Web Structure & Content

(not a theory talk)

Filippo Menczer

The University of Iowa

http://dollar.biz.uiowa.edu/~fil/

Research supported by NSF
CAREER Award IIS-0133124

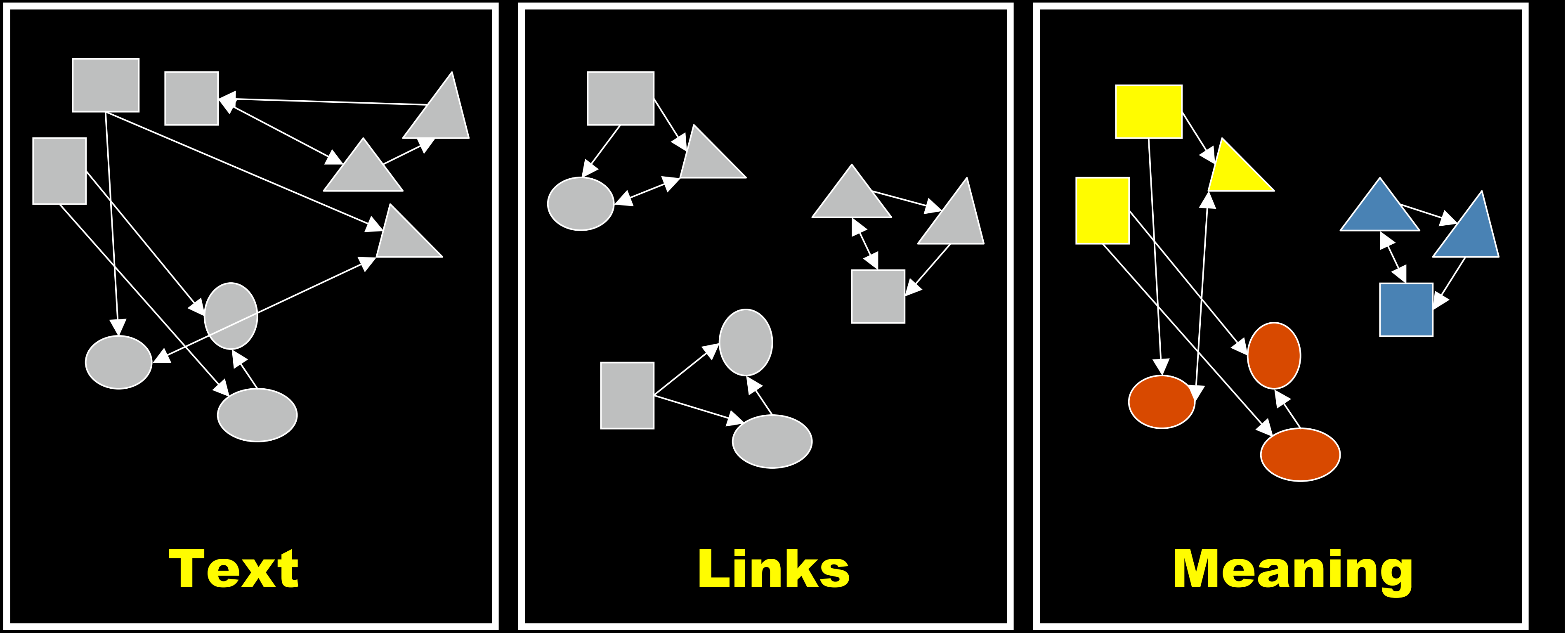
Exploiting the Web’s text and link cues

- *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

- *Pages that link to each other tend to be related*
  - Link-cluster conjecture (Menczer 1997)
    - Many formulations: Gibson & al 1998, Bharat & Henzinger 1998, Chakrabarti & al 1998, Dean & Henzinger 1999, Davison 2000, etc
  - Link eigenvalue analysis: HITS, hubs and authorities
    - (Kleinberg & al 1998 segg. @ Almaden etc.)
  - Google’s PageRank analysis
    - (Brin & Page 1998)
  - The whole second generation of search engines

Three topologies

What about the *relationship* between lexical / link cues and page meaning?



Text

Links

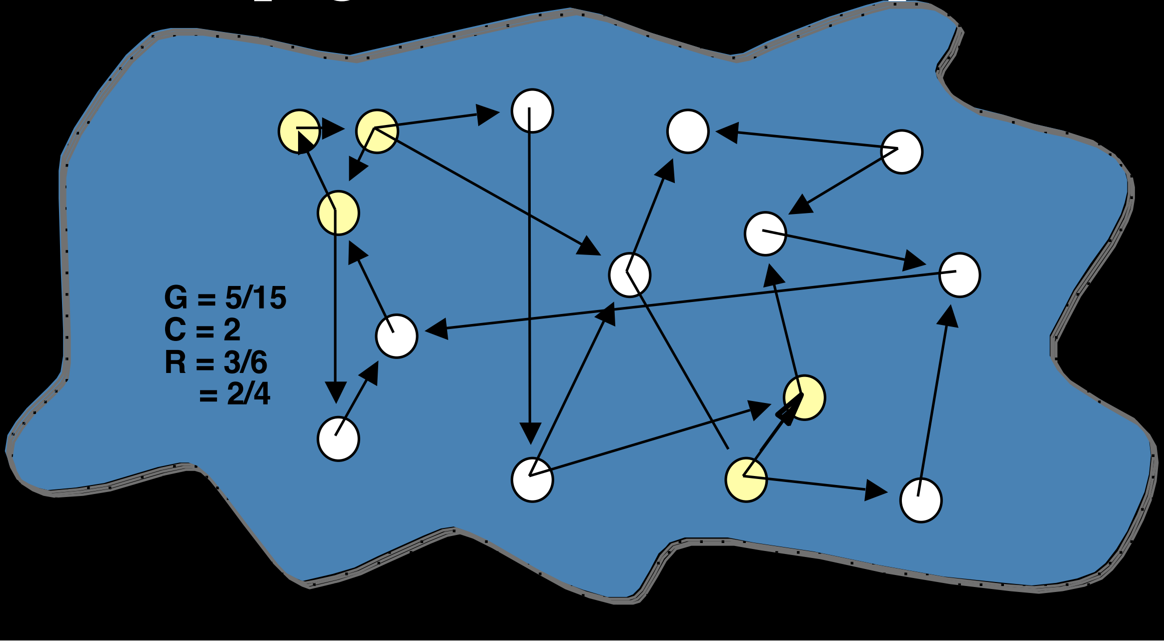
Meaning

Talk outline

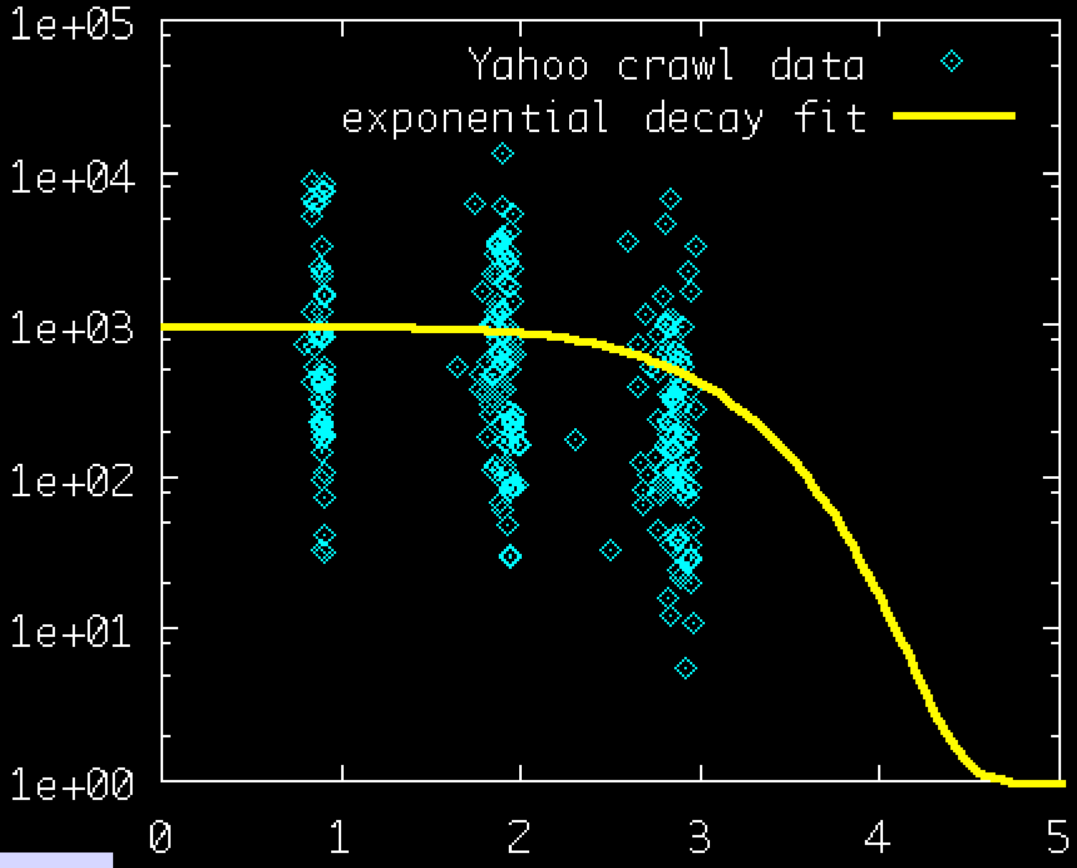
- The topologies of the Web
- Correlations, distributions, projections
- Power laws and Web growth models
- Navigating optimal paths
- Semantic maps (?)

The “link-cluster” conjecture

- Connection between semantic topology (relevance) and linkage 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***
  - Necessary and sufficient condition for a random crawler to find pages related to start points



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



Link-cluster conjecture

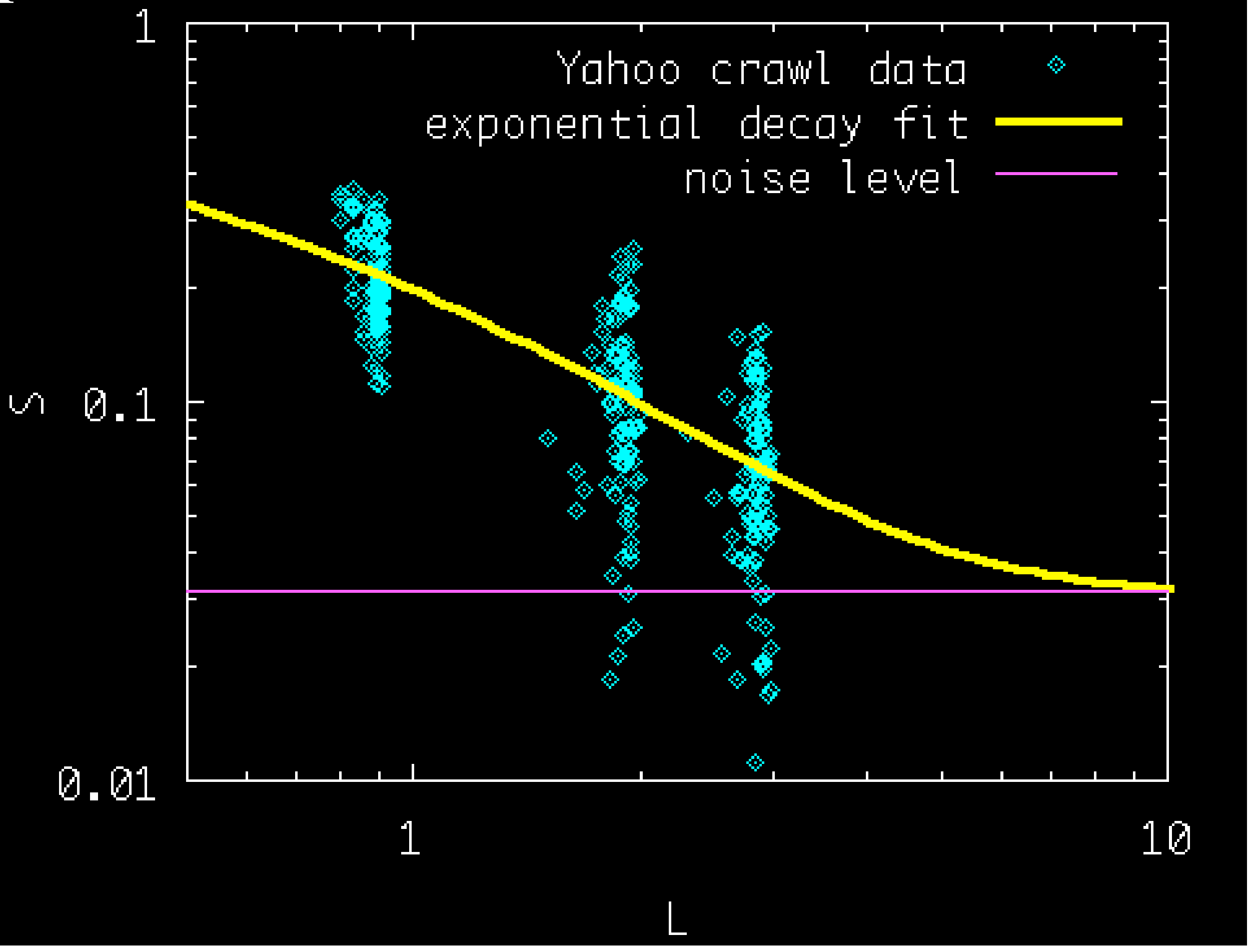
- Preservation of semantics (meaning) across links

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

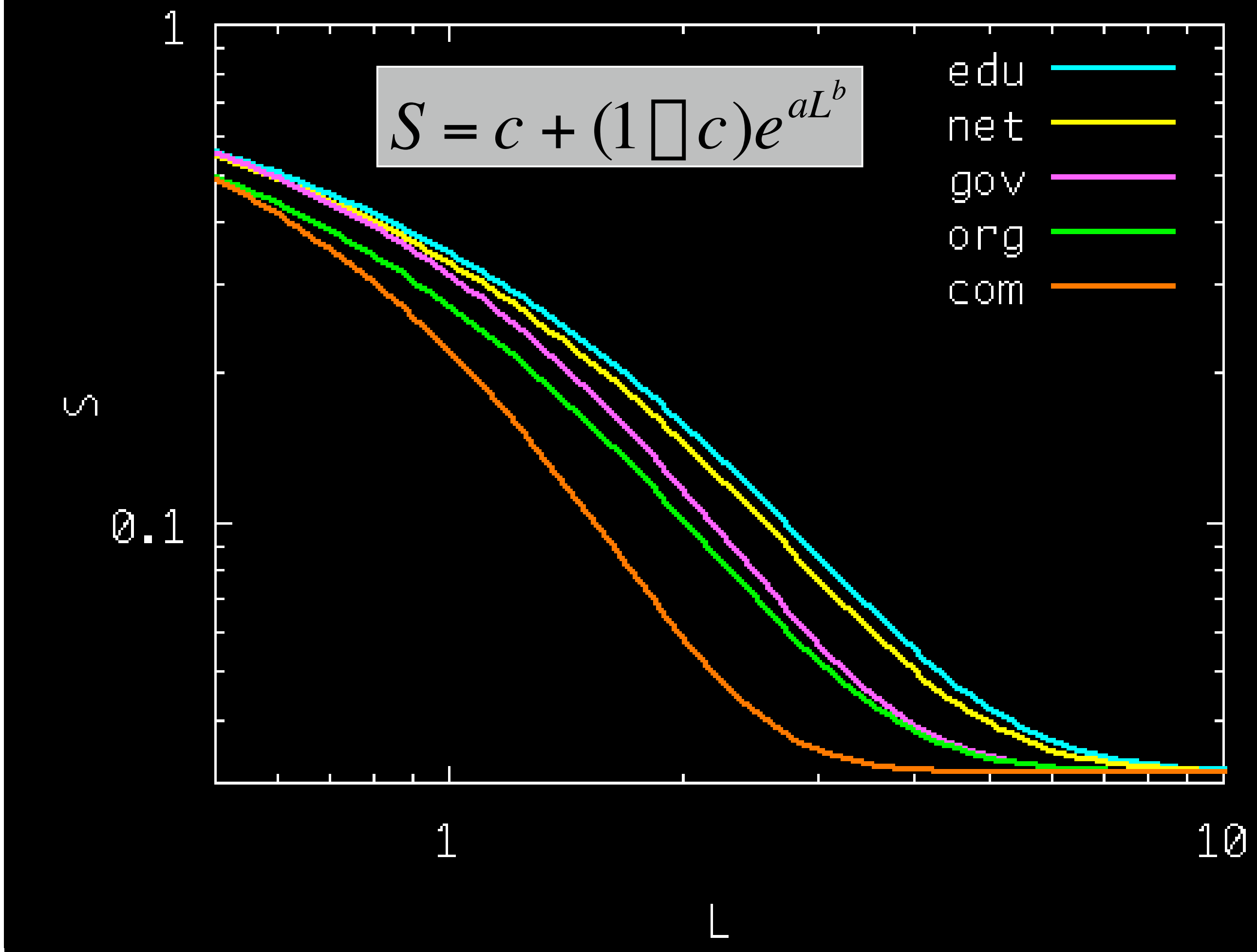
The “link-content” conjecture

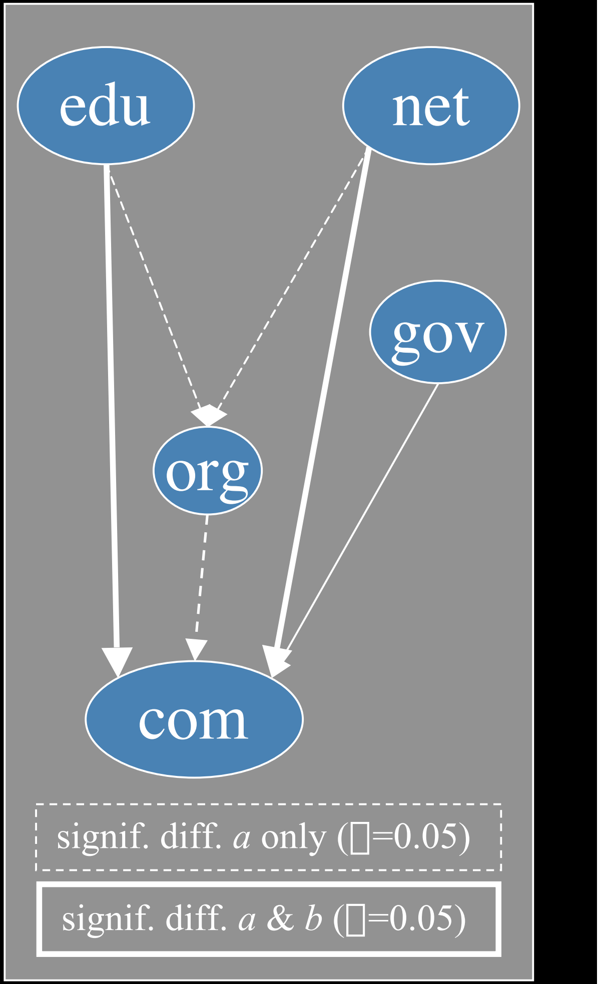
$$S(q,\delta) \equiv \frac{\sum_{\{p:\|path(q,p)\|\le\delta\}} sim(q,p)}{\left|\{p:\|path(q,p)\|\le\delta\}\right|}$$

- 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 lexical-linkage correlation


$$S = c + (1-c)e^{aL^b}$$



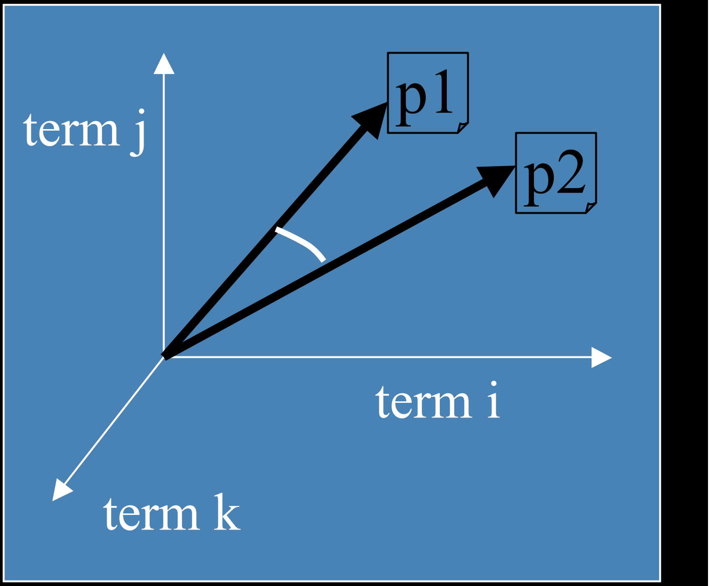
signif. diff. *a* only (□=0.05)

signif. diff. *a* & *b* (□=0.05)

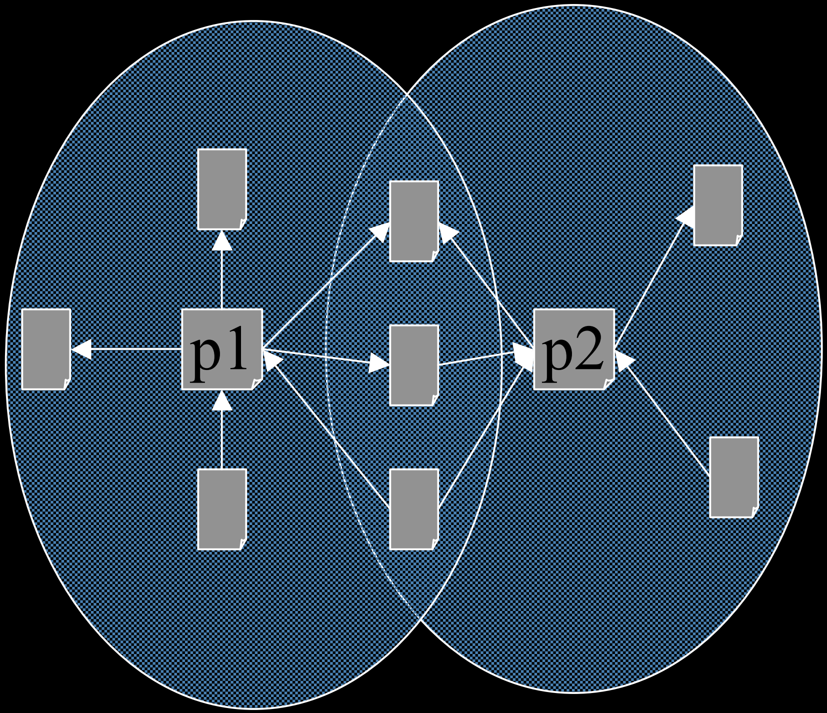
Mapping the relationship between topologies

- Any pair of pages rather than linked pages from crawl
- Data: Open Directory Project (dmoz.org)
  - RDF Snapshot: 2002-02-14 04:01:50 GMT
  - After cleanup: 896,233 URLs in 97,614 topics
  - After sampling: 150,000 URLs in 47,174 topics
    - 10,000 from each of 15 top-level branches
- Need ‘similarity’ or ‘proximity’ metric for each topology, given a pair of pages:
  - Content: textual/lexical (cosine) similarity
  - Link: co-citation/bibliographic coupling
  - Semantic: relatedness inferred from manual classification

$$\sigma_c(p_1,p_2) = \frac{\sum_{k \in p_1 \cap p_2} f_{kp_1} f_{kp_2}}{\sqrt{\sum_{k \in p_1} f_{kp_1}^2 \sum_{k \in p_2} f_{kp_2}^2}}$$



Content similarity



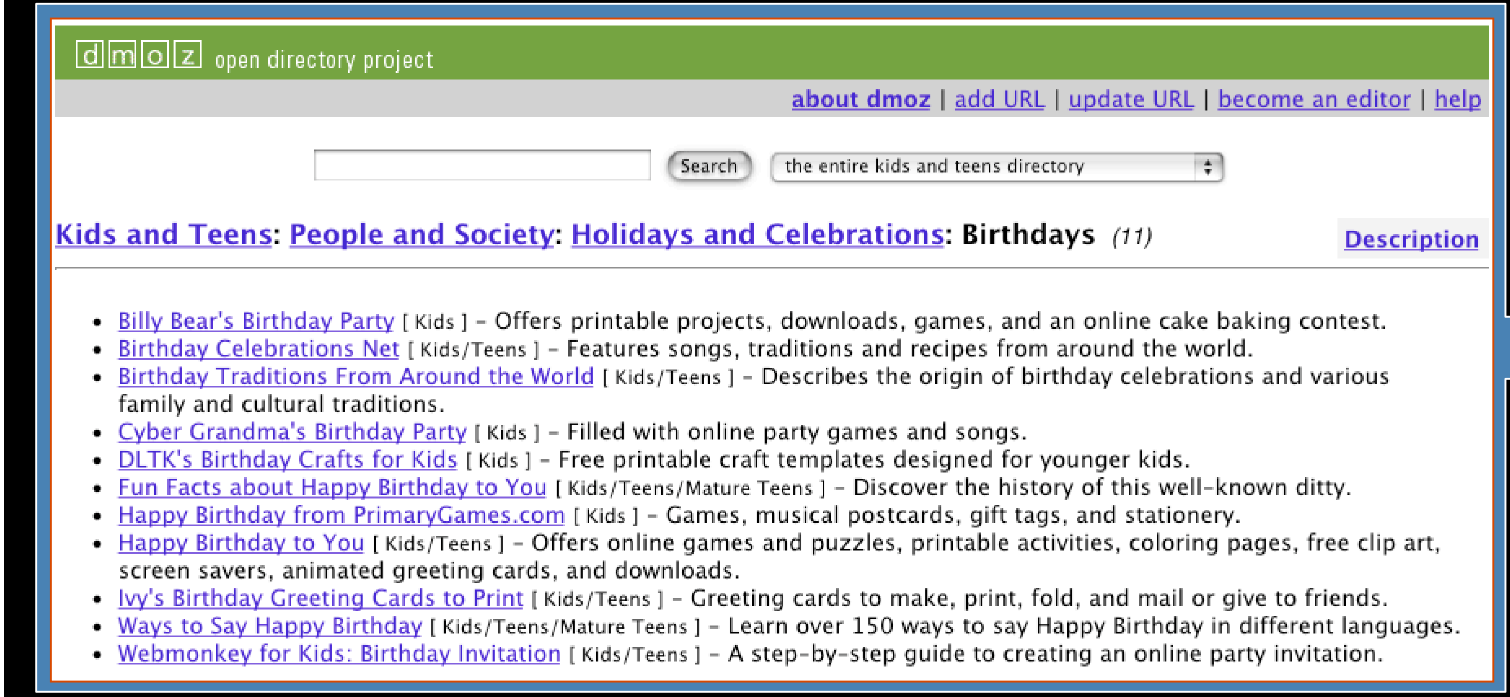
Link similarity

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

Semantic similarity

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

- Information-theoretic measure based on classification tree (Lin 1998)
- Classic path distance in special case of balanced tree



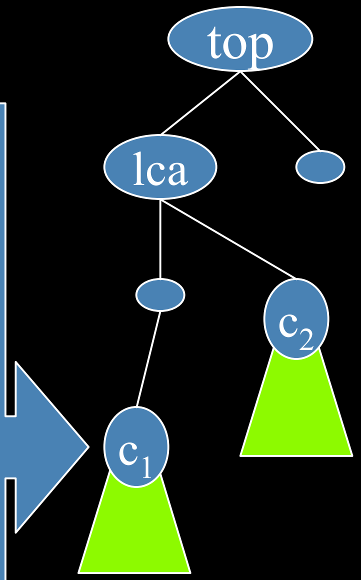
dmoz open directory project

about dmoz | add URL | update URL | become an editor | help

Search | the entire kids and teens directory

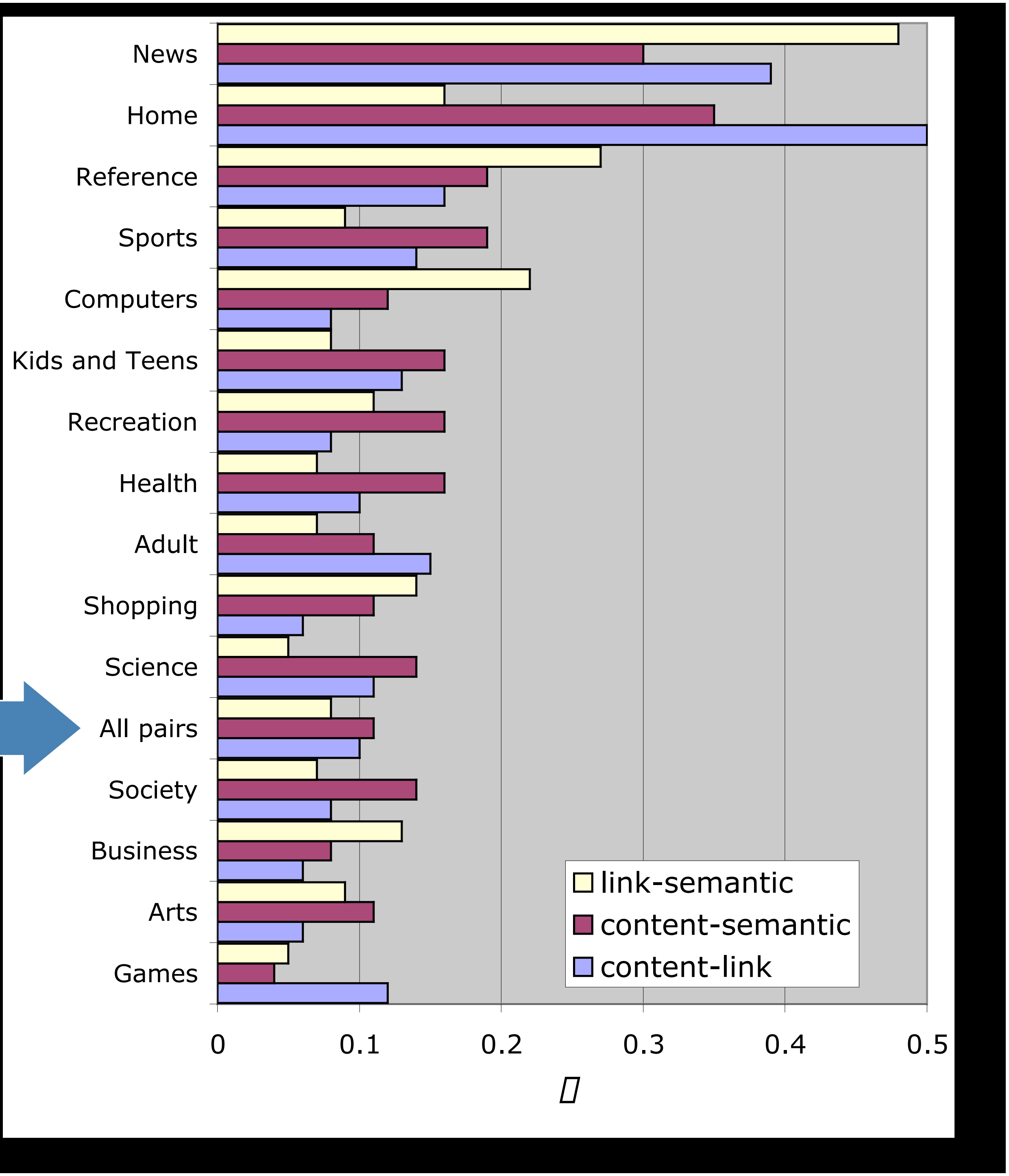
Kids and Teens: People and Society: Holidays and Celebrations: Birthdays *(11)* | Description

- Billy Bear's Birthday Party [Kids] - Offers printable projects, downloads, games, and an online cake baking contest.
- Birthday Celebrations Net [Kids/Teens] - Features songs, traditions and recipes from around the world.
- Birthday Traditions From Around the World [Kids/Teens] - Describes the origin of birthday celebrations and various family and cultural traditions.
- Cyber Grandma's Birthday Party [Kids] - Filled with online party games and songs.
- DLTK's Birthday Crafts for Kids [Kids] - Free printable craft templates designed for younger kids.
- Fun Facts about Happy Birthday to You [Kids/Teens/Mature Teens] - Discover the history of this well-known ditty.
- Happy Birthday from PrimaryGames.com [Kids] - Games, musical postcards, gift tags, and stationery.
- Happy Birthday to You [Kids/Teens] - Offers online games and puzzles, printable activities, coloring pages, free clip art, screen savers, animated greeting cards, and downloads.
- Ivy's Birthday Greeting Cards to Print [Kids/Teens] - Greeting cards to make, print, fold, and mail or give to friends.
- Ways to Say Happy Birthday [Kids/Teens/Mature Teens] - Learn over 150 ways to say Happy Birthday in different languages.
- Webmonkey for Kids: Birthday Invitation [Kids/Teens] - A step-by-step guide to creating an online party invitation.

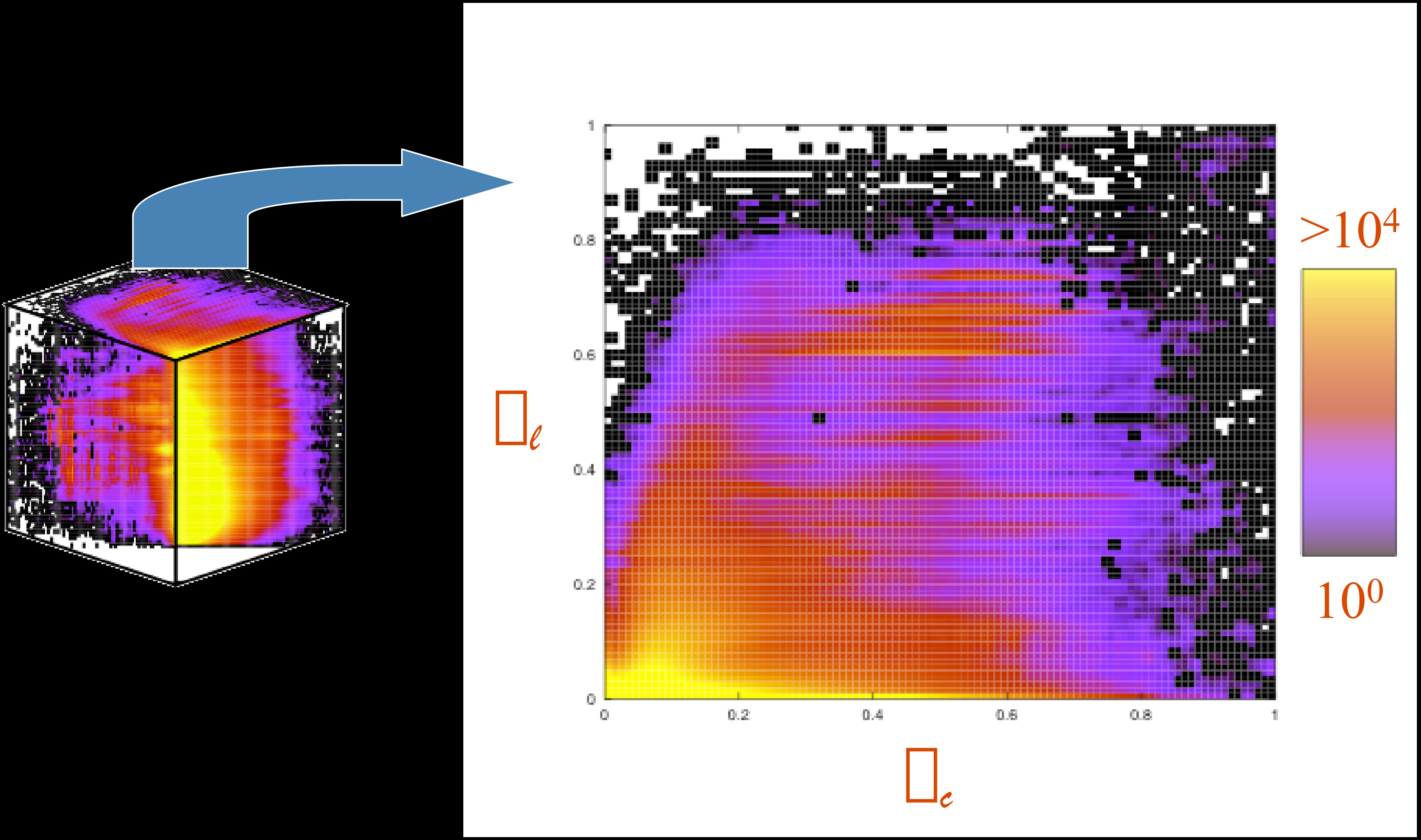


Correlations between similarities

3.84×10^9 pairs



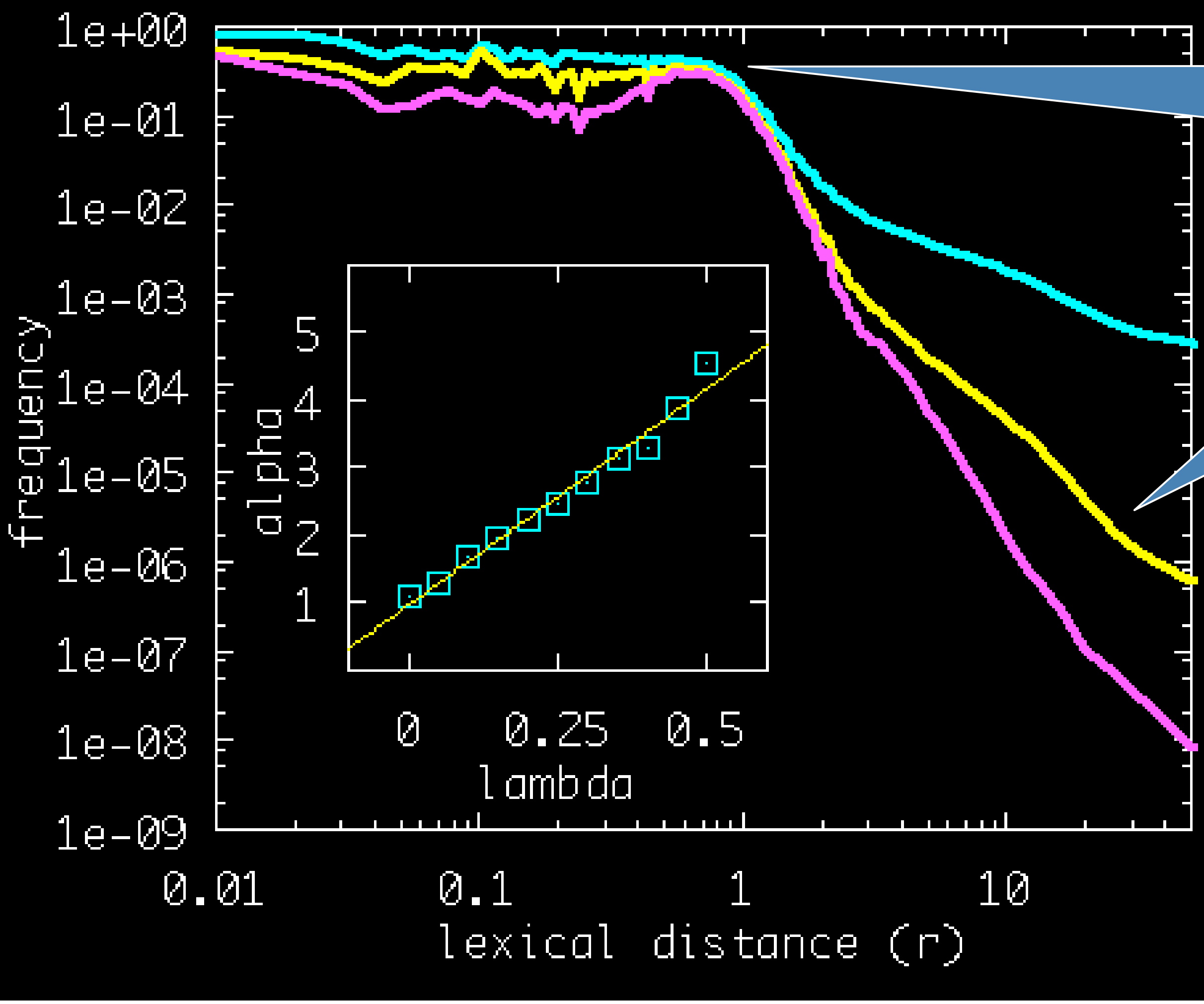
Joint distribution cube



Link probability vs lexical distance

$$r = \frac{1}{\square_c} \square 1$$

$$\Pr(\square \mid \square) = \frac{\left|(p,q) : r = \square \square \square_l > \square\right|}{\left|(p,q) : r = \square\right|}$$



Phase transition

$\square^*$

Power law tail

$\Pr(\square \mid \square) \sim \square^{\square\square(\square)}$

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

Web growth models

- Preferential attachment “BA”
  - At each step t add page p_t
  - Create m new links from p_t to $p_{i<t}$

  (Barabasi & Albert 1999, de Solla Price 1976)

$$\mathrm{Pr}(i) \quad k(i)$$

- Modified BA

  (Bianconi & Barabasi 2001, Adamic & Huberman 2000)

$$\mathrm{Pr}(i) \quad \square(i)k(i)$$

- Mixture

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

$$\mathrm{Pr}(i) \quad \square \cdot k(i) + (1 \square \square) \cdot c$$

- Web copying

  (Kleinberg, Kumar & al 1999, 2000)

$$\mathrm{Pr}(i) \quad \square \cdot \mathrm{Pr}(j \; \square \quad i) + (1 \square \square) \cdot c$$

- Mixture with Euclidean distance in graphs

  (Fabrikant, Koutsoupias & Papadimitriou 2002)

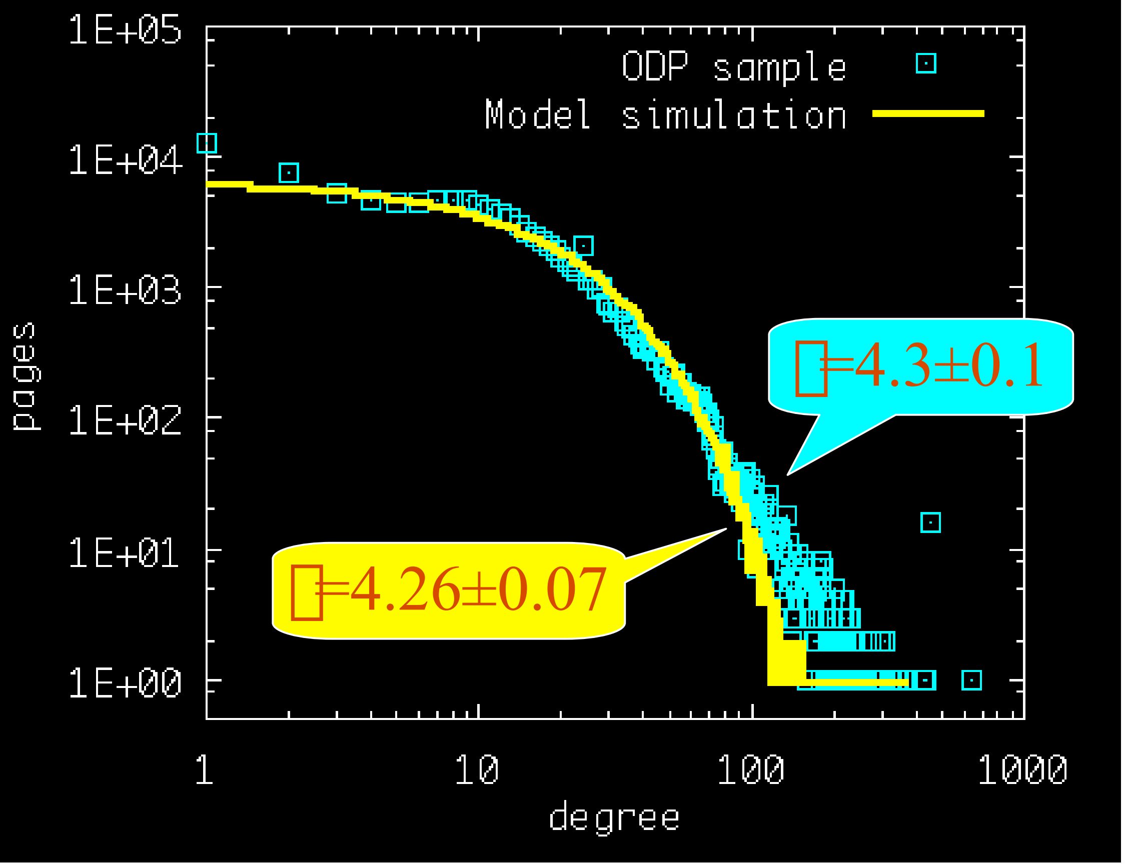
$$i = \arg\min(\square r_{it} + g_i)$$

Local content-based growth model

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

- Similar to preferential attachment (BA)
- At each step t add page p_t
- Create m new links from p_t to existing pages
- Use degree (k) info only for nearby pages

(popularity/importance of similar/related pages)



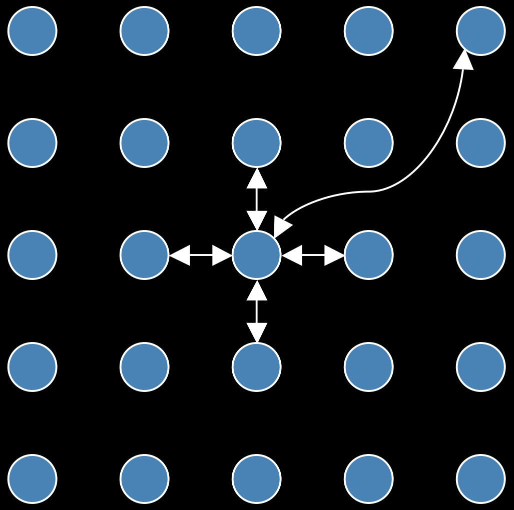
α=4.3±0.1

α=4.26±0.07

Efficient crawling algorithms?

- Theory: since the Web is a small world network, or has a scale free degree distribution, there exist short paths 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 *et al* 2001)

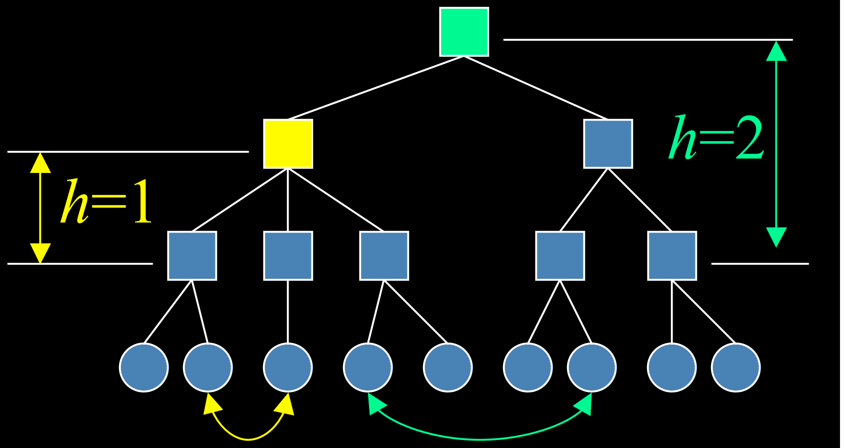
Exception # 1



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

$$t \sim \log^2 N \;\square\;\; \square = D$$

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)

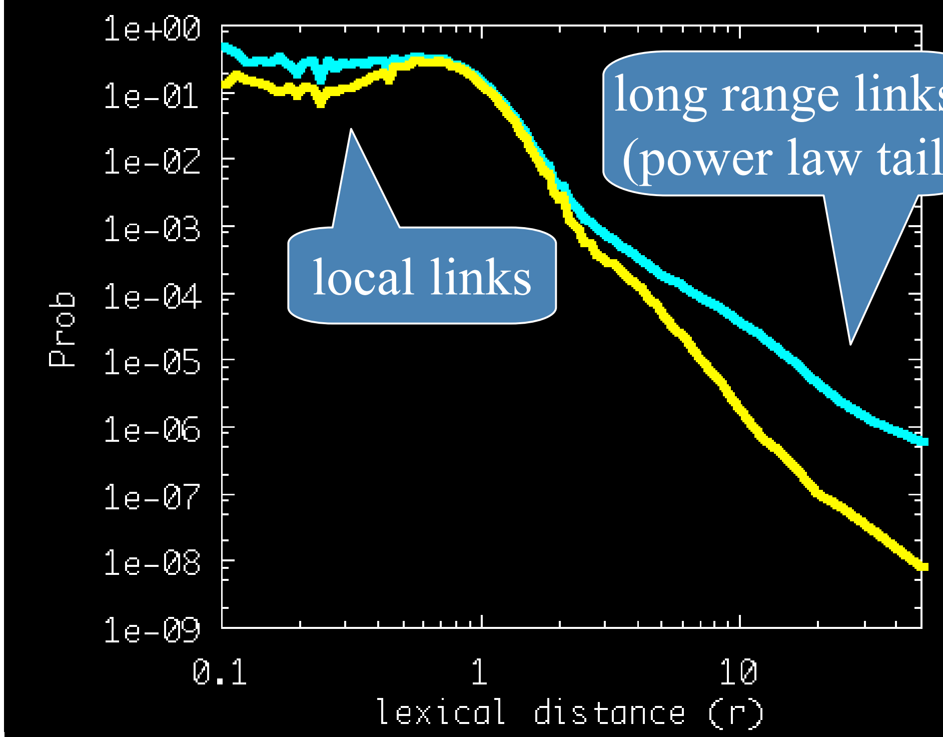
$$t \sim \log^{\square} N, \square \geq 1$$

Is the Web one of these exceptions?

Geographical model

- Replace lattice distance by lexical distance

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



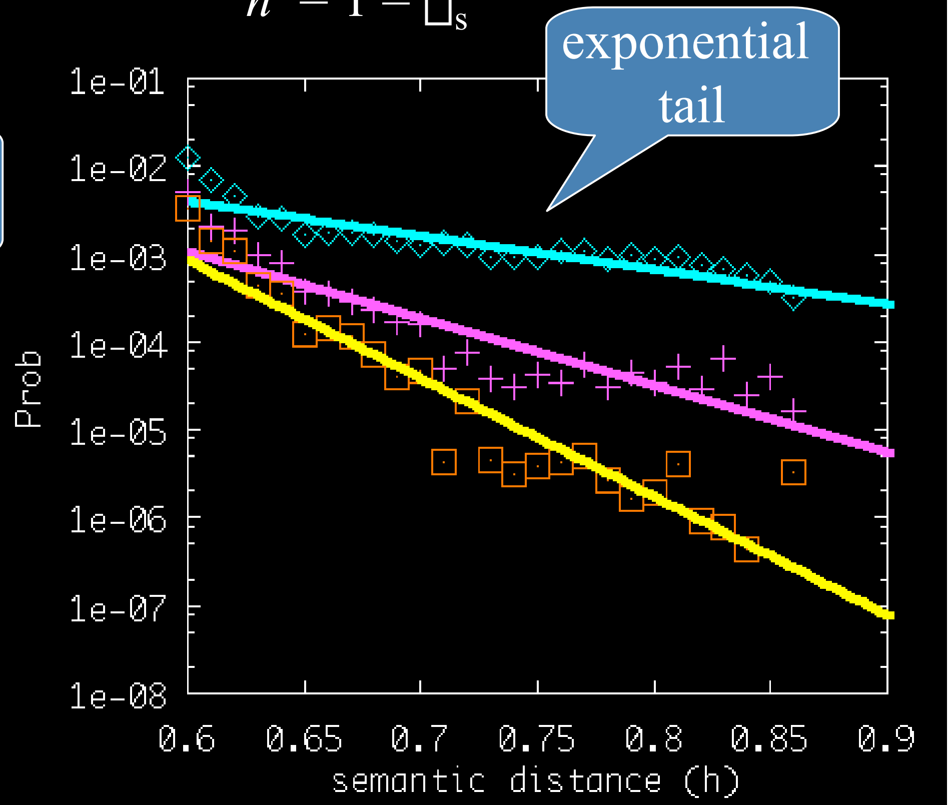
local links

long range links (power law tail)

Hierarchical model

- Replace tree distance by semantic distance

$h = 1 - \square_s$

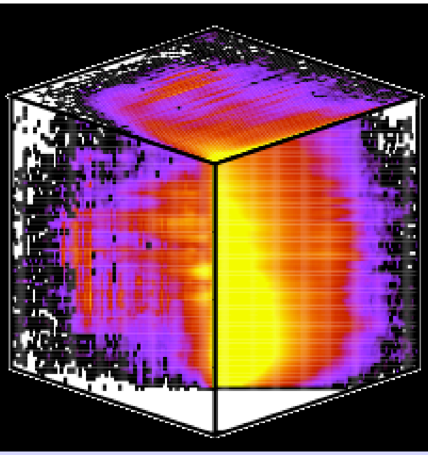


exponential tail

Talk outline

- The topologies of the Web
- Correlations, distributions, projections
- Power laws and Web growth models
- Navigating optimal paths
- Semantic maps

Semantic maps: define “local” Precision and Recall



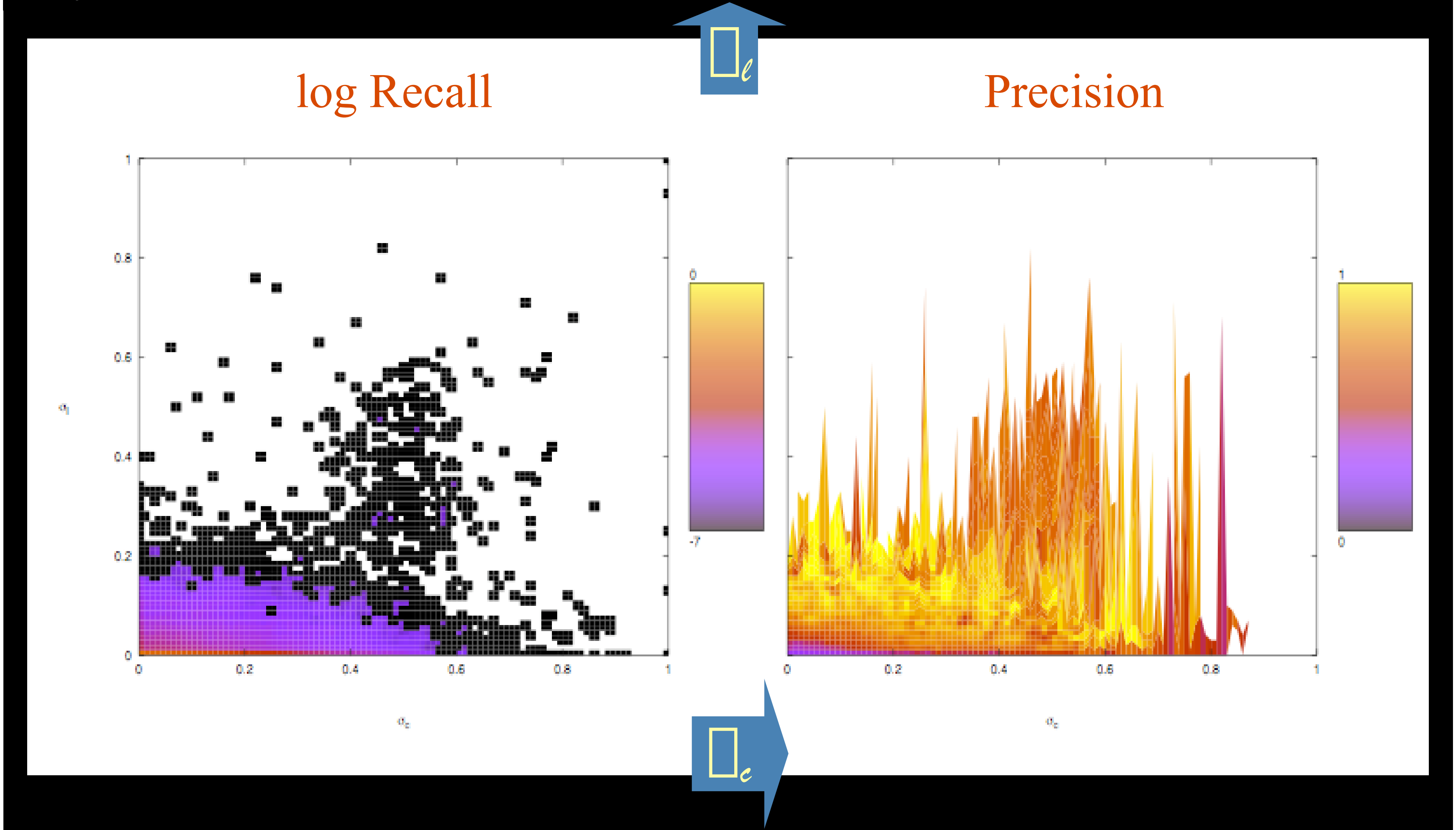
$$P(s_c,s_l) = \frac{\displaystyle\sum_{\{p,q:\square_c = s_c,\square_l = s_l\}} \square_s(p,q)}{\left|\{p,q:\square_c = s_c,\square_l = s_l\}\right|}$$

Averaging semantic similarity

$$R(s_c,s_l) = \frac{\displaystyle\sum_{\{p,q:\square_c = s_c,\square_l = s_l\}} \square_s(p,q)}{\displaystyle\sum_{\{p,q\}} \square_s(p,q)}$$

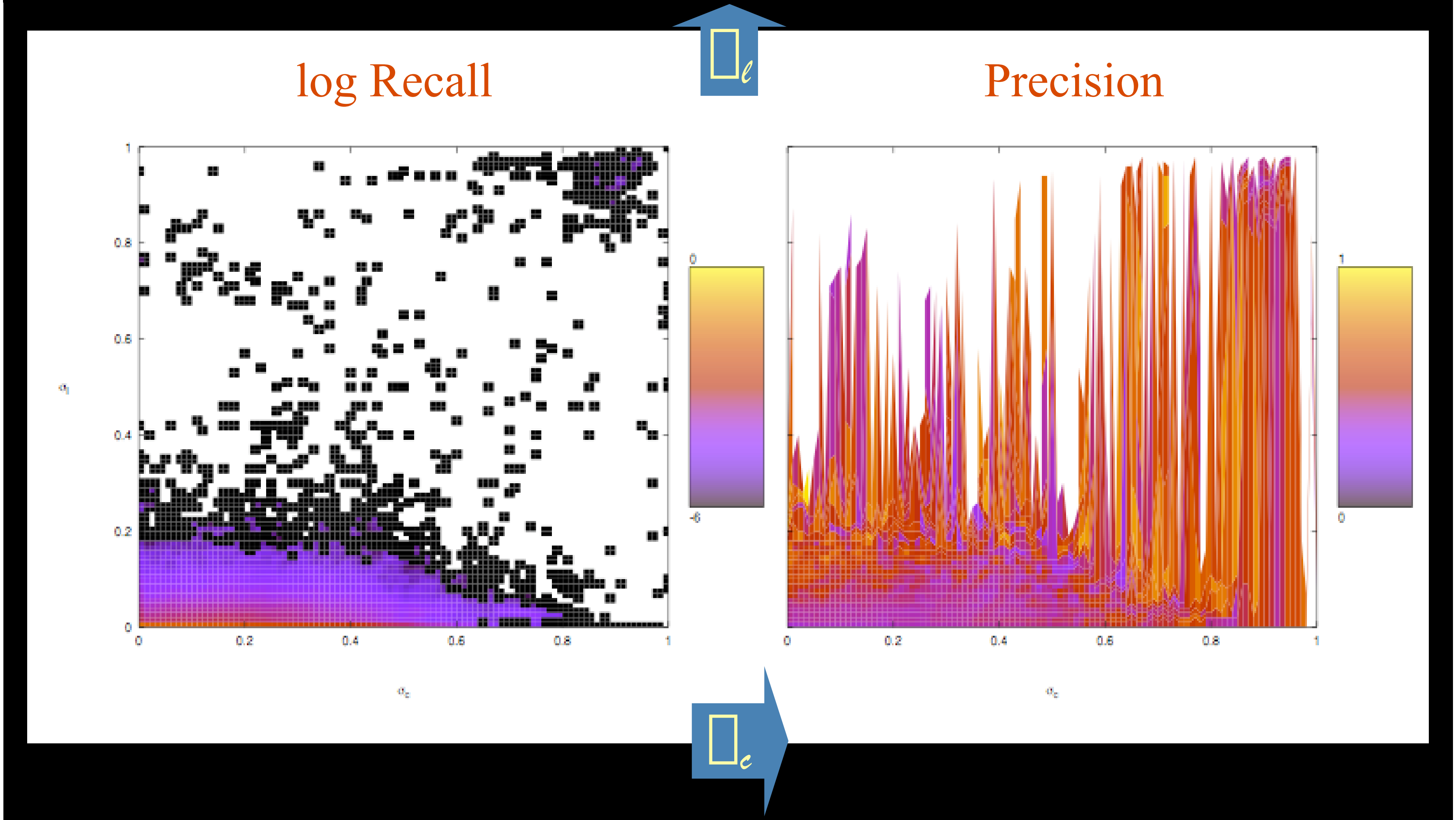
Summing semantic similarity

Semantic maps: *Business*


log Recall

Precision

Semantic maps: *Adult*



log Recall

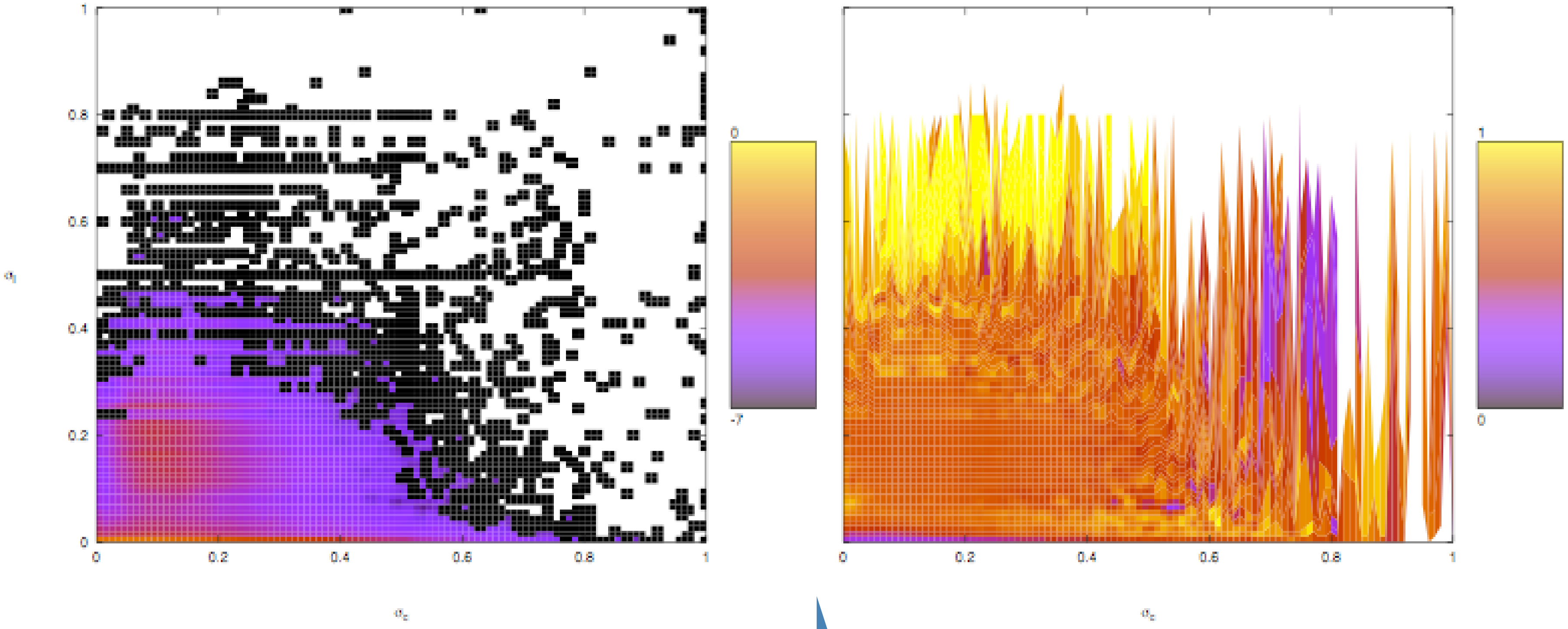
Precision

Semantic maps: *Computers*

$\square_\ell$

log Recall

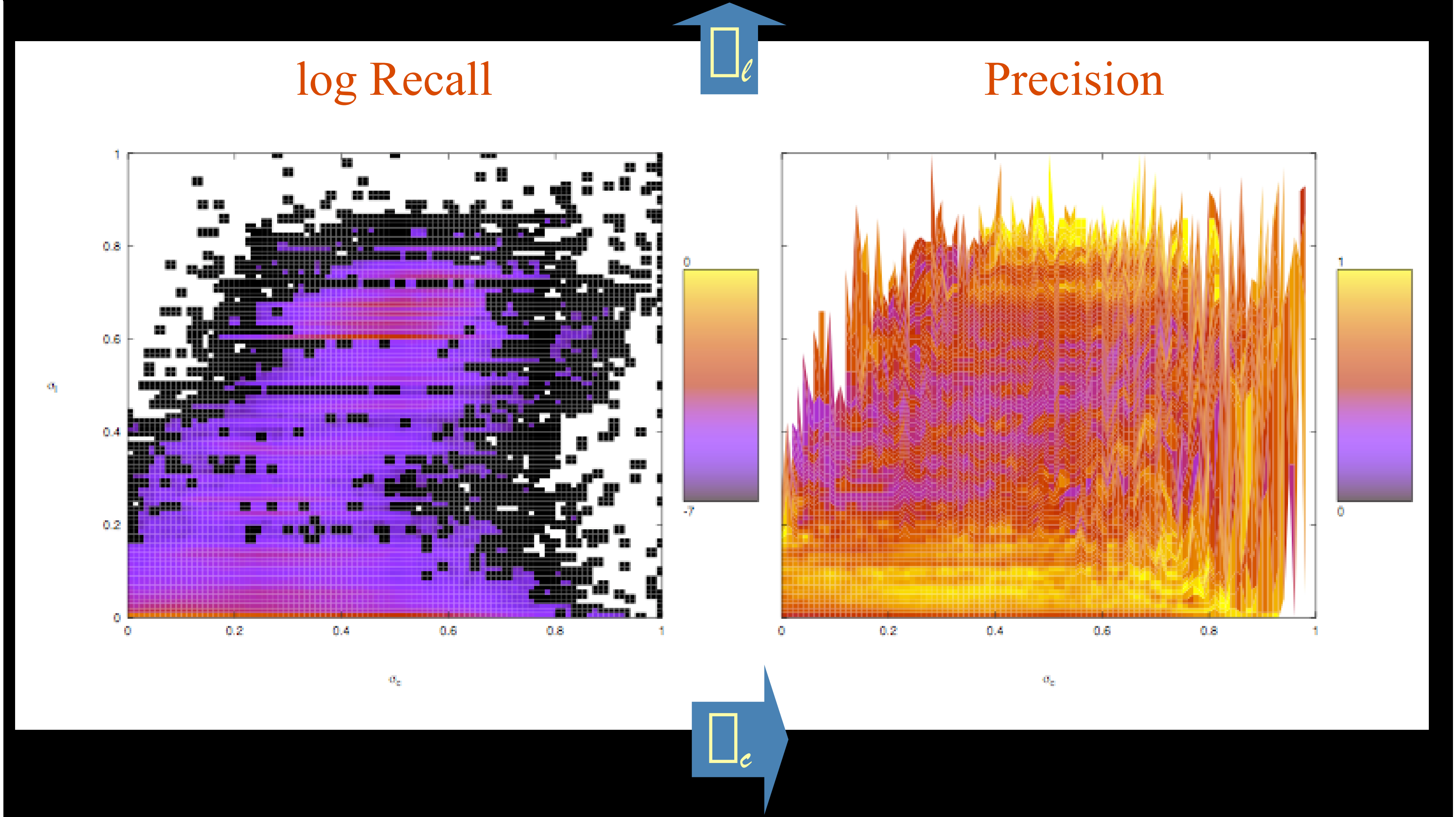
Precision



$\square_c$

WAW 2002

Semantic maps: *Home*



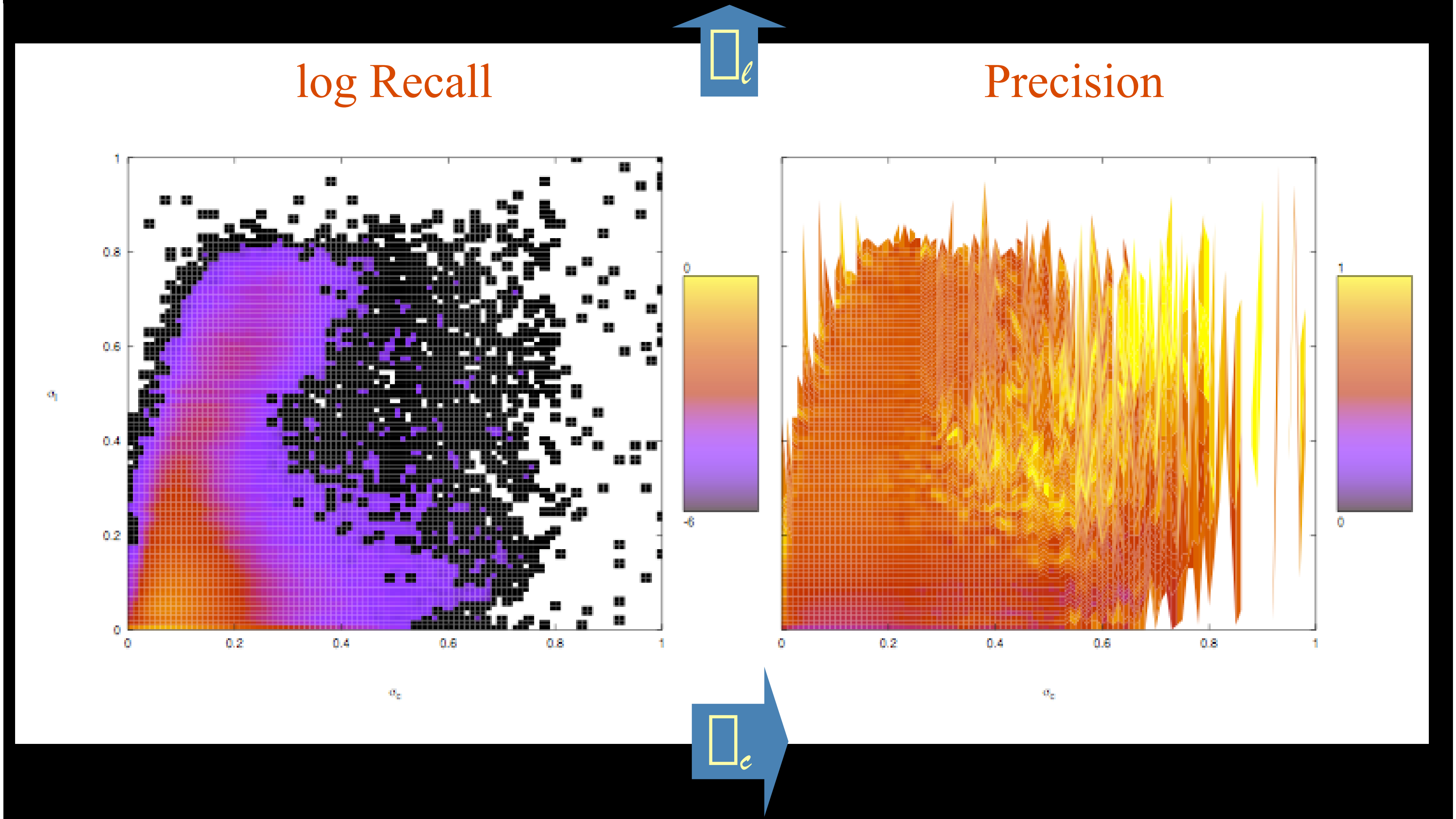
$\square_{\ell}$

log Recall

Precision

$\square_{c}$

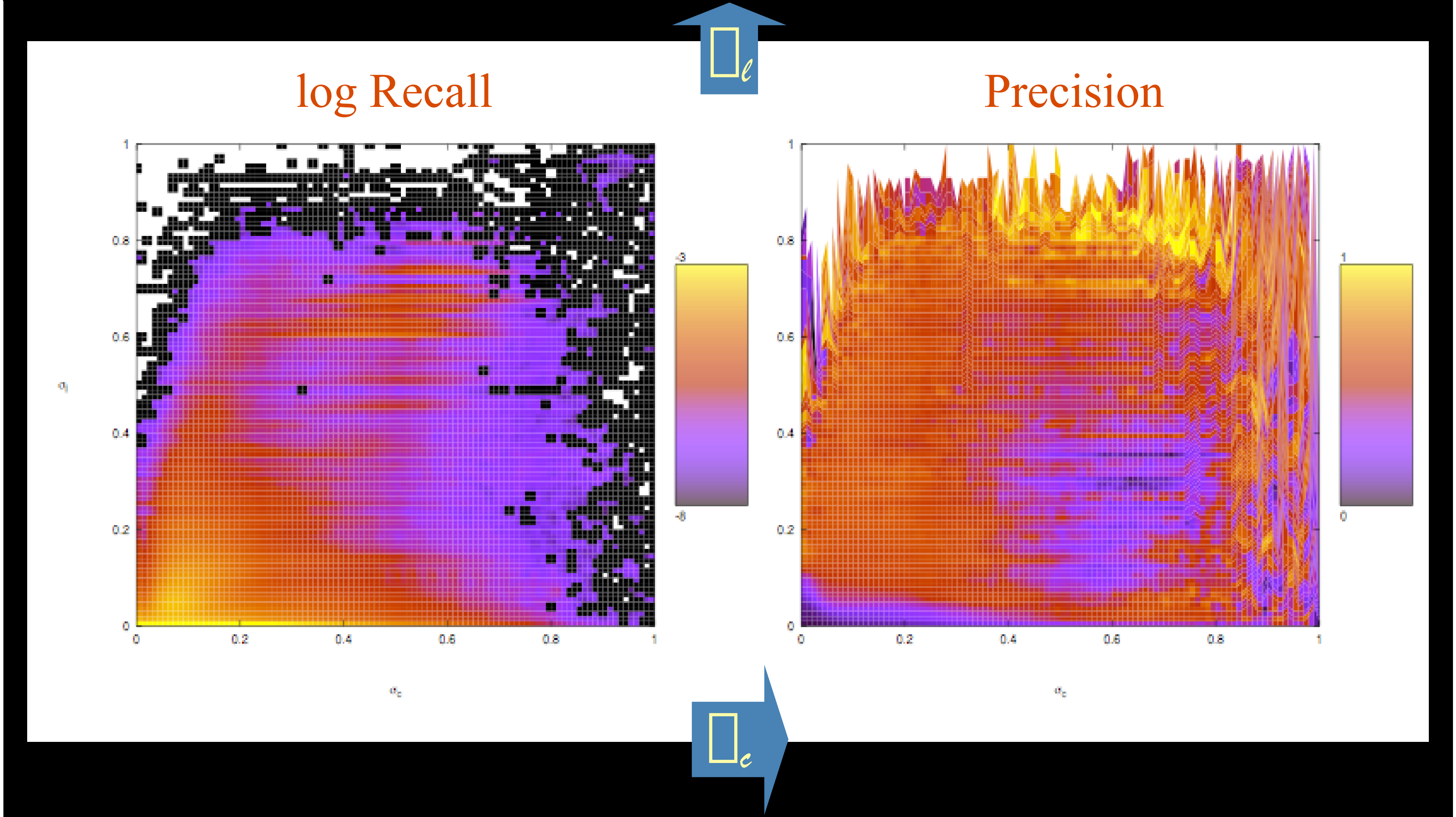
Semantic maps: *News*



log Recall

Precision

Semantic maps: all pairs



So what?

- Interpret performance of search engines
- Understand “ranking optimization”
  - vs filtering
  - vs combinations
- Topical signatures for topical/community portals
- Design “better” (and more scalable) crawlers
  - Topic driven
  - Query driven
  - User/community/peer driven
- Competitive intelligence, security applications

Talk outline

- The topologies of the Web
- Correlations, distributions, projections
- Power laws and Web growth models
- Navigating optimal paths
- Semantic maps
- Questions?

http://dollar.biz.uiowa.edu/~fil/

Research supported by NSF CAREER Award IIS-0133124