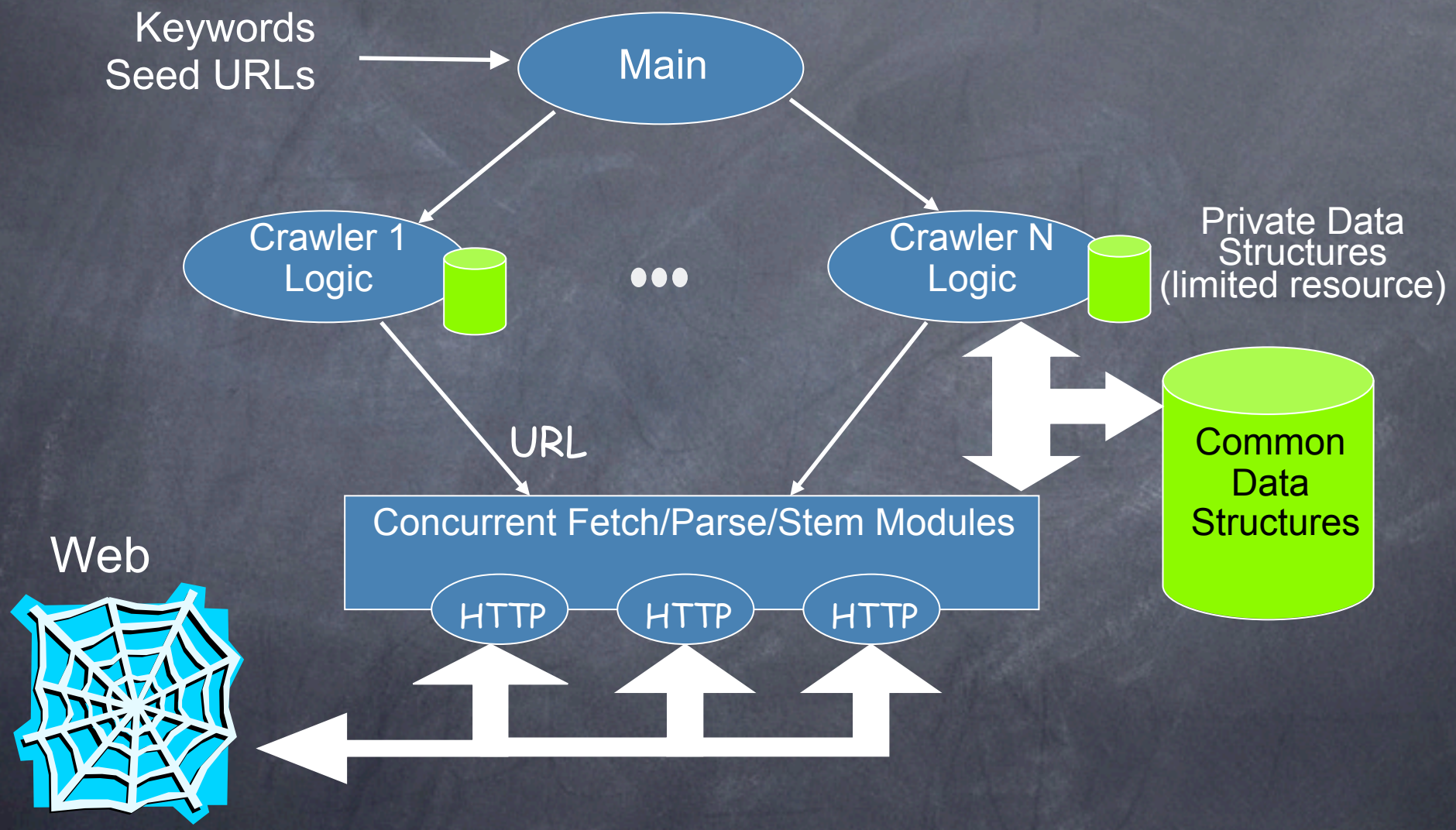
Performance matrix



Discussion:
What assumption (targets)?



Evaluating topical crawlers: Architecture

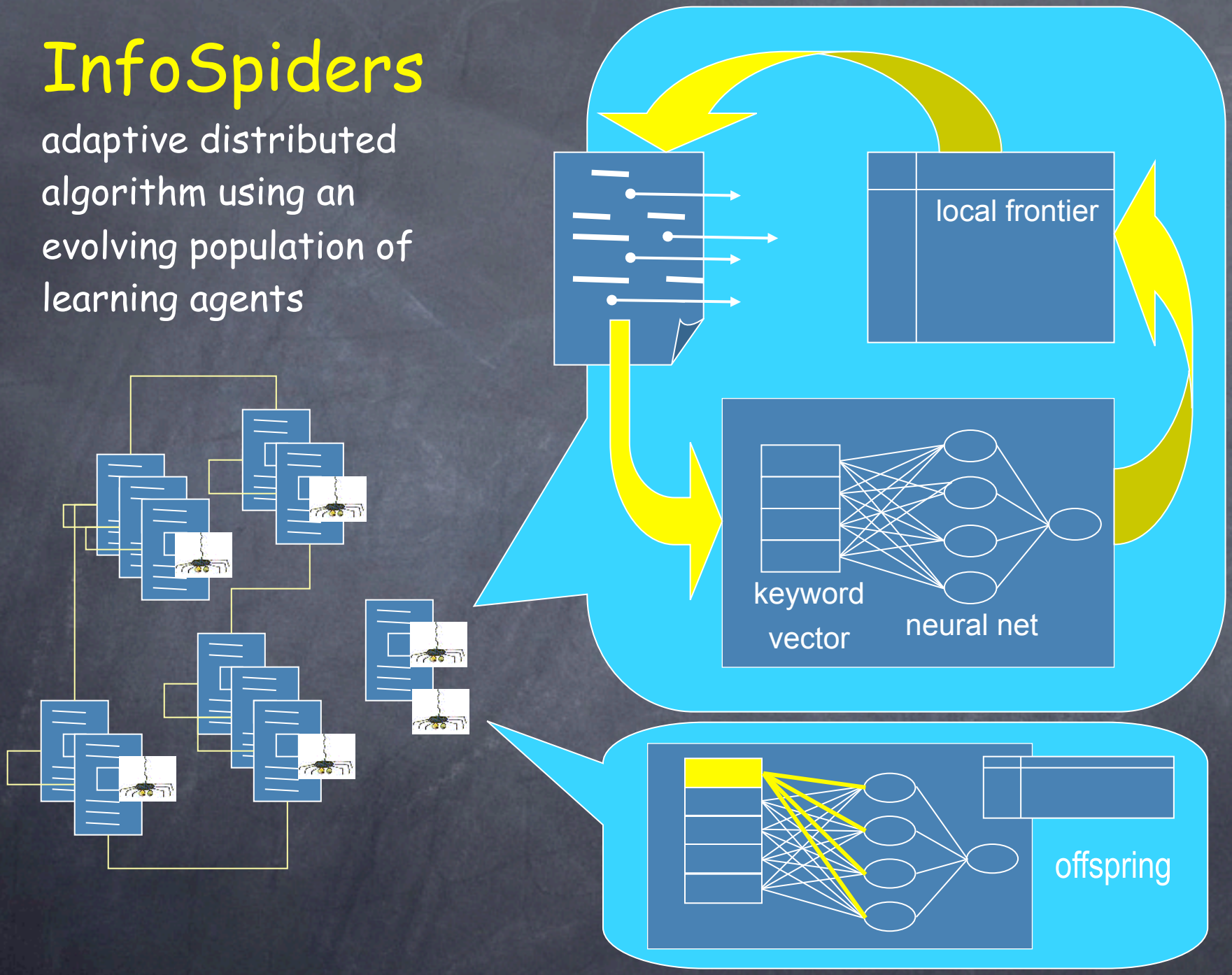


Examples of crawling algorithms

- 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

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

Impossible query:
The population goes extinct

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

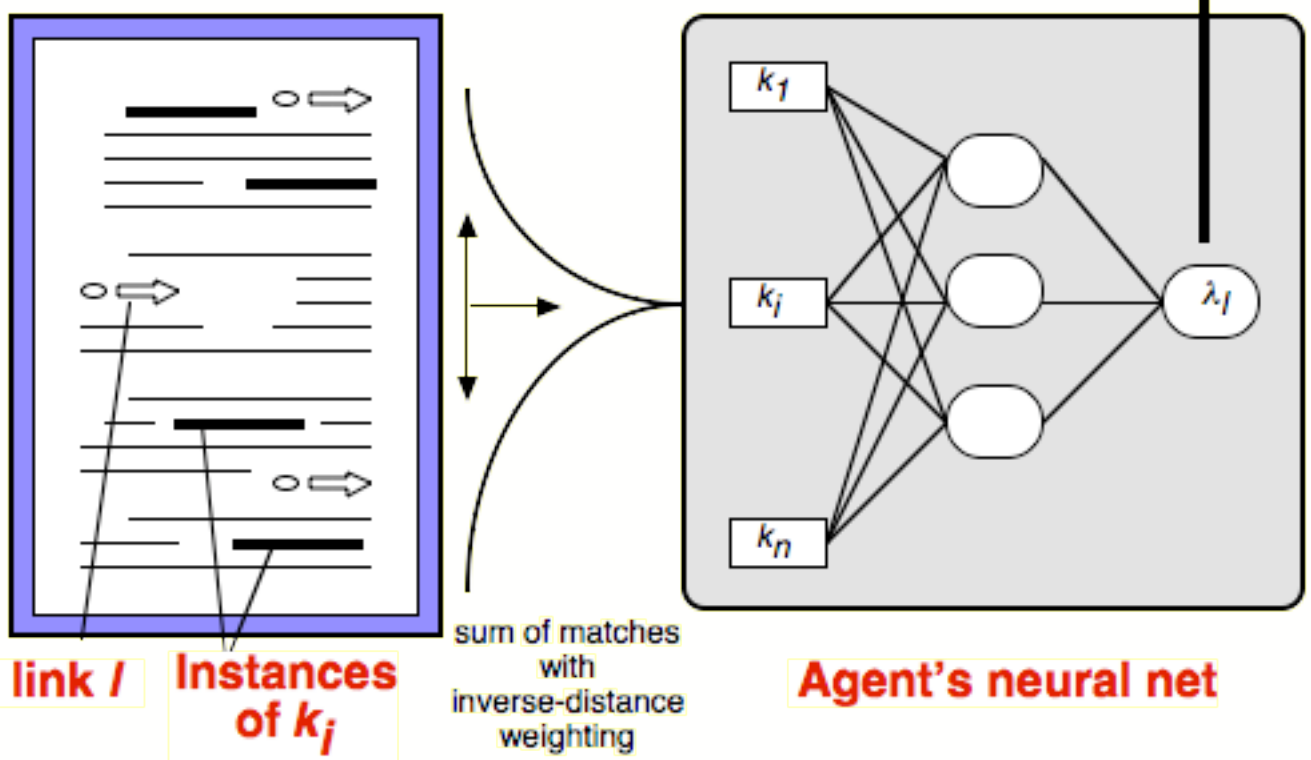
Action selection

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

Stochastic selector

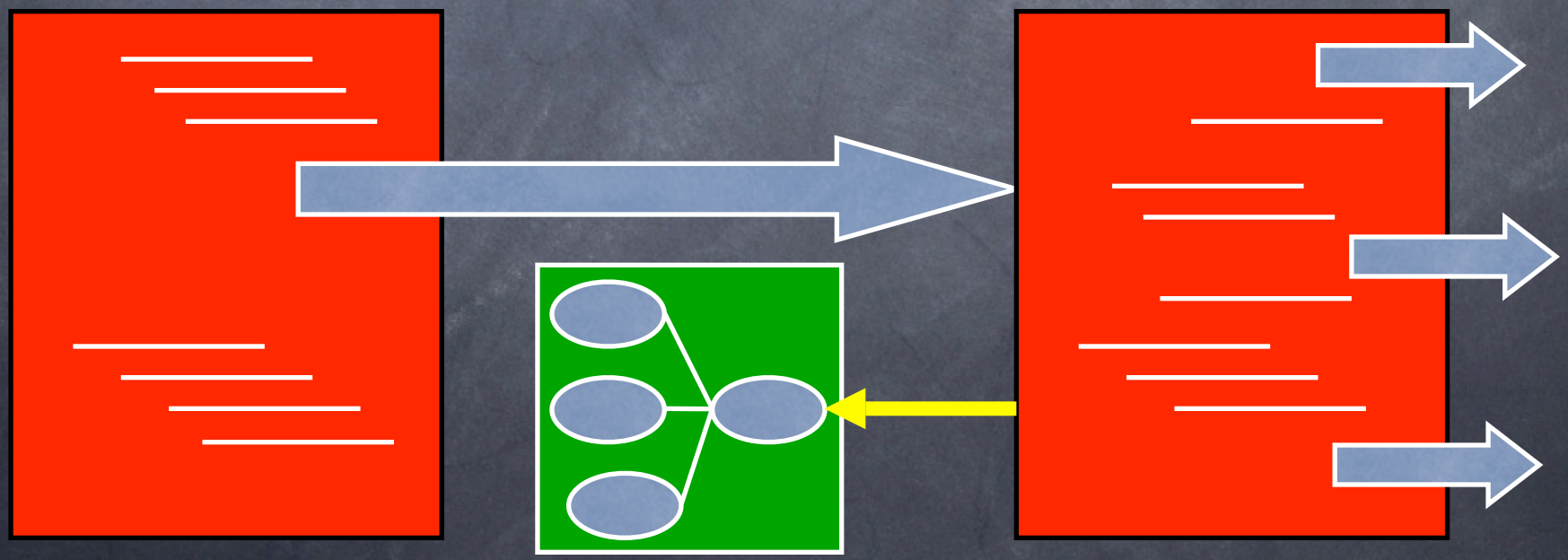
$$\lambda_l = nnet(in_1, \ldots, in_N)$$

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

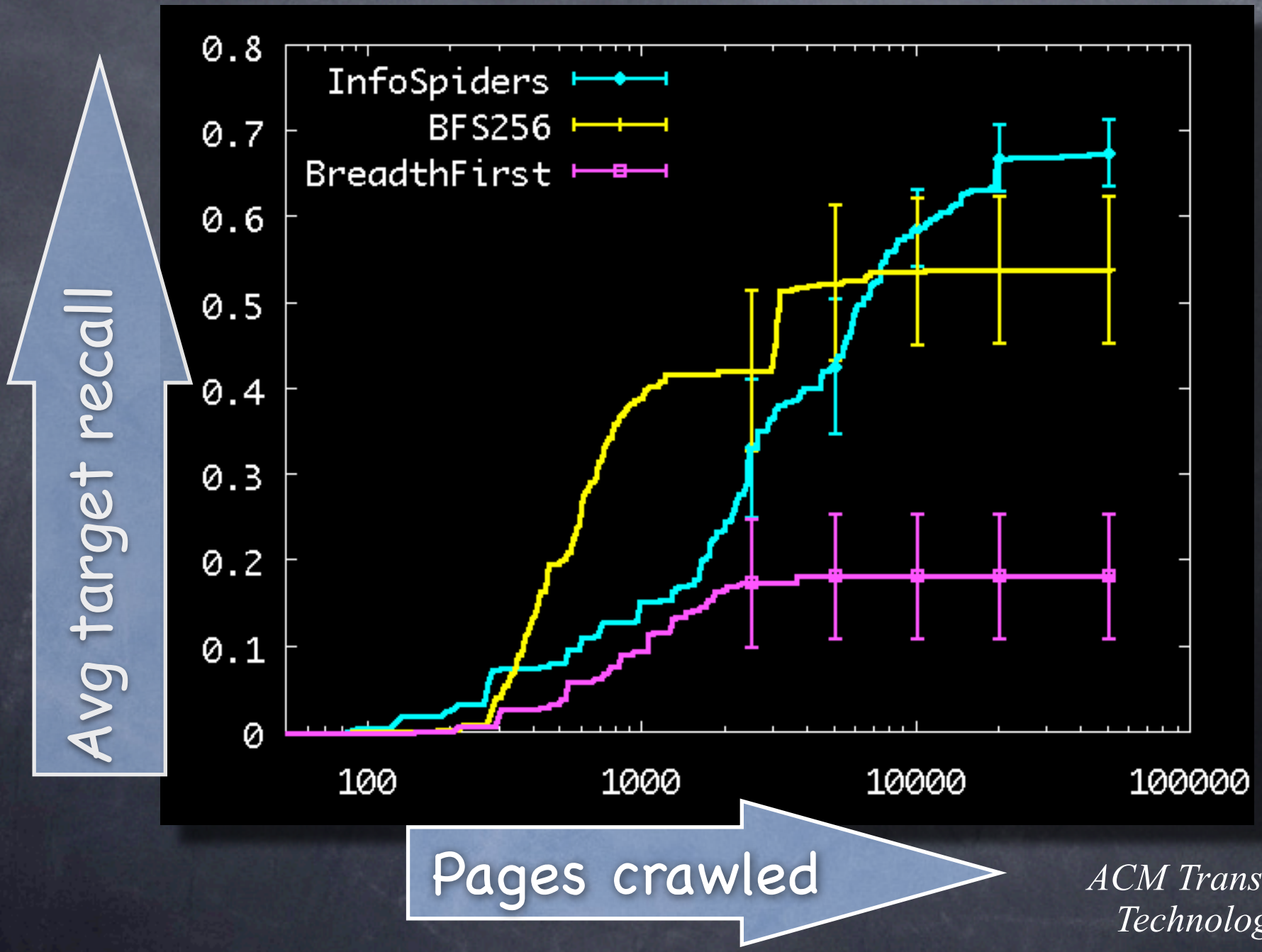


Q-learning

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

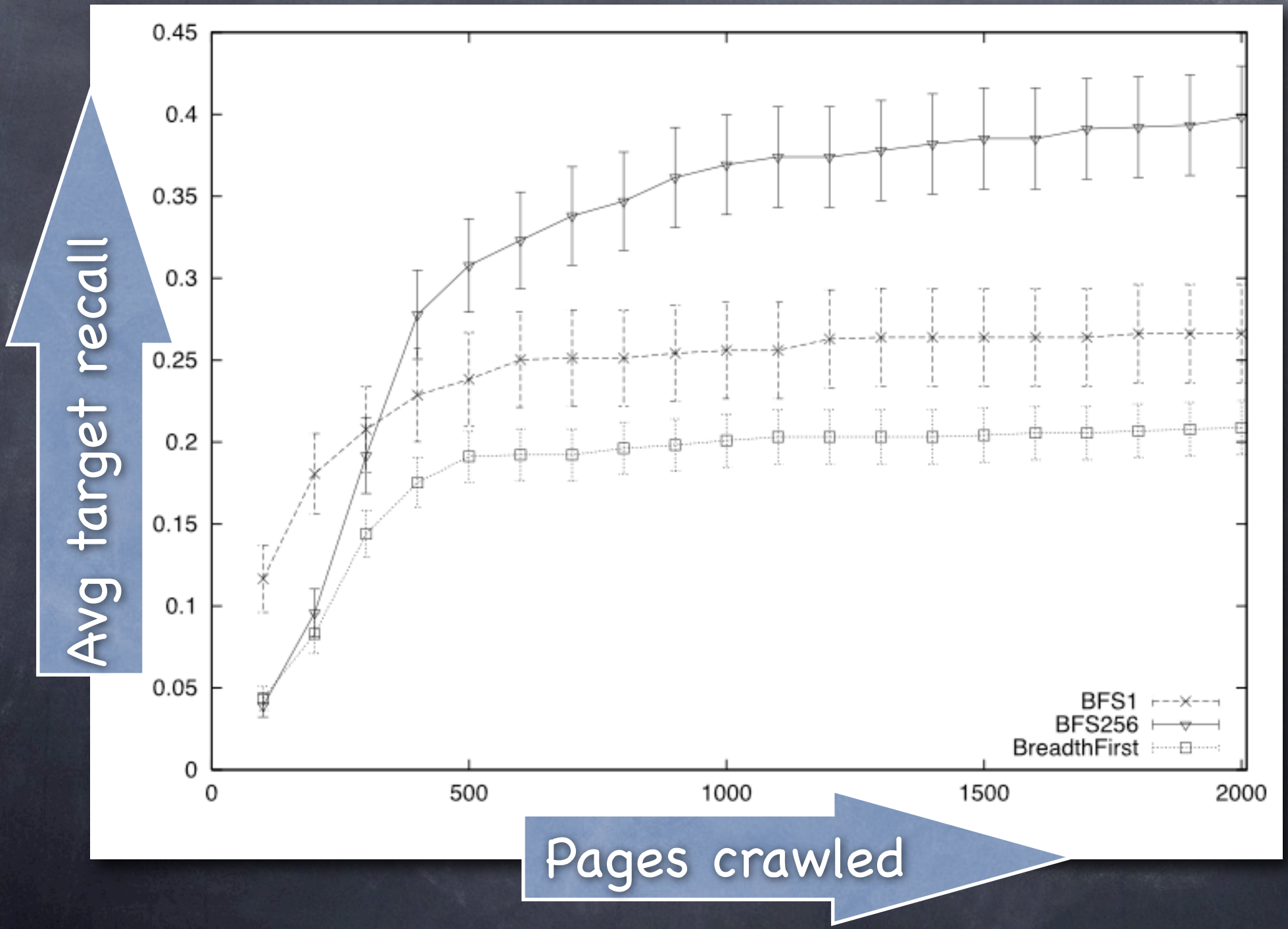


Performance

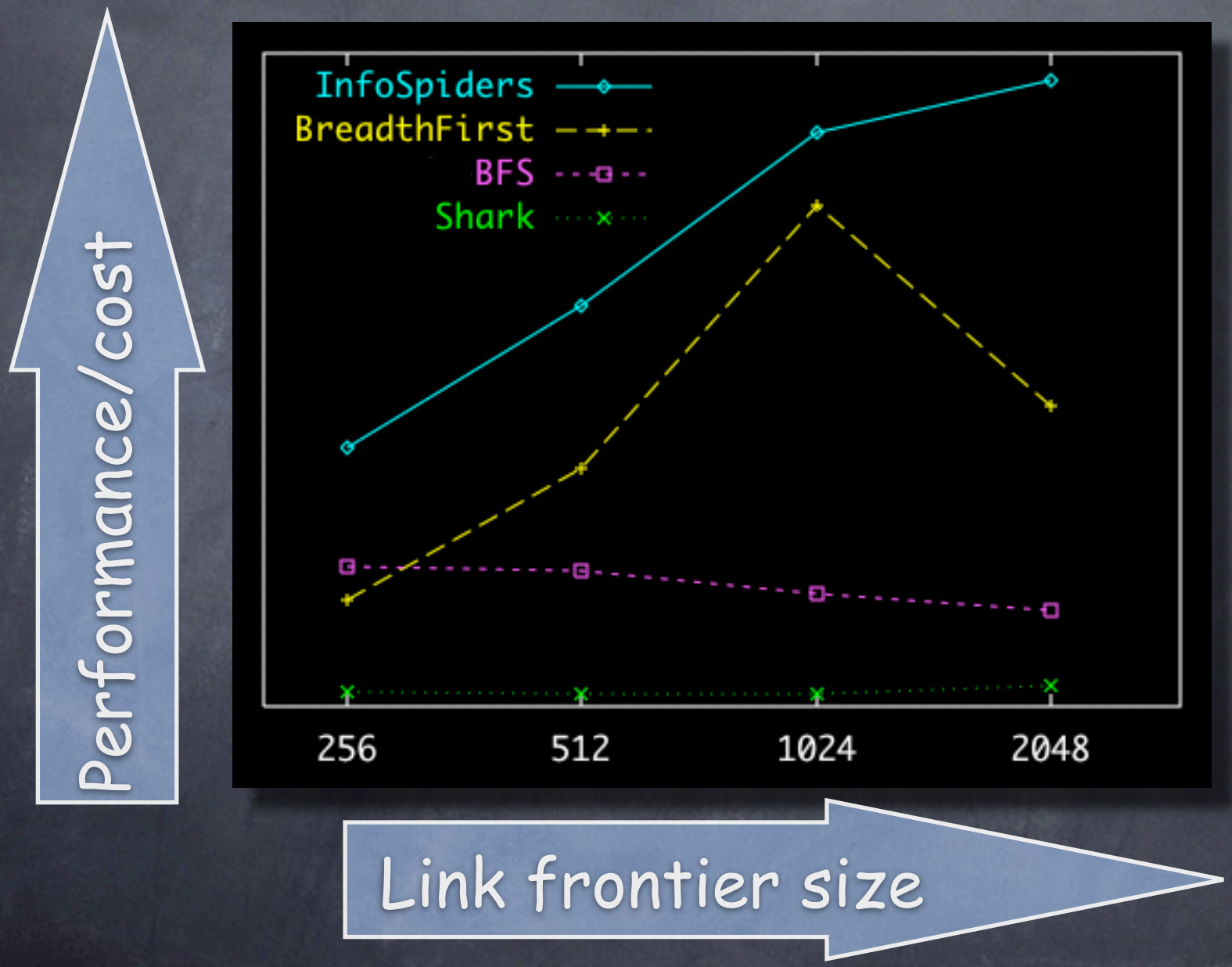


ACM Trans. Internet Technology 2003

Exploration vs. Exploitation



Efficiency & scalability



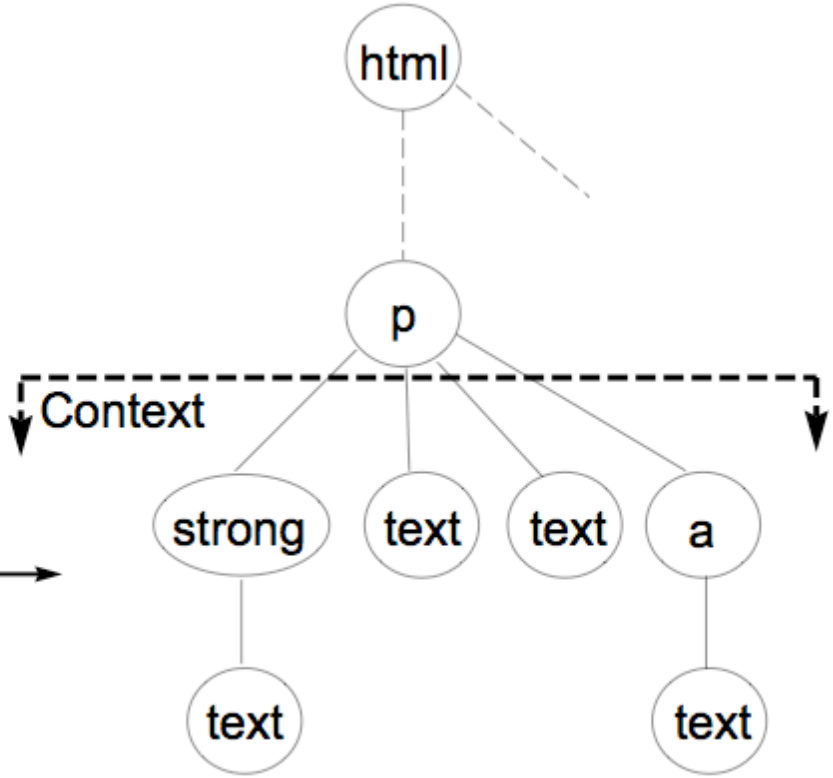
DOM context

Link score = linear combination between page-based and context-based similarity score

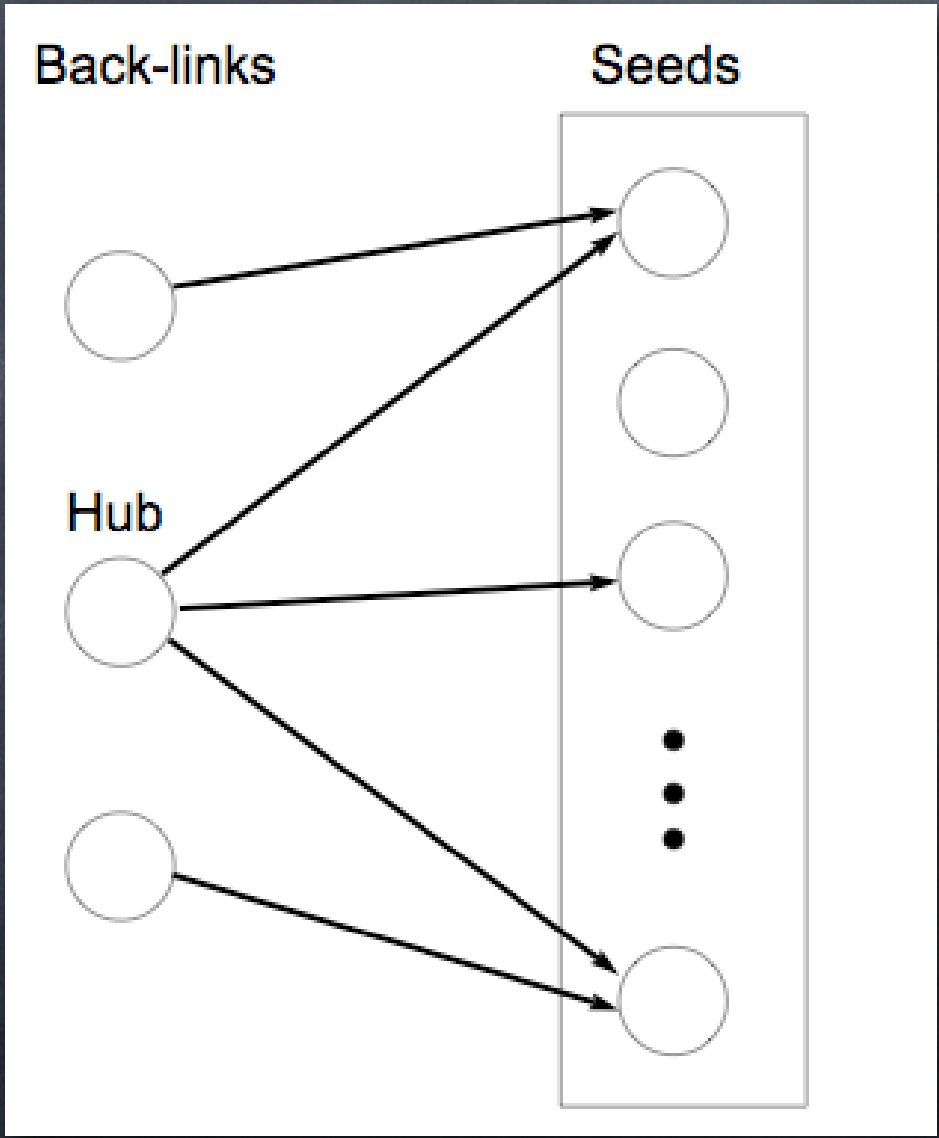
About Exelixis
Exelixis, Inc. is a leading genomics-based drug discovery company focused on product development through its expertise in comparative genomics and model system genetics. These technologies provide a rapid, efficient and cost effective way to move from DNA sequence data to knowledge about the function of genes and the proteins they encode. The company's technology is broadly applicable to all life sciences industries including pharmaceutical, diagnostic, agricultural biotechnology and animal health. Exelixis has partnerships with Aventis CropScience S.A., Bayer Corporation, Bristol-Myers Squibb Company, Elan Pharmaceuticals, Inc., Pharmacia Corporation, Protein Design Labs, Inc., Scios Inc. and Dow AgroSciences LLC, and is building its internal development program in the area of oncology. For more information, please visit the company's web site at www.exelixis.com.


```
<P class=MsoNormal>
<STRONG>About Exelixis</STRONG><BR>Exelixis, Inc. is a leading genomics-based drug discovery
company focused on product development through its expertise in comparative genomics and model system
genetics. These technologies provide a rapid, efficient and cost effective way to move from DNA sequence
data to knowledge about the function of genes and the proteins they encode. The company's technology is
broadly applicable to all life sciences industries including pharmaceutical, diagnostic, agricultural
biotechnology and animal health. Exelixis has partnerships with Aventis CropScience S.A., Bayer Corporation,
Bristol-Myers Squibb Company, Elan Pharmaceuticals, Inc., Pharmacia Corporation, Protein Design Labs,
Inc., Scios Inc. and Dow AgroSciences LLC, and is building its internal development program in the area of
oncology.<SPAN  style='mso-spacerun: yes'>  </SPAN>For more information, please visit the
company's web site at
<A href="http://www.exelixis.com/">www.exelixis.com</A>.<o:p></o:p>
</P>
```

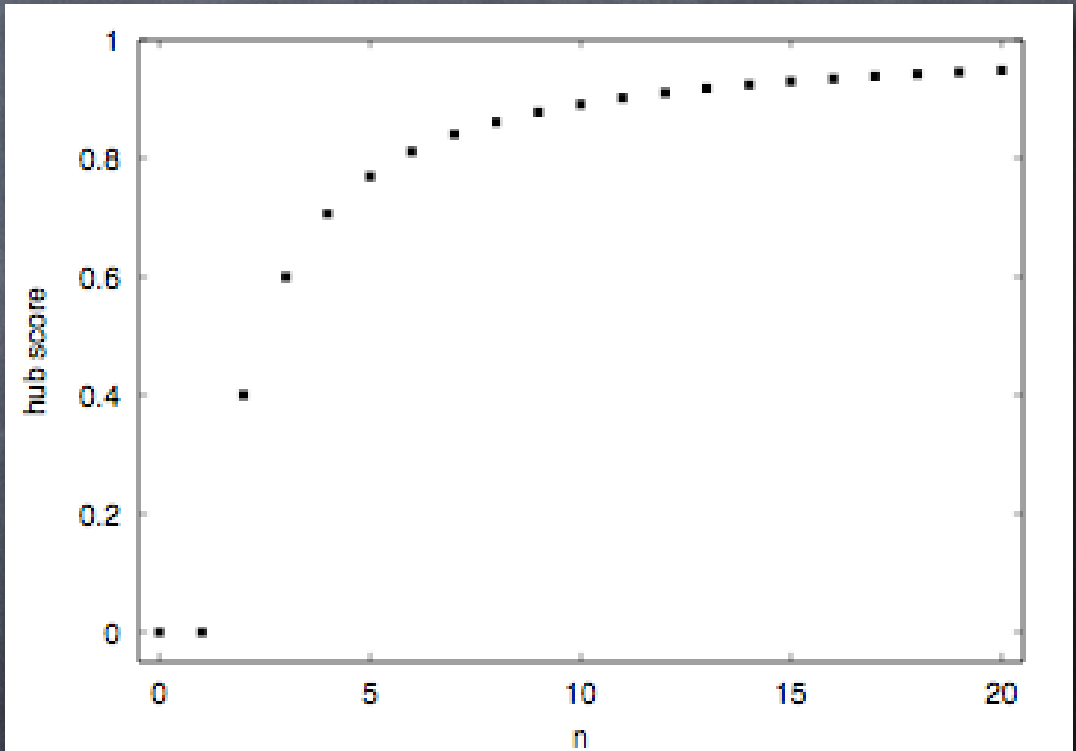
```
- <p>
  - <strong>
      <text>about exelixis</text>
    </strong>
    <text>exelixis inc is a leading genomics based drug discovery company focused on
      product development through its expertise in comparative genomics and model
      system genetics these technologies provide a rapid efficient and cost effective
      way to move from dna sequence data to knowledge about the function of genes
      andthe proteins they encode the company s technology is broadly applicable to
      all life sciences industries including pharmaceutical diagnostic agricultural
      biotechnology and animal health exelixis has partnerships with aventis
      cropscience s a bayer corporation bristol myers squibb company elan
      pharmaceuticals inc pharmacia corporation protein design labs inc scios inc and
      dow agrosciences llc and is building its internal development program in the area
      of oncology</text>
    <text>for more information please visit the company s web site at</text>
  - <a href="http://www.exelixis.com/">
      <text>www exelixis com</text>
    </a>
  </p>
```

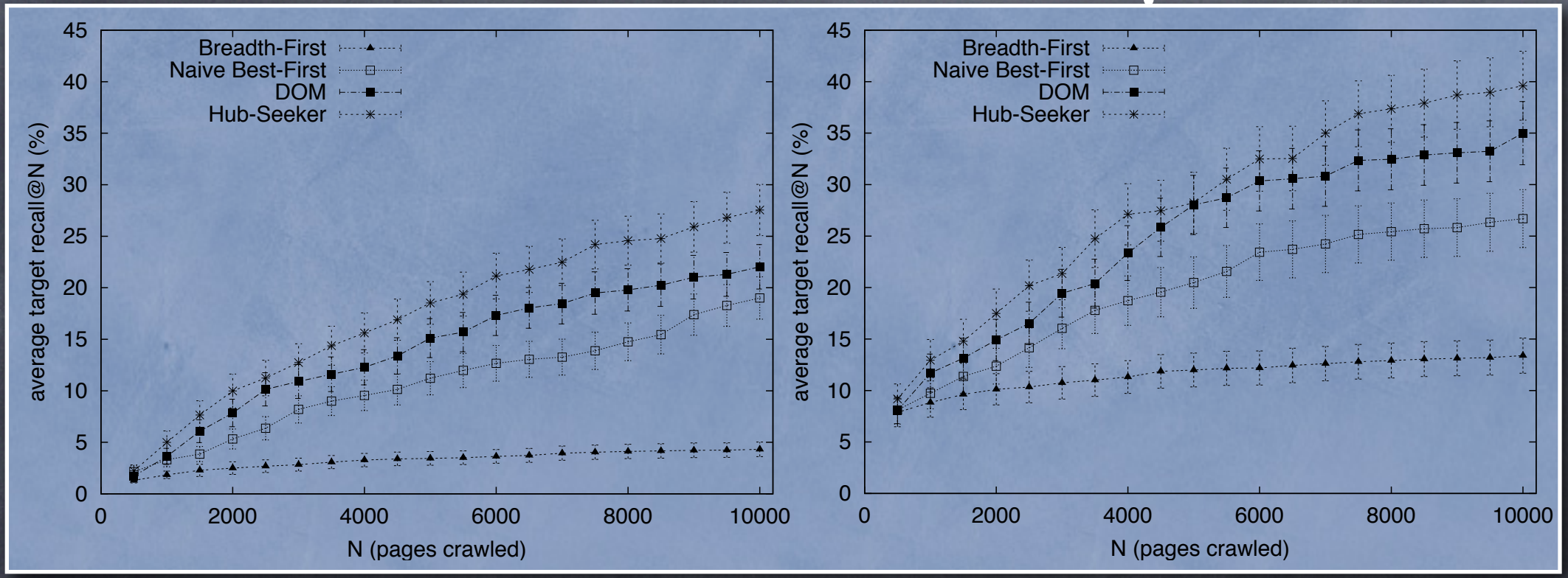
Co-citation: hub scores



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

Normative lessons about crawlers

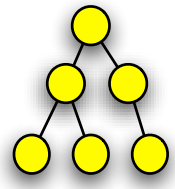
- Remember links from previous pages
- Exploration is important
- Adaptation can help
  - to focus search
  - to expand query
  - to learn link estimates
- Distributed algorithms
  - boost efficiency
  - but watch for premature exploitation
- Improve link evaluation by looking at
  - parent pages
  - DOM context

Discussion: Which crawler algorithm would you use in your app?

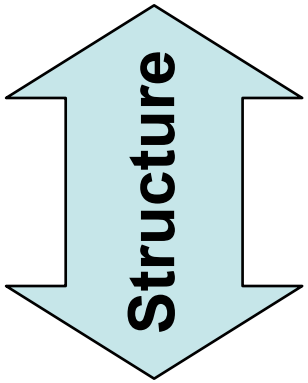
Outline

- ✓ Mapping
  - Topical locality
  - Content, link, and semantic topologies in the Web
- ✓ Modeling
  - How the Web evolves and why content matters
  - Consequences for navigation and crawling
- ✓ Mining
  - Topical Web crawlers
  - Adaptive, intelligent crawling techniques
- Mingling
  - Social Web search & recommendation
  - Distributed collaborative peer search

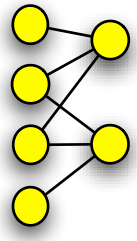
76788 links
1015 donations
117 registered users
Last updated: Apr 17, 2006



hierarchical

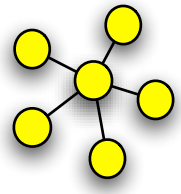


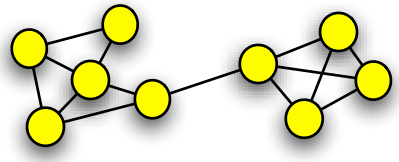
flat



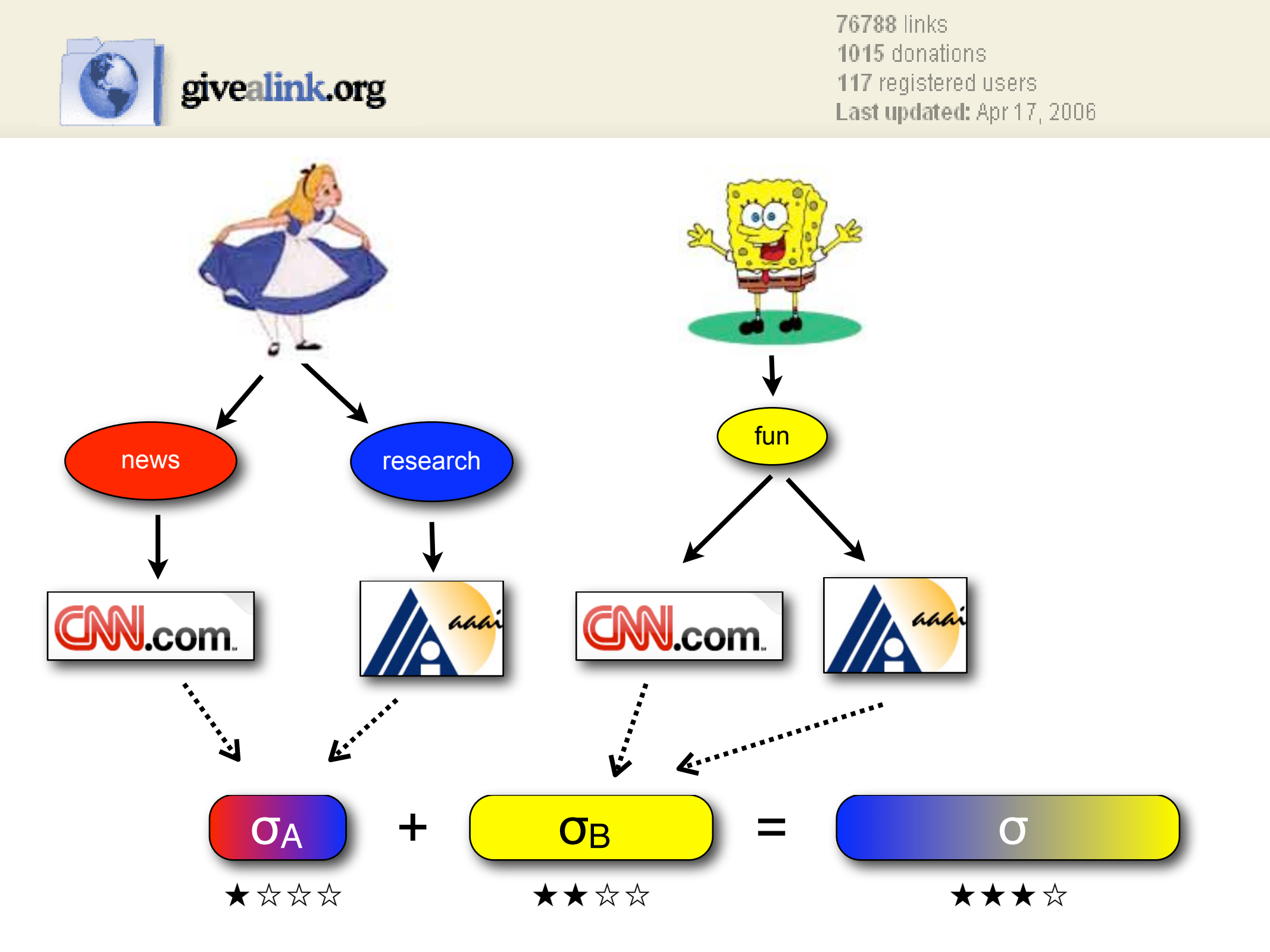
centralized

social



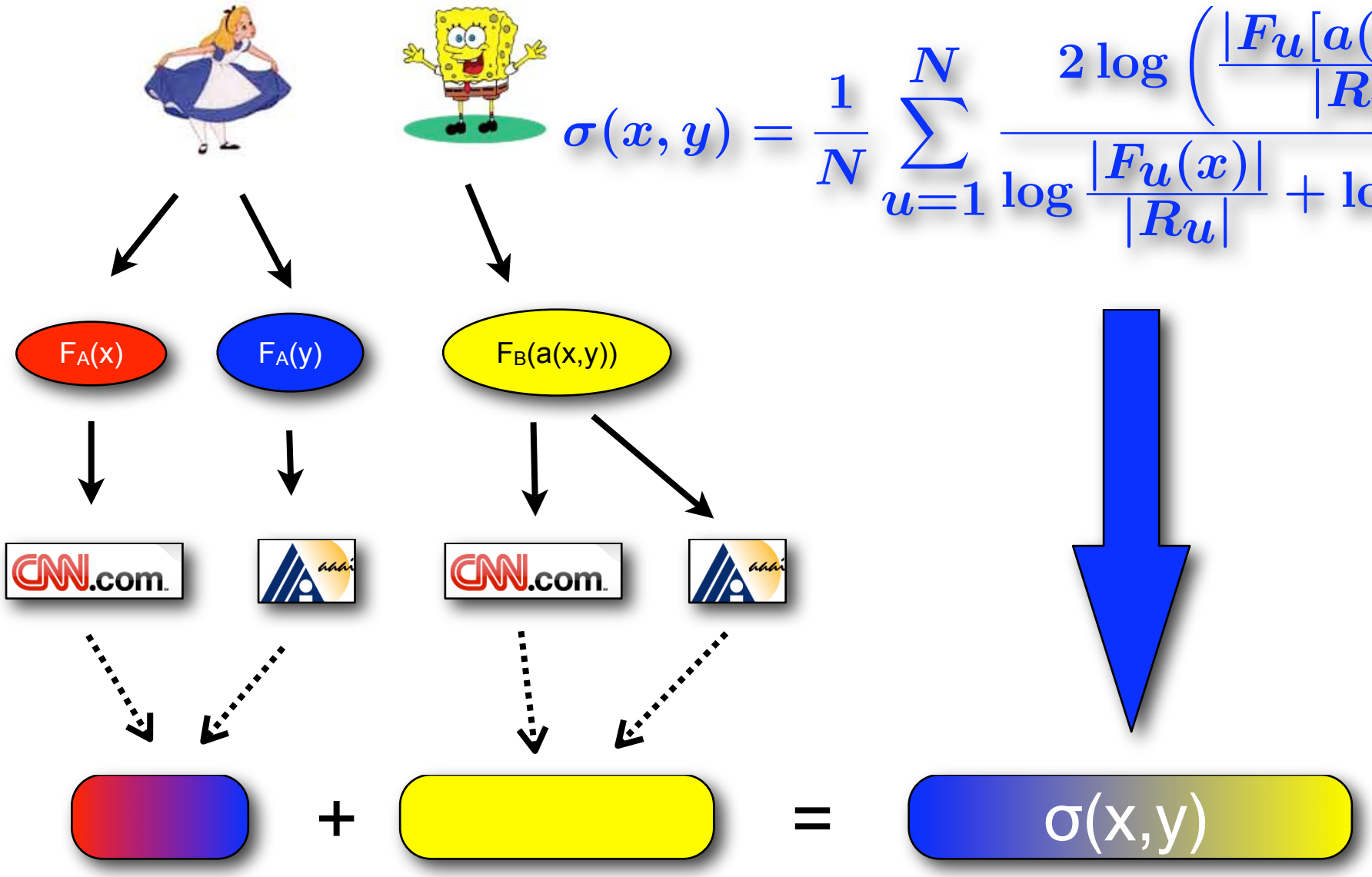


WWW2006, AAAI2006


givealink.org
76788 links
1015 donations
117 registered users
Last updated: Apr 17, 2006
news
research
fun
CNN.com
aaai
CNN.com
aaai
σA
+
σB
=
σ

givealink.org

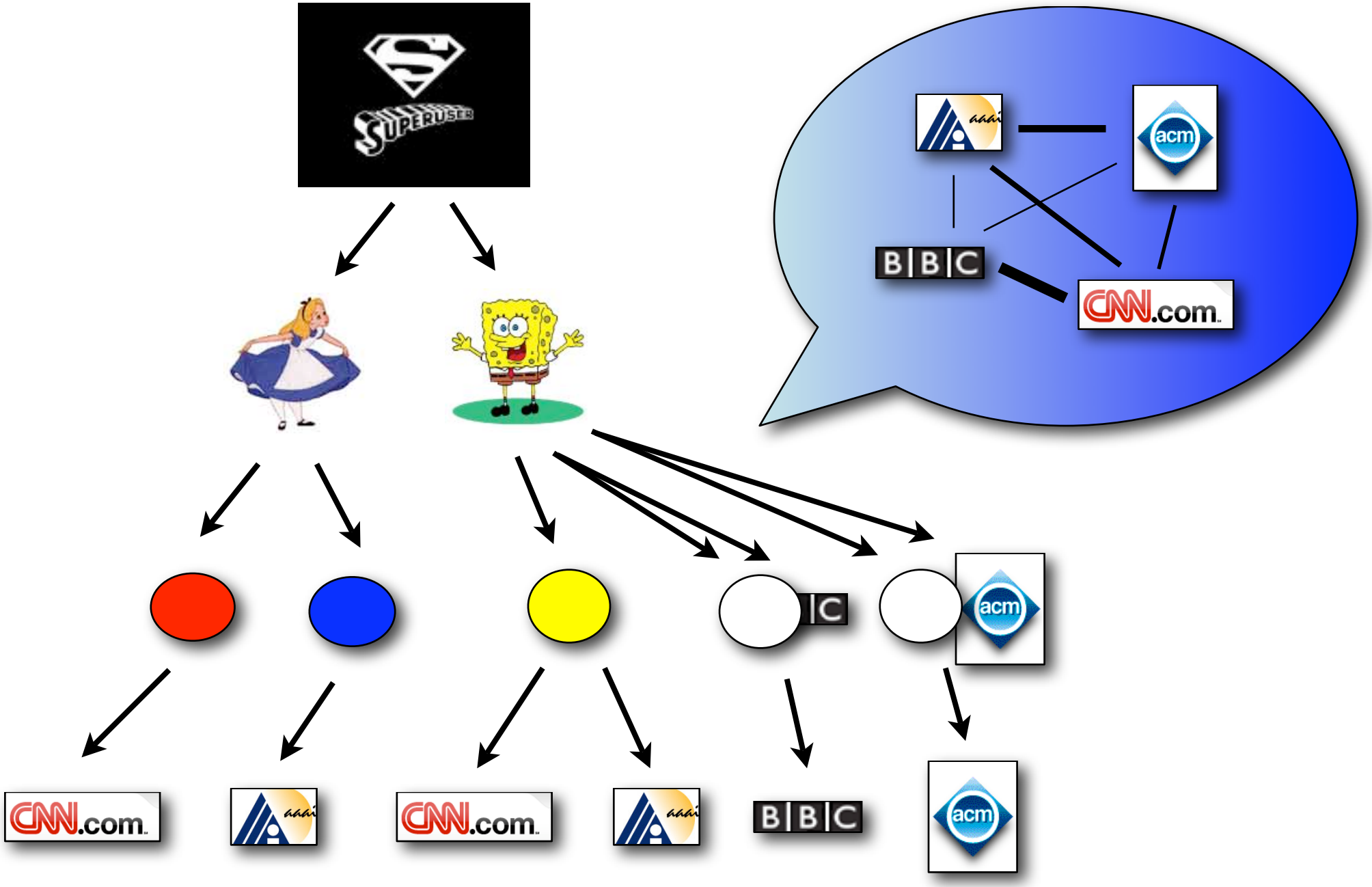
76788 links
1015 donations
117 registered users
Last updated: Apr 17, 2006

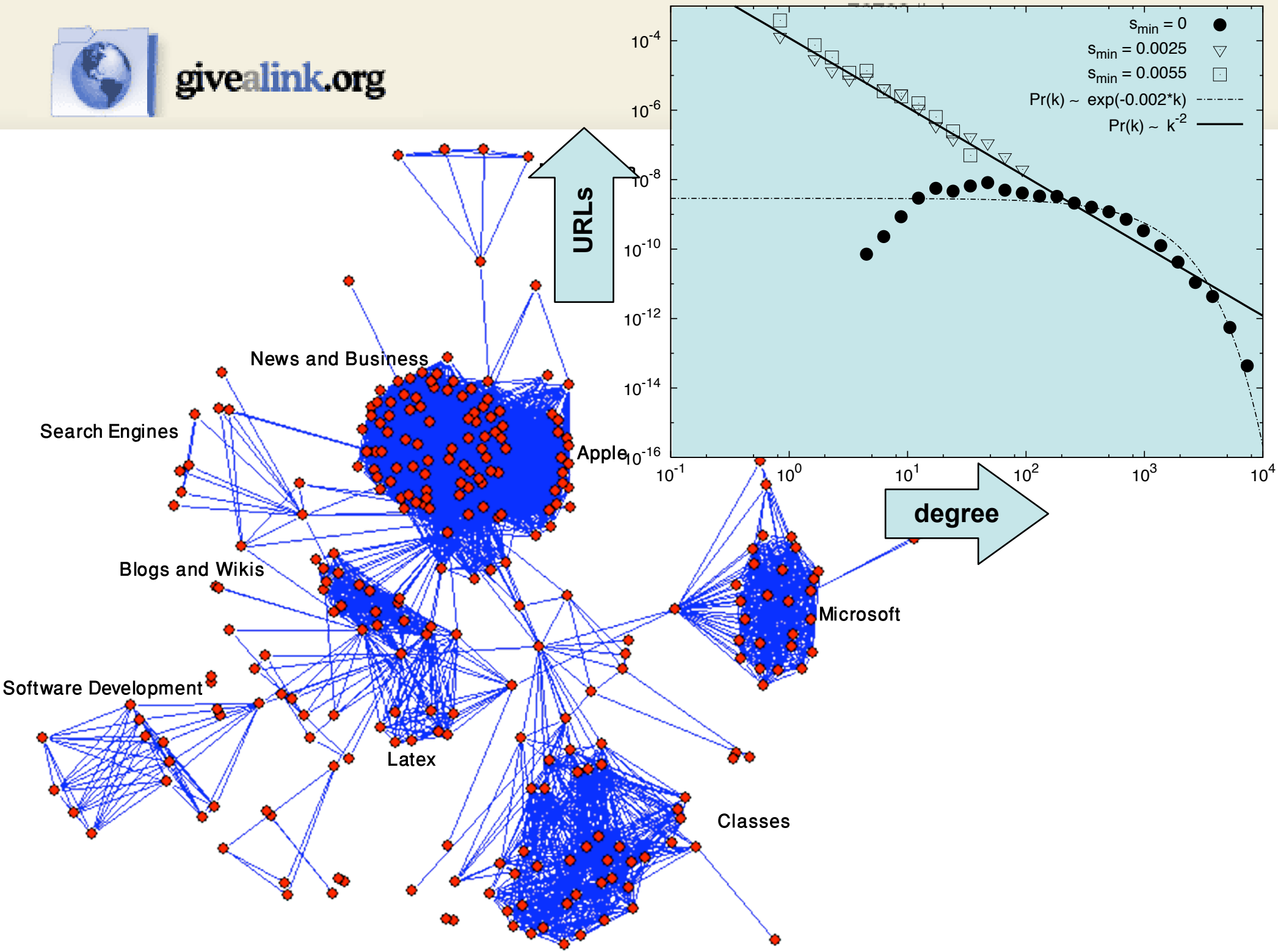


$$\sigma(x,y) = \frac{1}{N}\sum_{u=1}^{N} \frac{2\log\left(\frac{|F_u[a(x,y)]|}{|R_u|}\right)}{\log\frac{|F_u(x)|}{|R_u|} + \log\frac{|F_u(y)|}{|R_u|}}$$

givealink.org

76788 links
1015 donations
117 registered users
Last updated: Apr 17, 2006




givealink.org
URLs
degree
$s_{min} = 0$
$s_{min} = 0.0025$
$s_{min} = 0.0055$
Pr(k) ~ exp(-0.002*k)
Pr(k) ~ k^{-2}
News and Business
Search Engines
Apple
Blogs and Wikis
Microsoft
Software Development
Latex
Classes

givealink.org

76788 links
1015 donations
117 registered users
Last updated: Apr 17, 2006

Home Apps FAQ News About Us

Link Recommender

Enter URL or keyword to get related bookmarks. (More Options)

webir.org search

Sort by Bookmark: ☑ Similarity ? ☐ Novelty ? ☐ Prestige ?

Similar URLs sorted by Similarity **1-10** of **360** recommendations for **http://www.webir.org/**

http://www.parc.xerox.com/istl/groups/iea/dynamics.shtml
http://www.parc.xerox.com/istl/groups/iea/dynamics.shtml

http://www.springer.de/cgi-bin/search_book.pl?isbn=3-540-65112-8
http://www.springer.de/cgi-bin/search_book.pl?isbn=3-540-65112-8

Web Research Collections - Web Track
http://es.csiro.au/TRECWeb/

DIMACS Workshop on Internet and WWW Measurement, Mapping and Modeling
http://dimacs.rutgers.edu/Workshops/Internet/

http://www.ibm.com/java/fetuccino/
http://www.ibm.com/java/fetuccino/

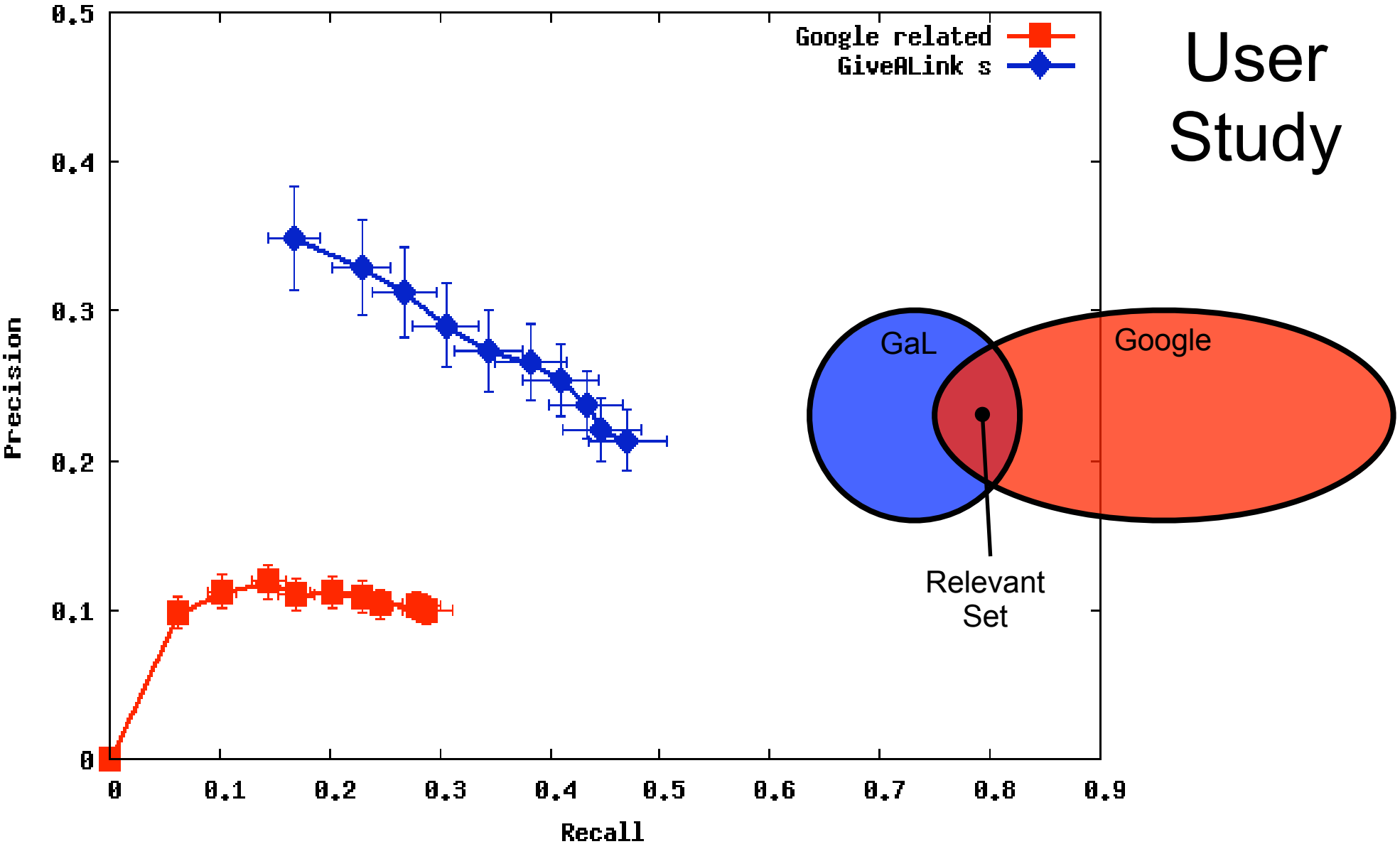
Web Term Document Frequency Form
http://elib.cs.berkeley.edu/docfreq/index.html

Finding Out About
http://www.cs.ucsd.edu/~rik/foa/

Bibliometrics of the World Wide Web: An Exploratory Analysis of the Intellectual Structure of Cyberspace
http://sherlock.berkeley.edu/asis96/asis96.html

Terabyte TREC Homepage
http://www-nlpir.nist.gov/projects/terabyte/

http://www.neci.nj.nec.com/homepages/lawrence/websize.html
http://www.neci.nj.nec.com/homepages/lawrence/websize.html


givealink.org
76788 links
1015 donations
117 registered users
Last updated: Apr 17, 2006
User Study
Google related
GiveALink s
Precision
Recall
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
GaL
Google
Relevant Set

givealink.org

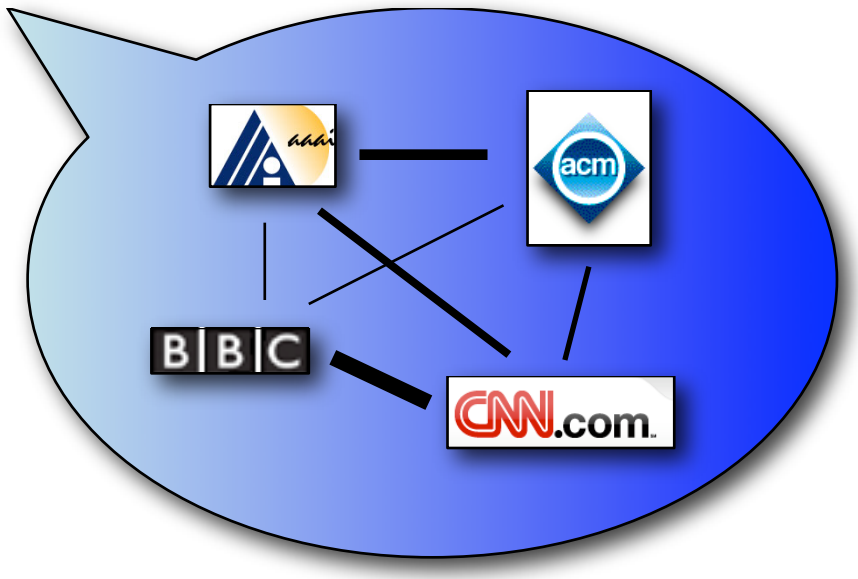
76788 links
1015 donations
117 registered users
Last updated: Apr 17, 2006

$$c(x) = \frac{1}{|U|} \sum_{y \in U} \left[1 + \min_{x \rightsquigarrow y} \sum_{(u,v) \in x \rightsquigarrow y} \left(\frac{1}{\sigma(u,v)} - 1 \right) \right]^{-1}$$

Centrality

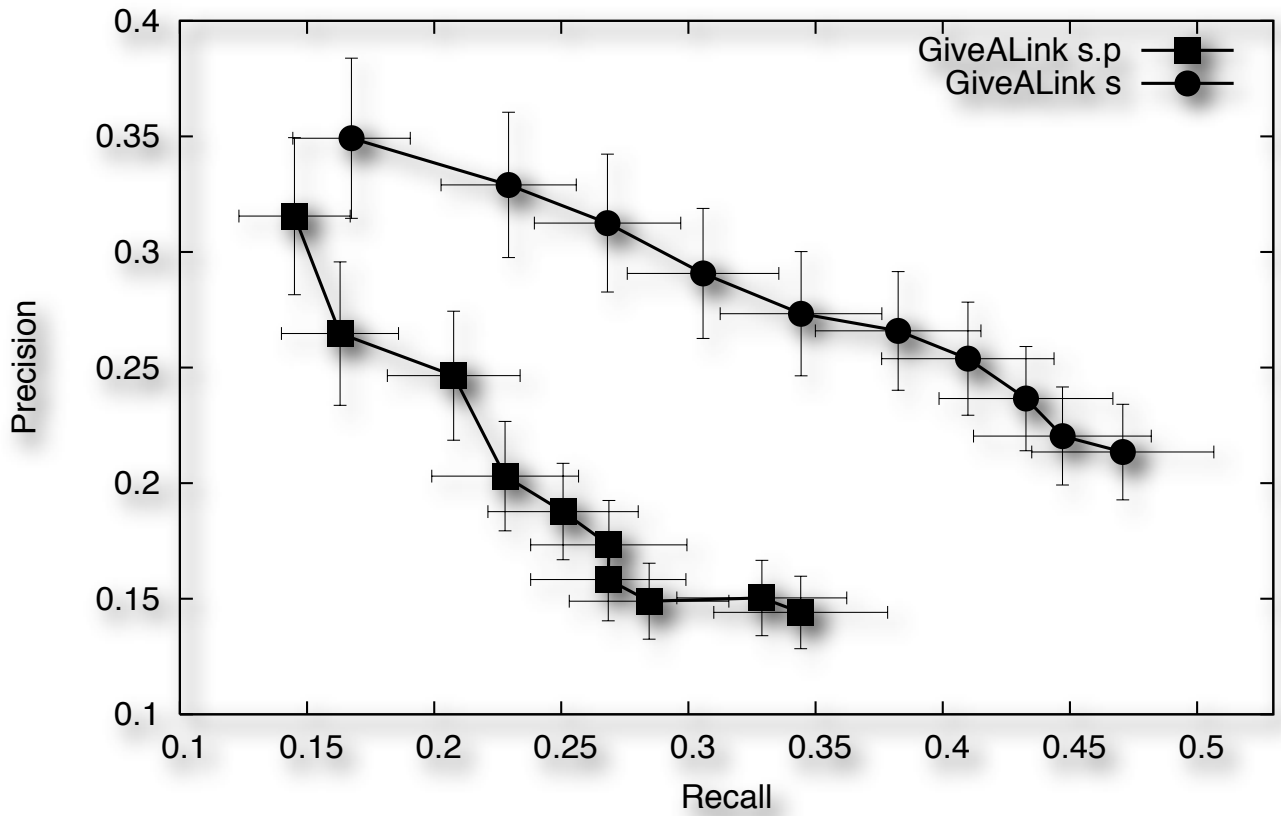
&

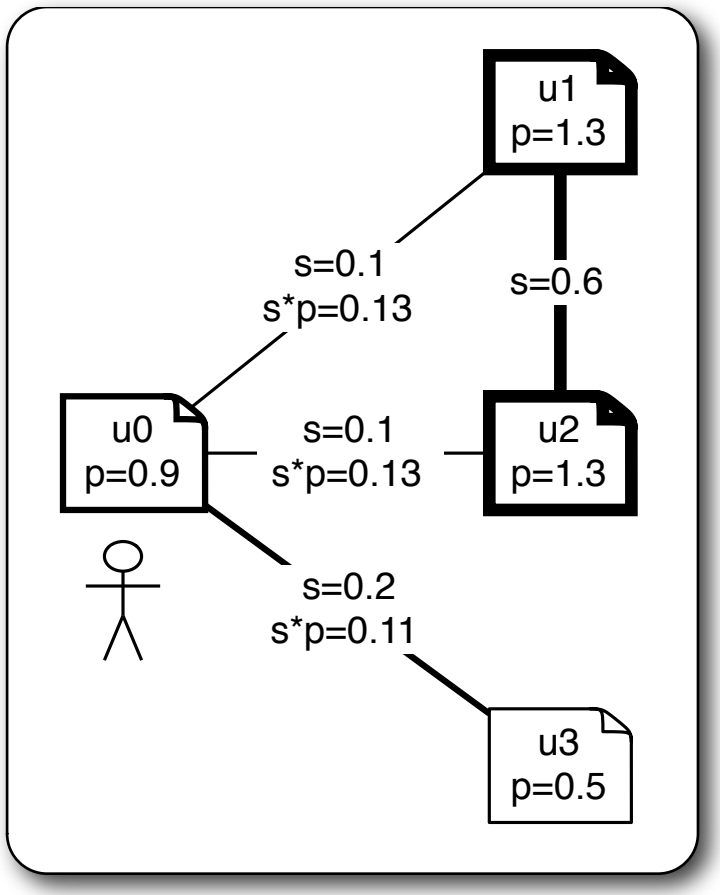
Prestige

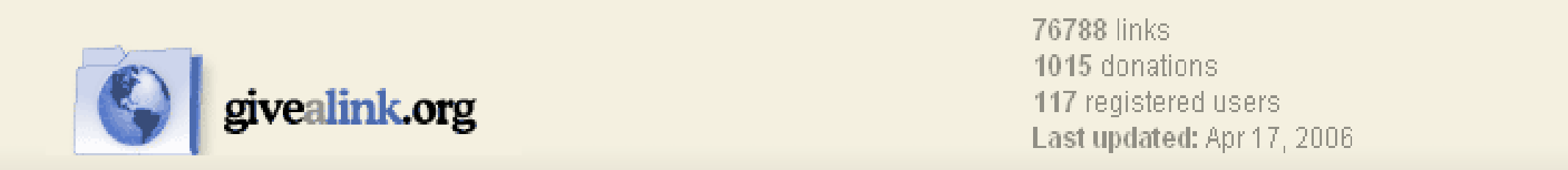


$$p_{t+1}(x) = (1 - \alpha) + \alpha \cdot \sum_{y \in U} \frac{\sigma(x,y) \cdot p_t(y)}{\sum_{z \in U} \sigma(y,z)}$$

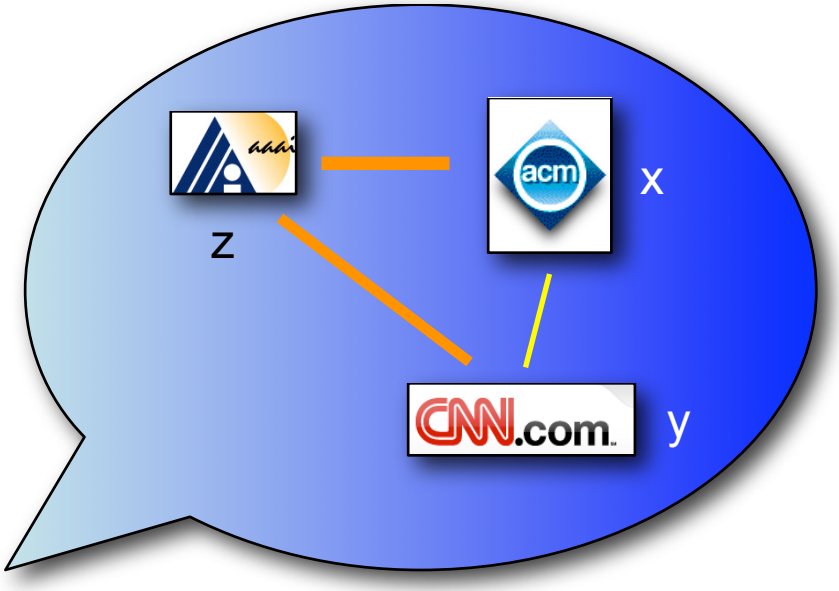
Ranking by Similarity * Prestige







Ranking & Recommendation by Novelty



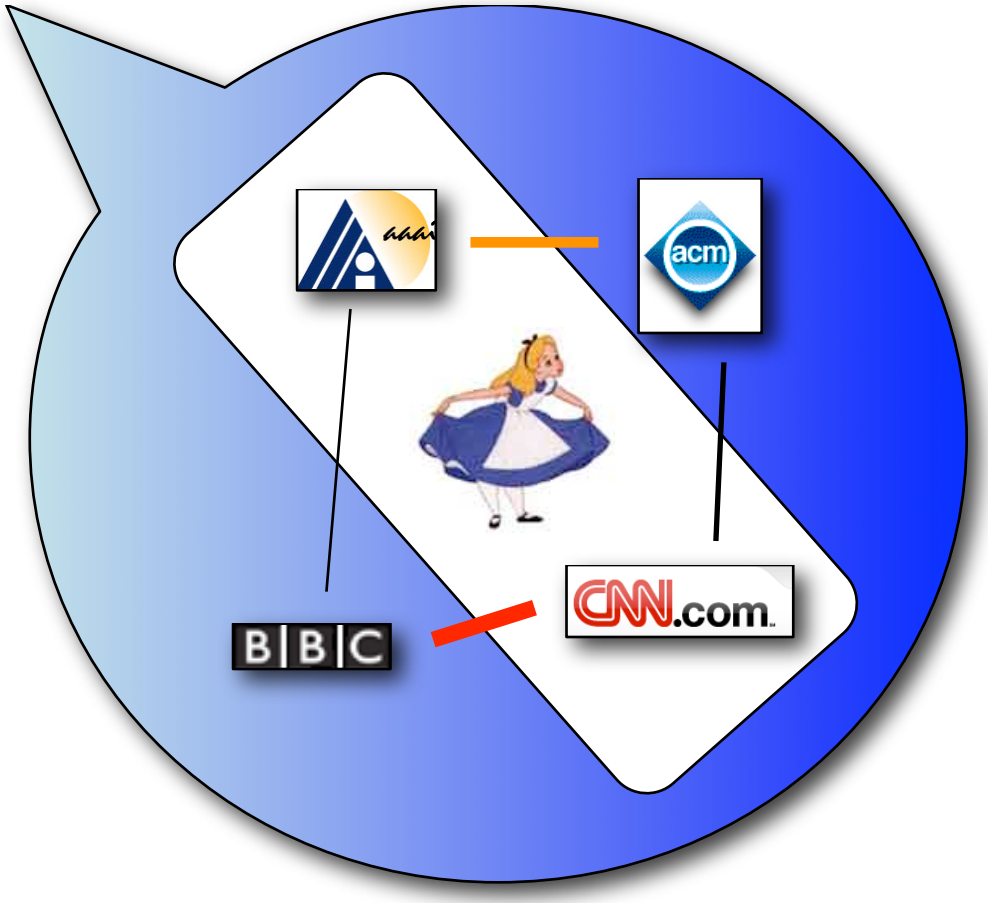
$$\nu(x,y) = \frac{\left[1 + \min_{z \in U}\left(\frac{1}{\sigma(x,z)} + \frac{1}{\sigma(z,y)} - 2\right)\right]^{-1}}{\sigma(x,y)}$$

givealink.org

76788 links
1015 donations
117 registered users
Last updated: Apr 17, 2006

$$\eta(x, A) = \max_{y \in A} \left[\sigma(x, y) \cdot \log \left(\frac{N}{N(y)} \right) \right]$$

Personalization



givealink.org

76788 links
1015 donations
117 registered users
Last updated: Apr 17, 2006

collaborative filtering, social semantic similarity, unlinked pages, multimedia content, trust

scalability, density

spam

givealink.org

I donated my bookmarks to science. Did you?

Username: fil@indiana.edu Register ?

Password: •••••••• Log In

About: Project Help FAQ Privacy **Links:** Give Manage Download RSS

General

http:// Related Links

keywords Related Links

Surprise Me

http:// Novel Links

keywords Novel Links

Getting Started

1. Register
2. Donate your bookmarks to science
3. Get personalized recommendations/search results
   Also manage your bookmarks or add a new URL while you browse
4. Spread the word!
   More users means more data means better results!

Top Ten Bookmarks ?

1. http://www.iuma.com/
2. Sonicnet.com
3. Error
4. the prog organ - progressive rock reviews
5. Hard Rock and Heavy Metal Radio - Video
6. GODS OF MUSIC - Music Reviews For The Independent Music Scene
7. RUTHLESS REVIEWS: MUSIC
8. Welcome! -- Rate Your Music
9. Tiny Mix Tapes
10. Reviews of Indie Albums on Irish music webzine CLUAS.com

Stats

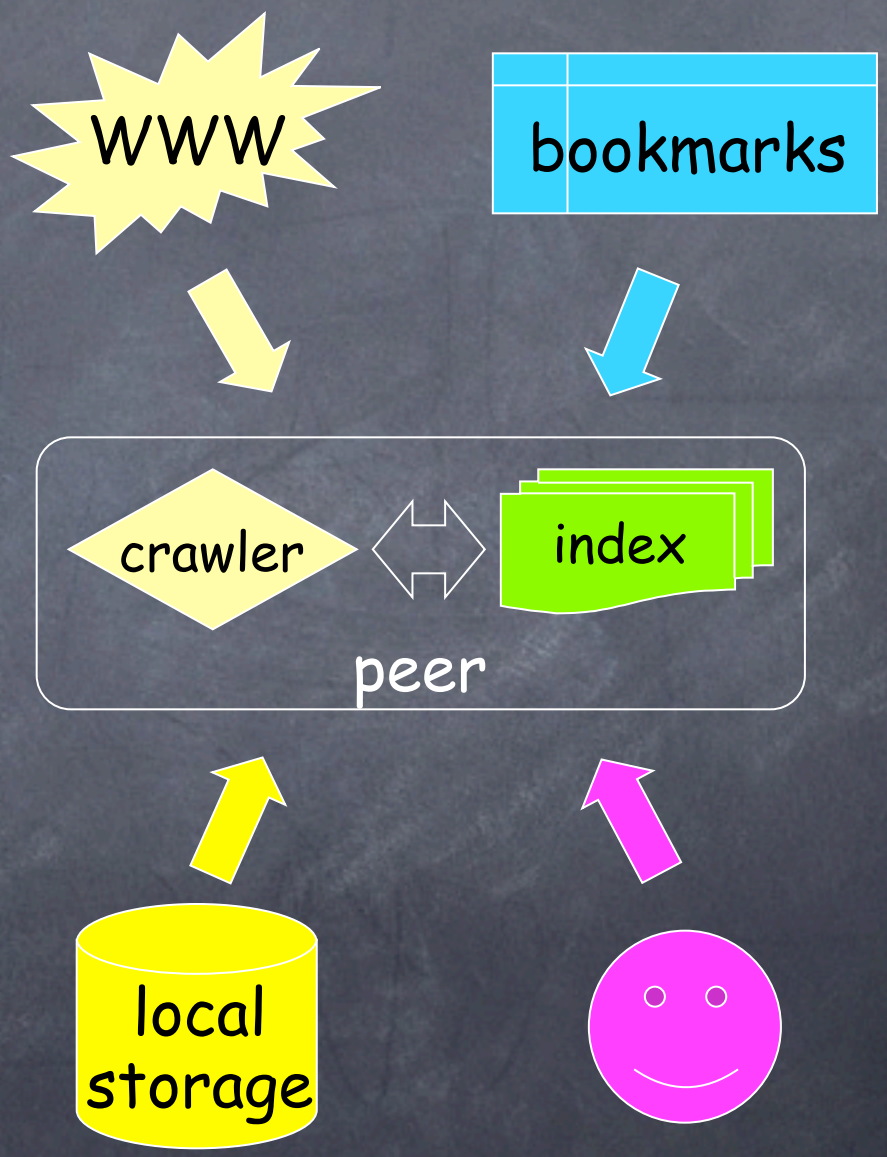
81949 links
3277880 relationships
1049 donations
184 registered users
Last Updated: Sat Aug 5 16:10:58 2006

Donate!
givealink.org

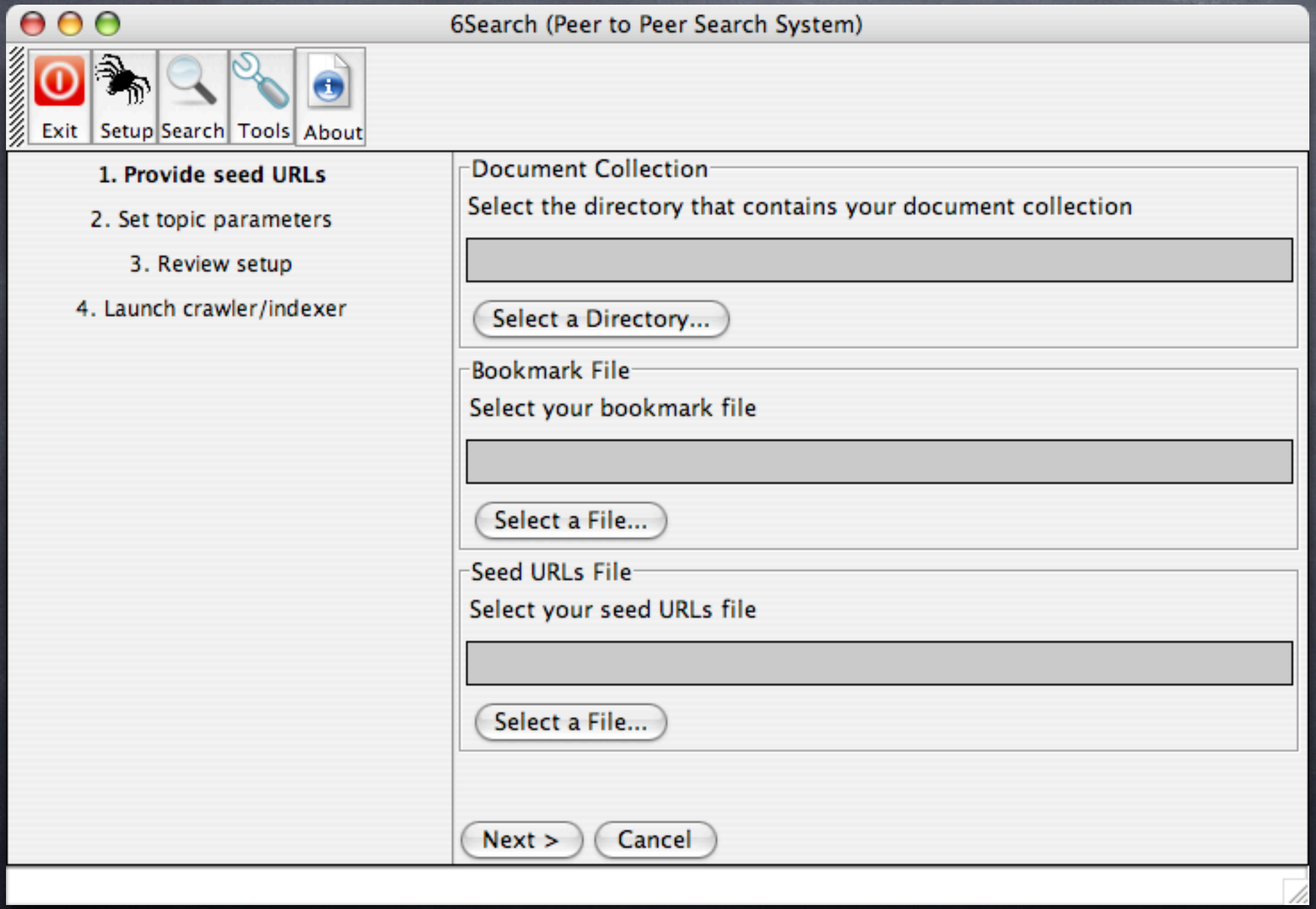
Discussion: Critical mass?

Peer distributed crawling and collaborative searching

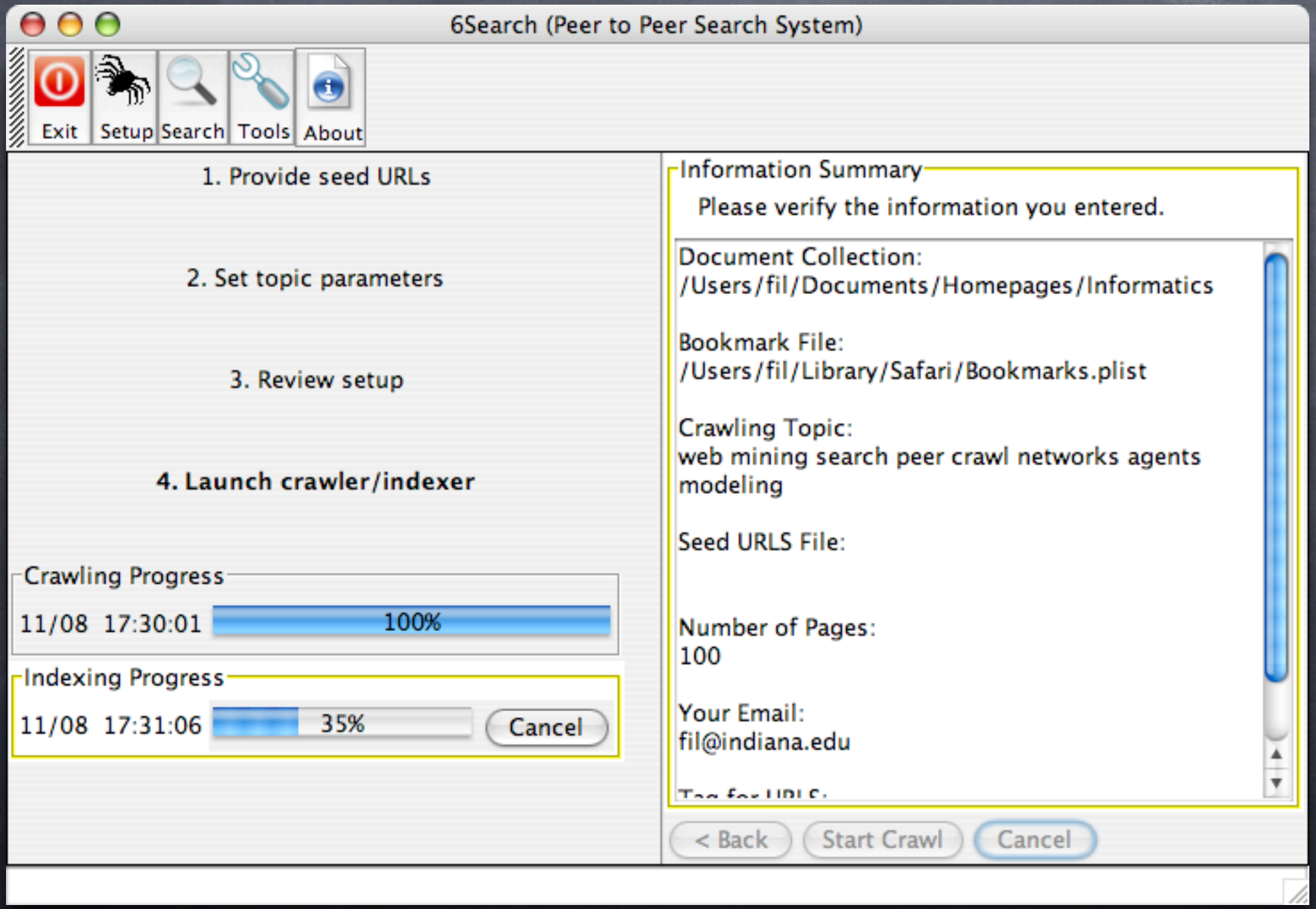
- Crawl when idle
  - Start from bookmarks
  - Past queries or selected documents as topics
- Index locally
  - User relevance feedback is natural
  - No centralized coordination bottleneck
    - Unlike grub.org, hyperbee.com



http://homer.informatics.indiana.edu/~nan/6S/



http://homer.informatics.indiana.edu/~nan/6S/



http://homer.informatics.indiana.edu/~nan/6S/

6Search (Peer to Peer Search System)

Exit | Setup | Search | Tools | About

Indexing Date	Document Title	Document Url	Document Tag
2006/11/08 ...	SiliconBeat: News about tech money and in...	http://www.siliconbeat.com/	web mining search pe...
2006/11/08 ...	Investment Management Research Services	http://www.round.table.com/i...	web mining search pe...
2006/11/08 ...	VentureBeat » Bold start-up, Powerset, ab...	http://venturebeat.com/2006/...	web mining search pe...
2006/11/08 ...	Yahoo!	http://www.yahoo.com/	web mining search pe...
2006/11/08 ...	The J Curve	http://jurvetson.blogspot.com/	web mining search pe...
2006/11/08 ...		http://www.webconferences.co...	web mining search pe...
2006/11/08 ...	Search Boards for threads - Forum search ...	http://www.boardtracker.com/...	web mining search pe...
2006/11/08 ...	Techcrunch » Blog Archive » Will Powerset ...	http://www.techcrunch.com/20...	web mining search pe...
2006/11/08 ...	Jeff Clavier	http://softtechvc.blogs.com/ab...	web mining search pe...
2006/11/08 ...	Challengelist	http://www.research.att.com/%...	web mining search pe...
2006/11/08 ...		http://www.gutenberg.org/dirs...	web mining search pe...
2006/11/08 ...		http://www.w3.org/WAI/UA/W...	web mining search pe...
2006/11/08 ...	My AOL	http://feeds.my.aol.com/?url=...	web mining search pe...
2006/11/08 ...	Donor Suitability Workshop: Donor History ...	http://www.fda.gov/cber/minu...	web mining search pe...
2006/11/08 ...	ODP - Open Directory Project	http://www.dmoz.org/	web mining search pe...
2006/11/08 ...	The CC Chemokine Thymus-derived Chem...	http://www.jem.org/cgi/conten...	web mining search pe...
2006/11/08 ...	absintelagent : Messages : 258-287 of 287	http://tech.groups.yahoo.com/...	web mining search pe...
2006/11/08 ...	Survey: Information Gathering and Knowled...	http://www.cio.com/research/...	web mining search pe...
2006/11/08 ...	British Blogs	http://www.britishblogs.co.uk/	web mining search pe...
2006/11/08 ...	NJDEP Compliance & Enforcement - A decl...	http://www.state.nj.us/dep/en...	web mining search pe...
2006/11/08 ...	Barney Pell's Weblog: Search Archives	http://www.barneypell.com/ar...	web mining search pe...
2006/11/08 ...	Elatable	Bradley Horowitz	http://www.elatable.com/blog/	web mining search pe...

Advanced Tool | Delete row | Undelete All | Delete All

http://homer.informatics.indiana.edu/~nan/6S/

6Search (Peer to Peer Search System)

Exit | Setup | Search | Tools | About

6search

web mining

Search

Local Search Only ☐ Yahoo Search ☑ Google Search ☑ Peer network connected

6S search | Google search | Yahoo search

Data Mining - Home Page (Misc)
... Data **Mining** Software (index) Data **Mining** Events (index) Data **Mining** General/Misc (index) People working ... OnLine Analytical Processing (OLAP) , Data ...
http://www.the-data-mine.com/ Similar pages (Power by GiveALink)
Contributors: wls_iceman#1,wls_angel#3,wls_beast#7

KDnuggets: Data Mining, Web Mining, and Knowledge Discovery Guide
... Data **Mining**, **Web Mining**, Text **Mining**, and Knowledge Discovery ... KDnuggets: Data **Mining**, **Web Mining**, and Knowledge Discovery ...
http://www.kdnuggets.com/ Similar pages (Power by GiveALink)
Contributors: wls_iceman#1,wls_angel#3

Data Mining Software in the Yahoo! Directory
... Data **Mining** Software in the ...
http://dir.yahoo.com/Business_and_Economy/Business_to_Business/Computers/Software/Databases/Data_Mining/ Similar pages (Power by GiveALink)
Contributors: wls_iceman#1,wls_beast#7

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/ Similar pages (Power by GiveALink)
Contributors: wls_iceman#1,wls_beast#7

Done

http://homer.informatics.indiana.edu/~nan/6S/

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 Windo

Google
Yahoo
6Search (local)
6Search (peers)
Amazon.com
Answers.com
Creative Commons
eBay
Manage Search Engines...

KDnuggets: Data Mining, Web Mining, and Knowledge Discovery Guide
Newsletter on the data mining and knowledge industries, offering information on data mining, know
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
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/
Contributors: Unknown#7,wls_iceman#1
Similar pages (Power by GiveALink)

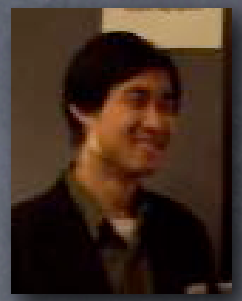
UMBC Agent Web -- news and information on software agent technology
... UMBC Agent **Web** -- news and information ...
http://agents.umbc.edu/Topics/Related_Topics/Information_retrieval_and_knowledge_management/
Contributors: Unknown#7,wls_iceman#1
Similar pages (Power by GiveALink)

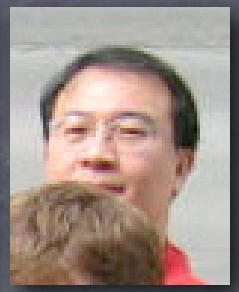
BUBL LINK: Information retrieval
... and Forward Knowledge Approach **Data** Mine: **Data Mining** and Knowledge Discovery ... the Gaps KD Mine: **Data** ...
http://bubl.ac.uk/link/i/informationretrieval.htm
Contributors: Unknown#7,wls_iceman#1
Similar pages (Power by GiveALink)

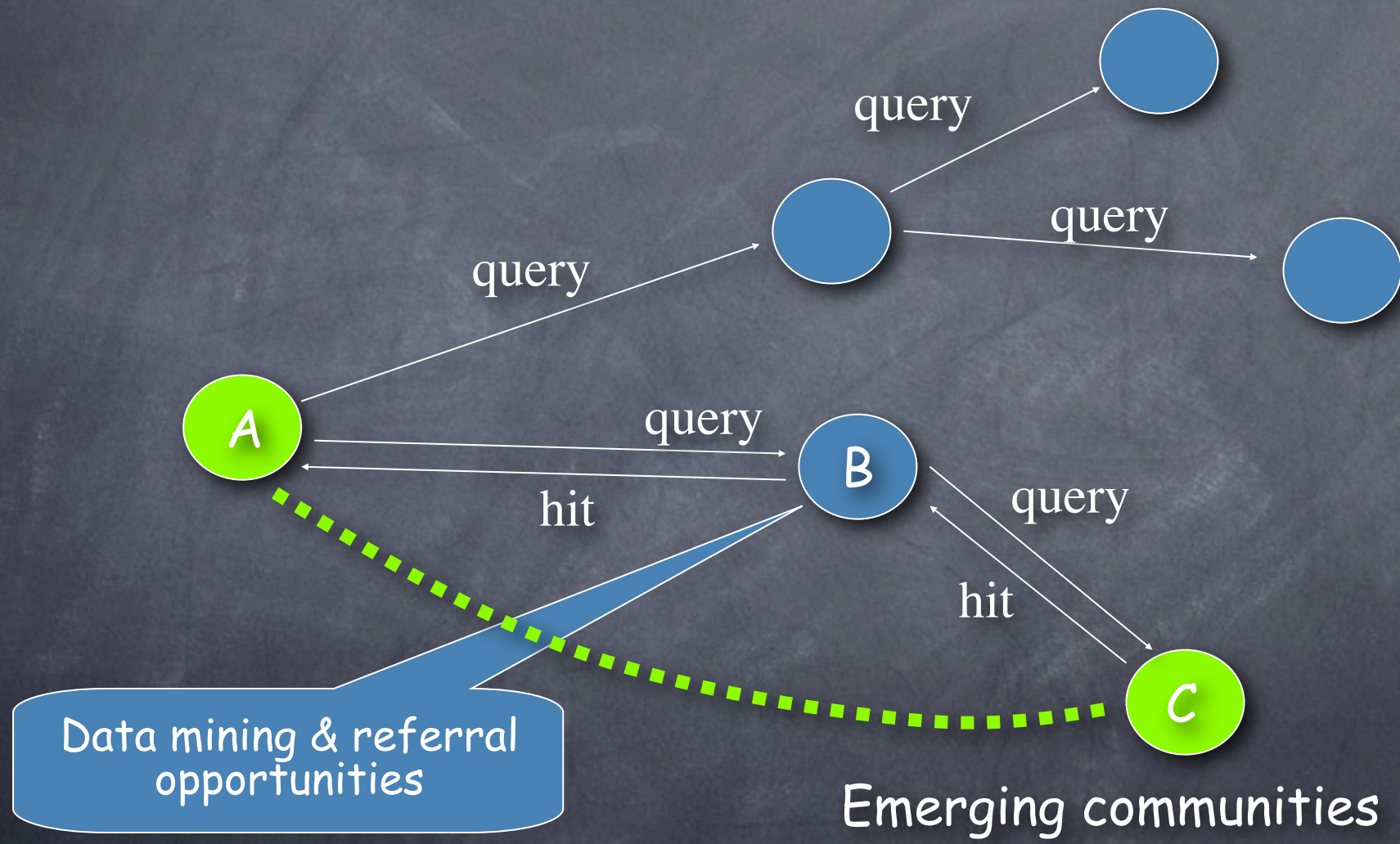
Library and Information Science > Information Retrieval in the Yahoo! Directory

Done

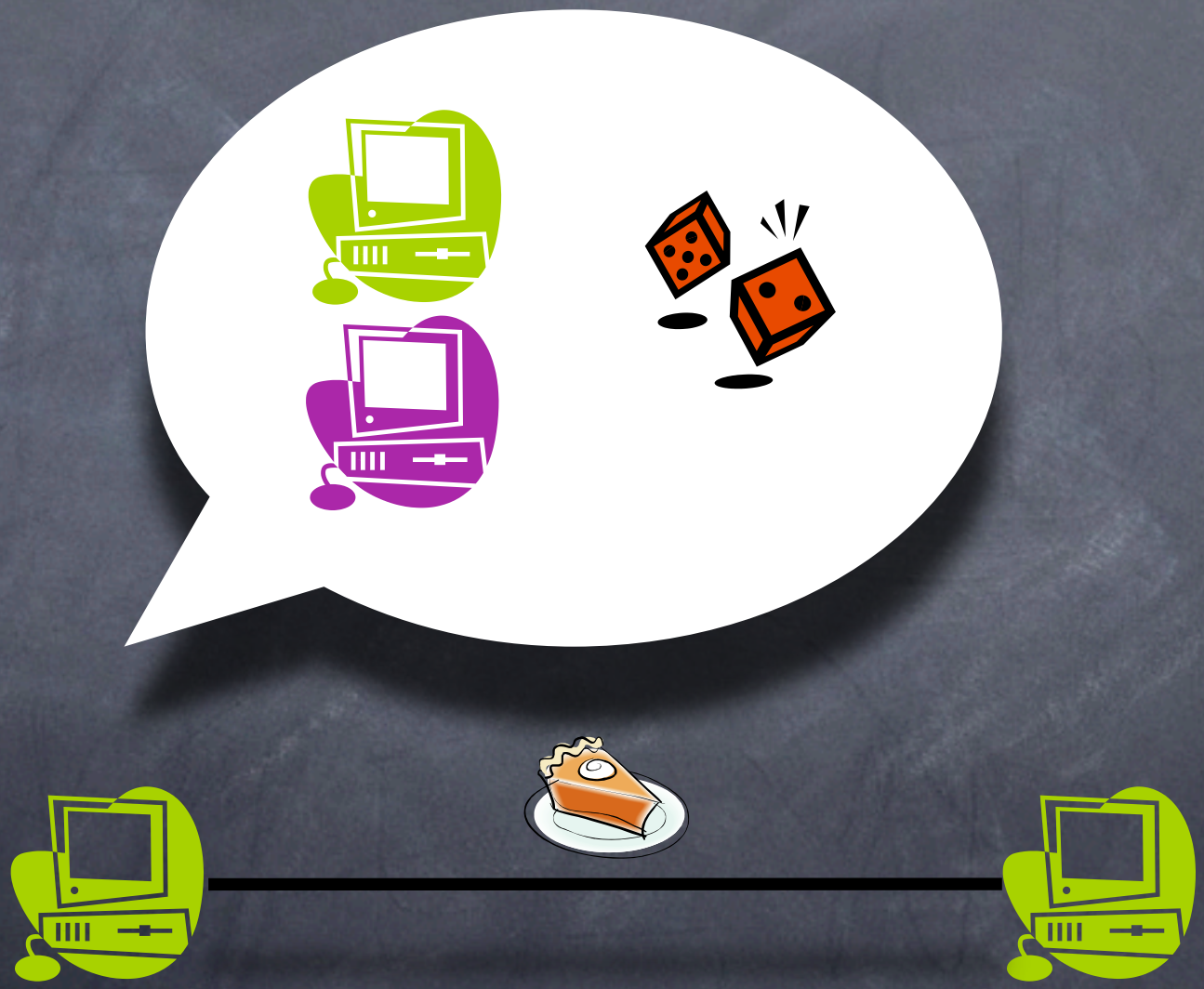
6S: peer distributed crawling and collaborative searching



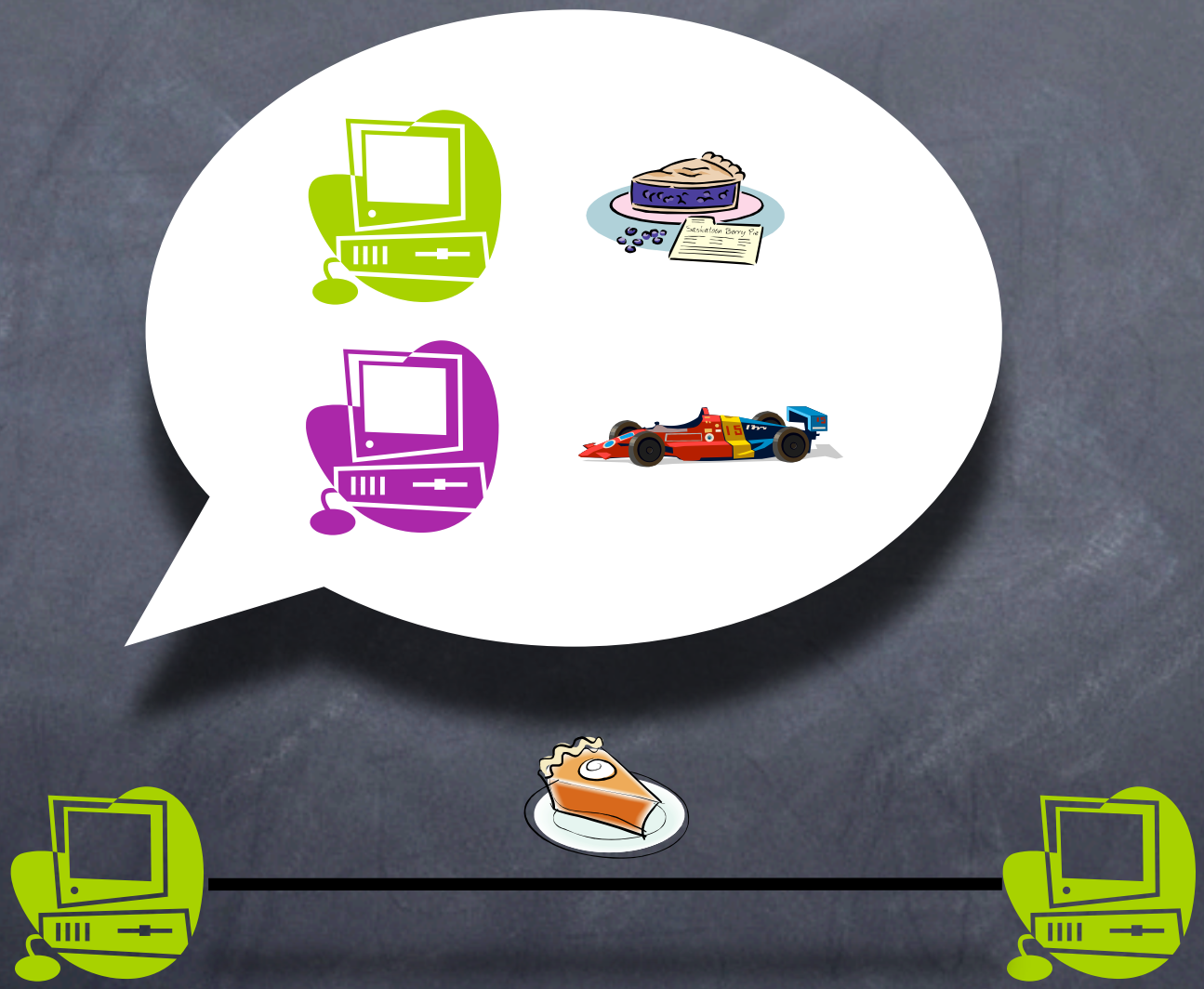




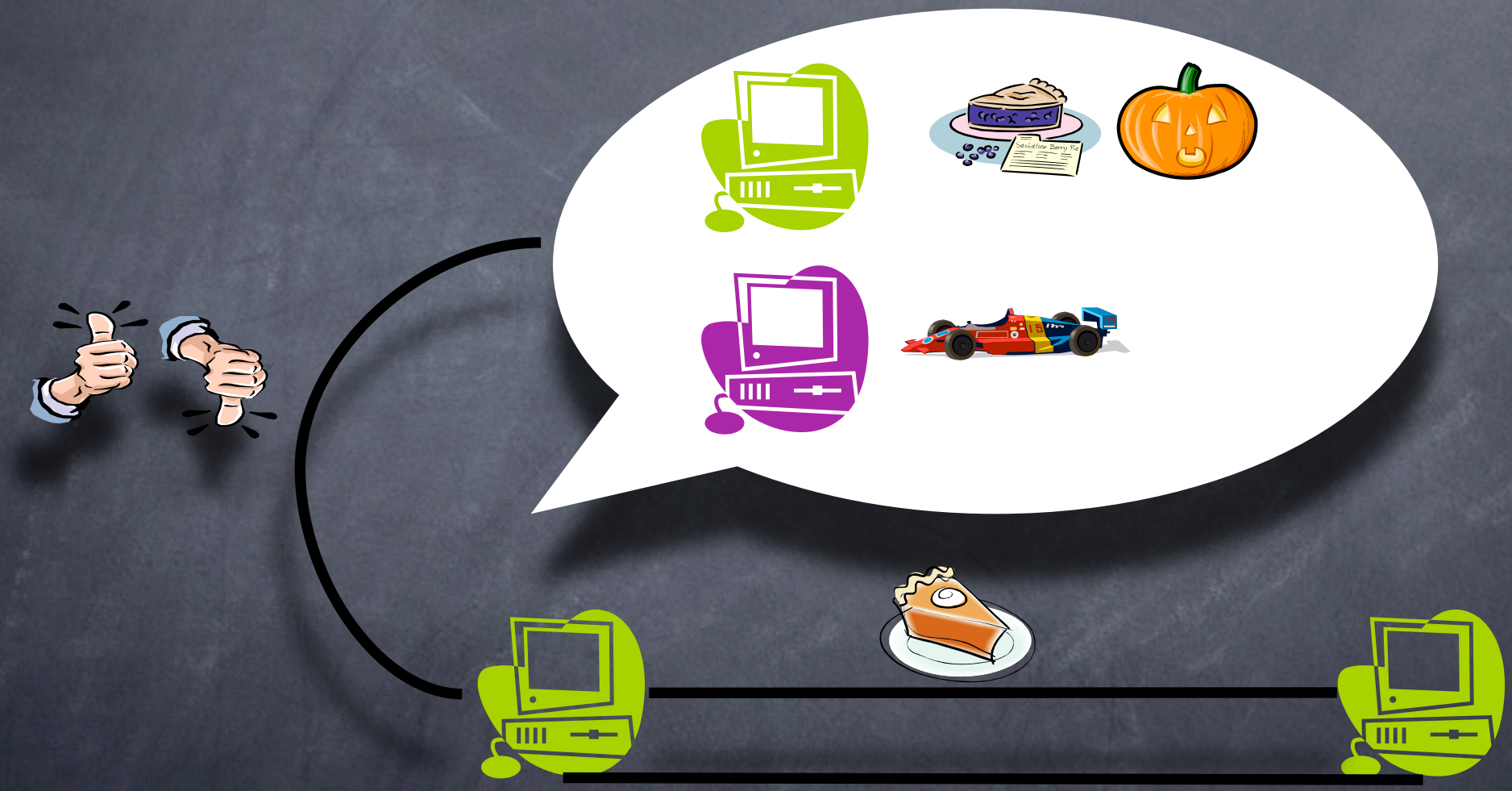
Algorithm 1: Random Known



Algorithm 2: Greedy



Algorithm 3: Reinforcement



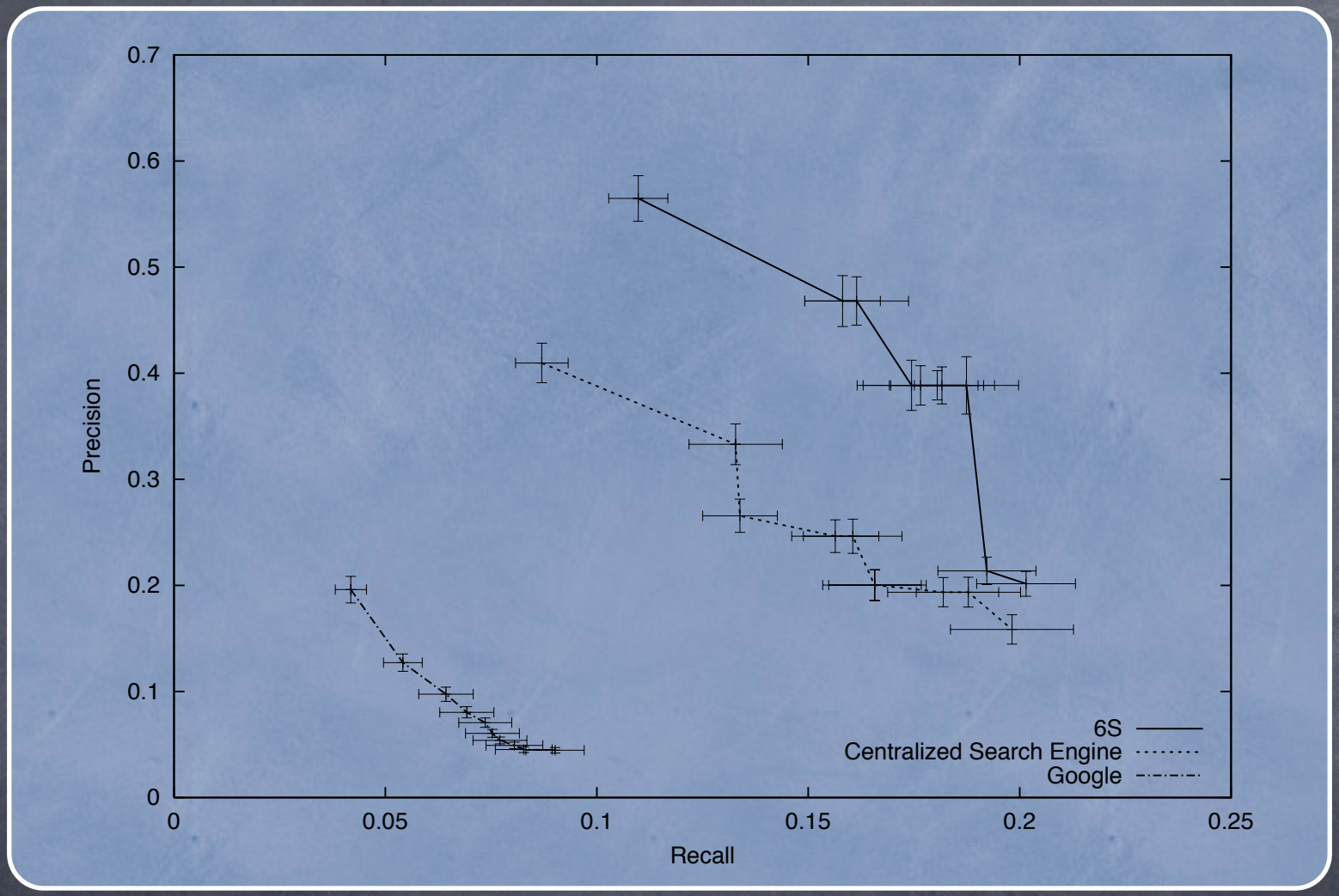
Learning Peers

- A focused profile W^f of other peers for terms in queries
- An expanded profile W^e of other peers for terms not in queries, but that co-occur with query terms and with higher frequency

$$w_{i,p}(t+1) = (1-\gamma) \cdot w_{i,p}(t) + \gamma \cdot \left(\frac{S_p + 1}{S_l + 1} - 1 \right)$$

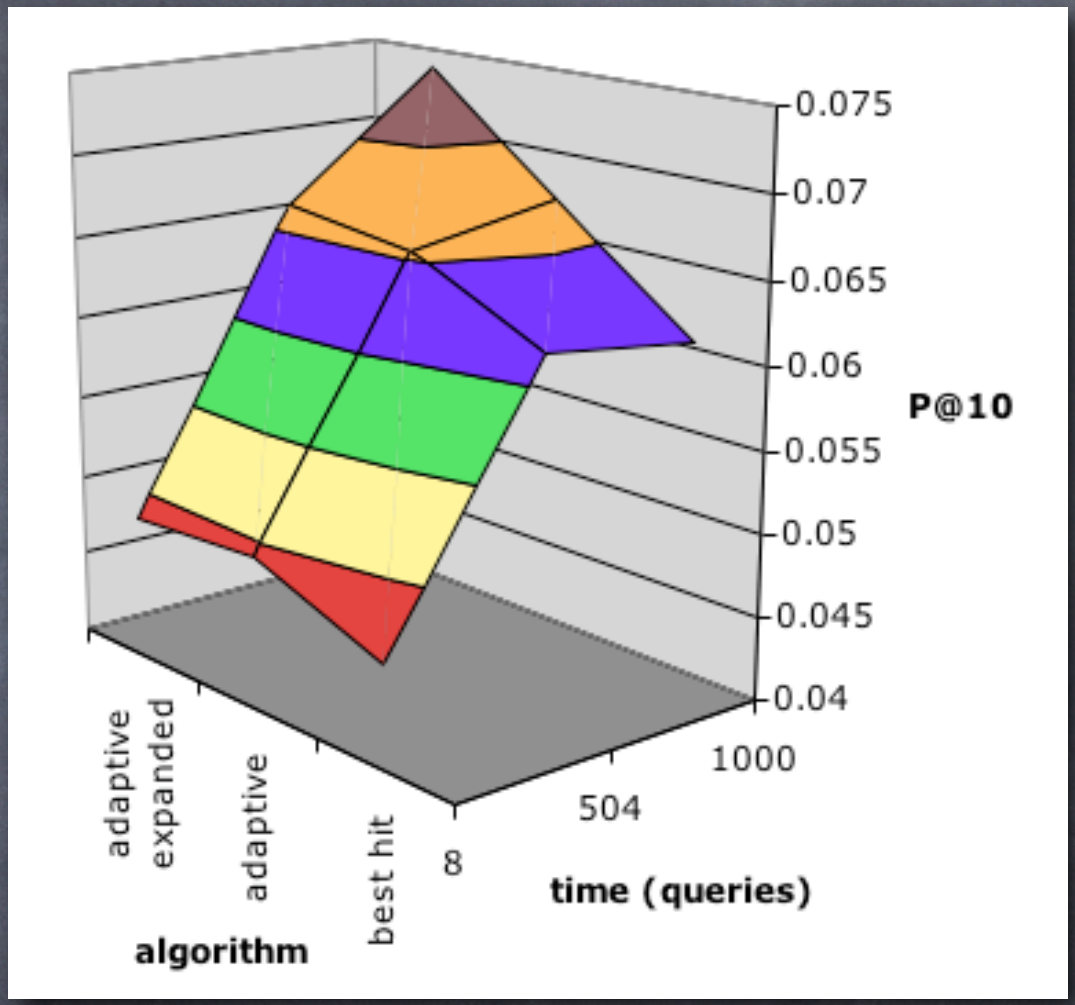
- Query routing:

$$\sigma(p, Q) = \sum_{i \in Q} \left[\alpha \cdot w^f_{i,p} + (1-\alpha) \cdot w^e_{i,p} \right]$$



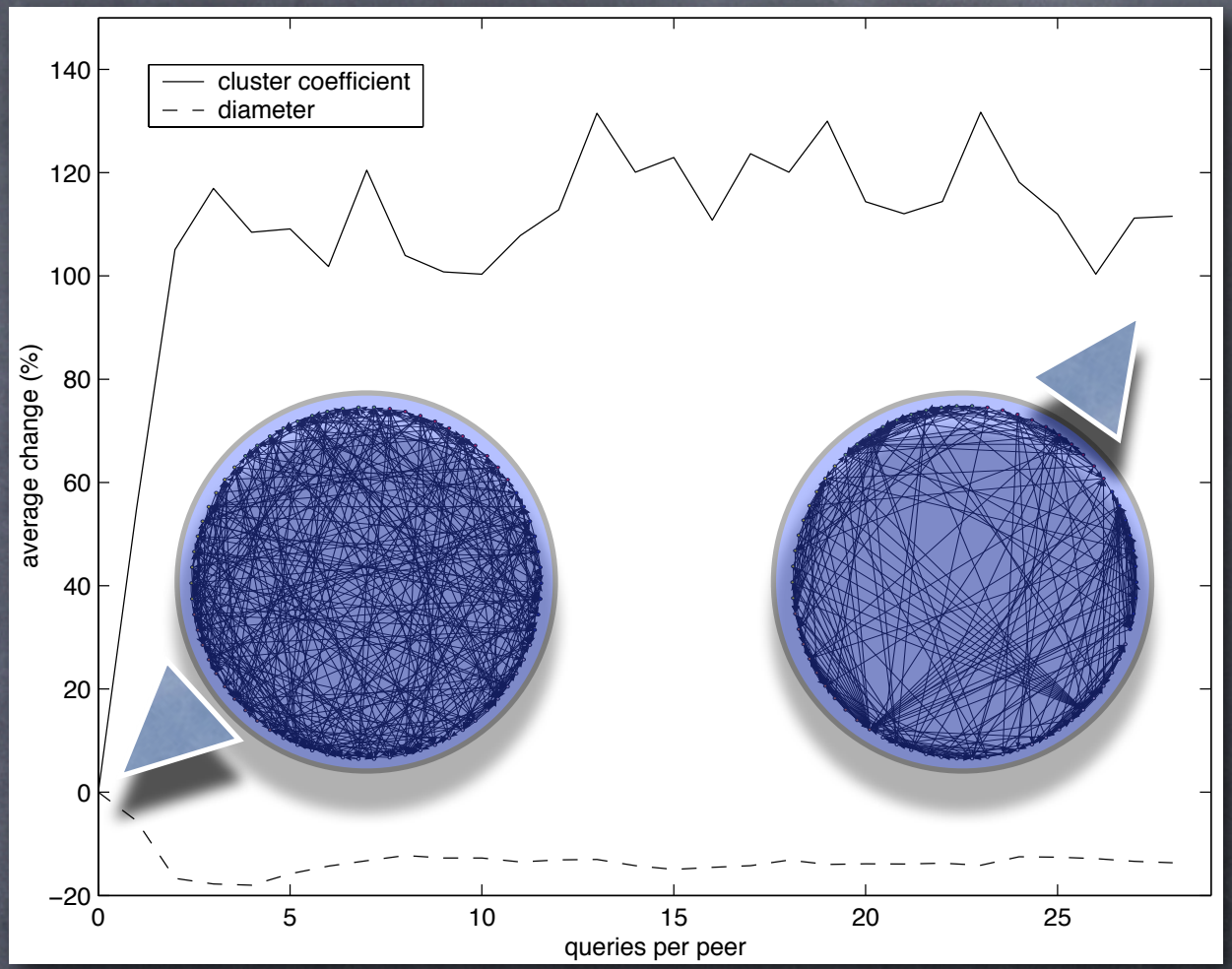
Distributed vs Centralized

WTAS 2005


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

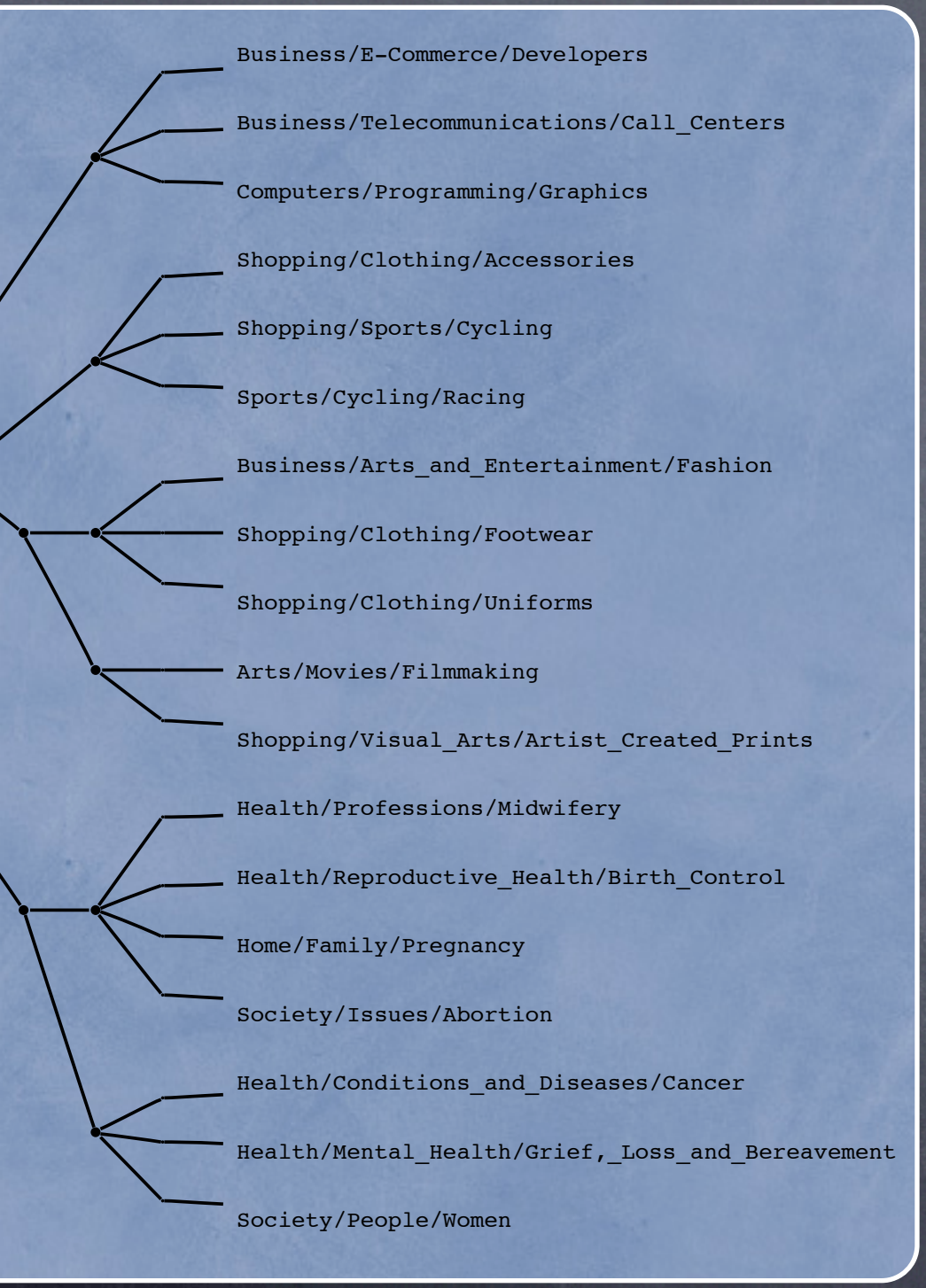
WWW 2005

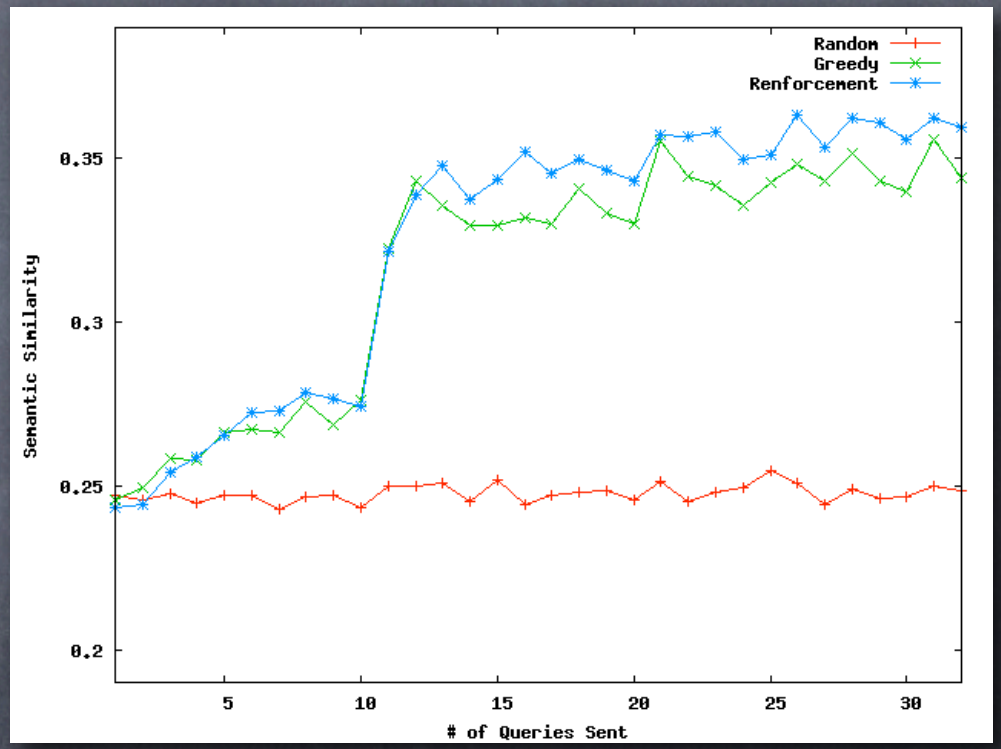


Small-world

WWW 2004

Semantic similarity





P2PIR 2006

Discussion:

Would you use 6S?
Why or why not?

Thank you!
Questions?

informatics.indiana.edu/fil

homer.informatics.indiana.edu/~nan

Research supported by NSF
CAREER Award IIS-0348940

Indiana University Department of ComputerScience

Indiana University School of informatics

Data and Search Institute

Next generation tools and ideas for searching and manipulating large volumes of data

www.dataandsearch.org