Web Networks

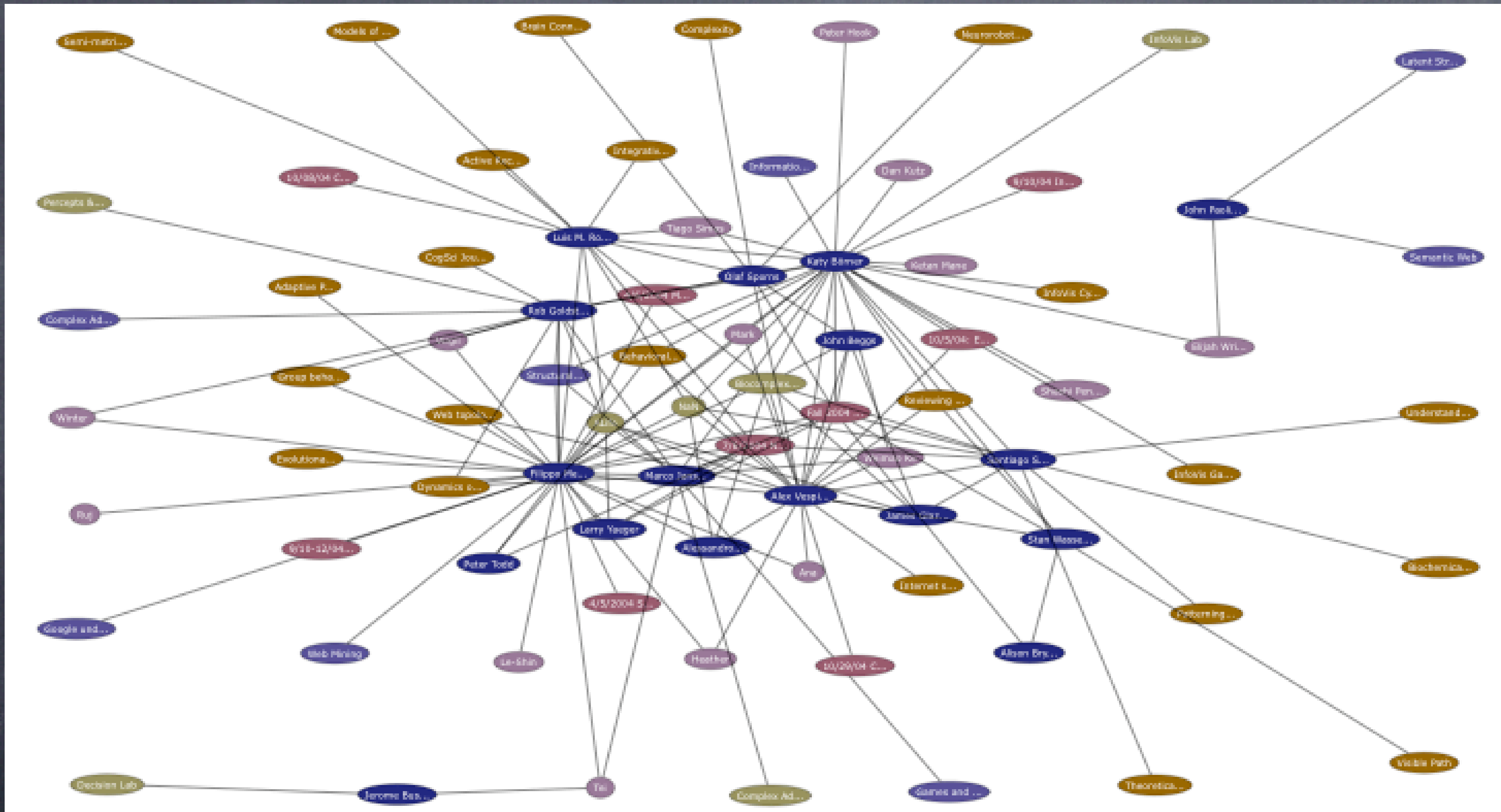
Filippo Menczer

Department of Computer Science
School of Informatics

Indiana University

Research supported by NSF
CAREER Award IIS-0348940

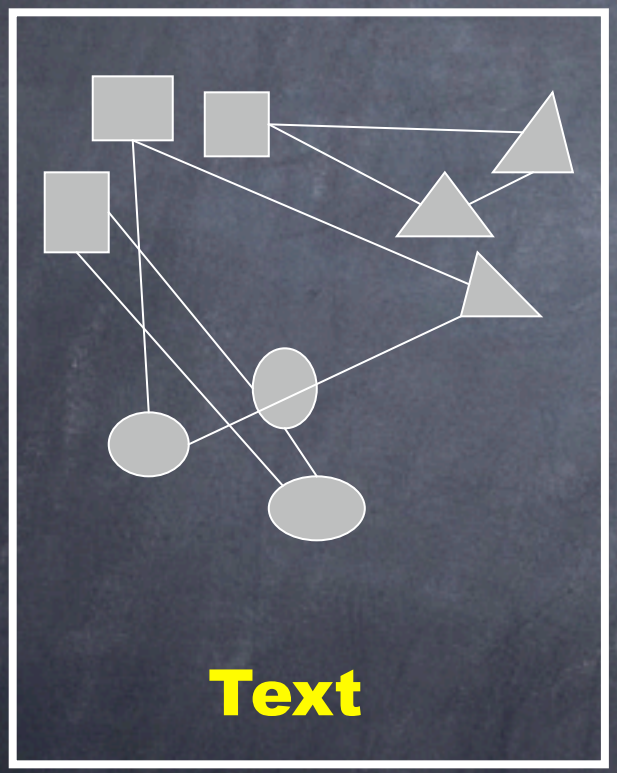
csn.indiana.edu

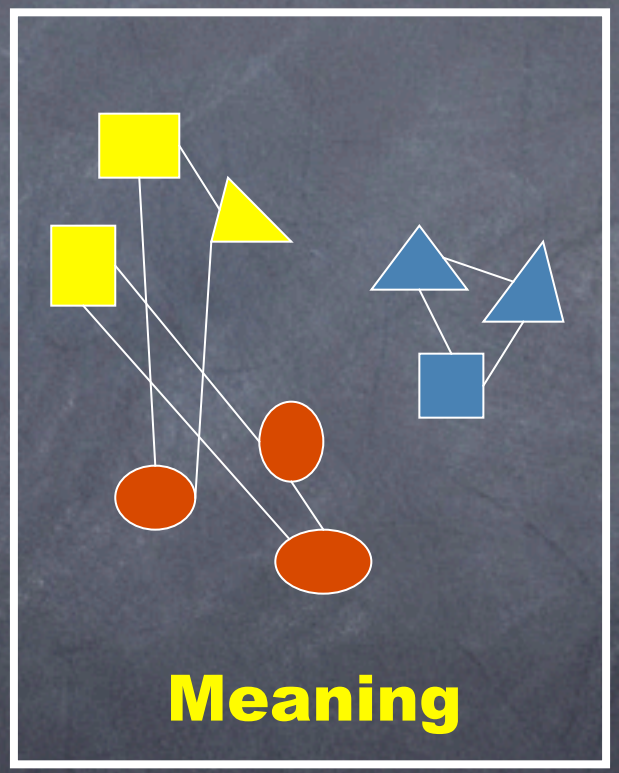


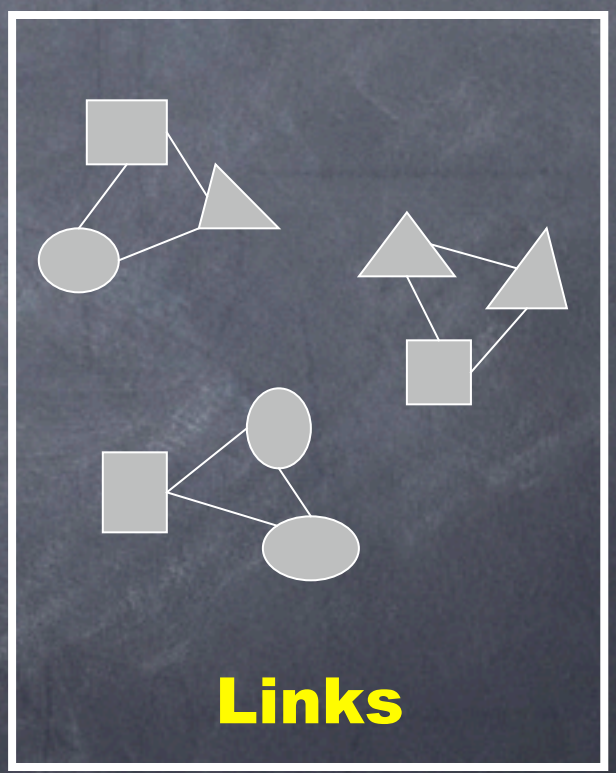
Outline

- ✓ Link network
- ✓ Lexical network
- ✓ Growth models
- Semantic network
- Peer search network
- Traffic network

Three network topologies

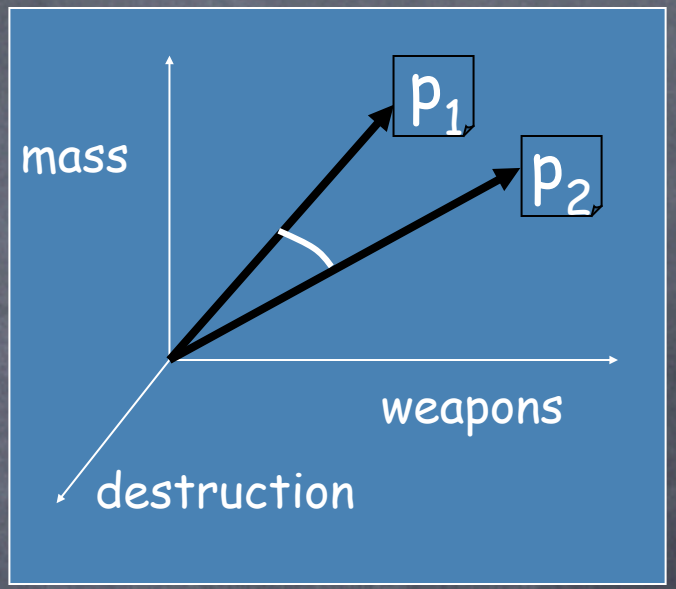




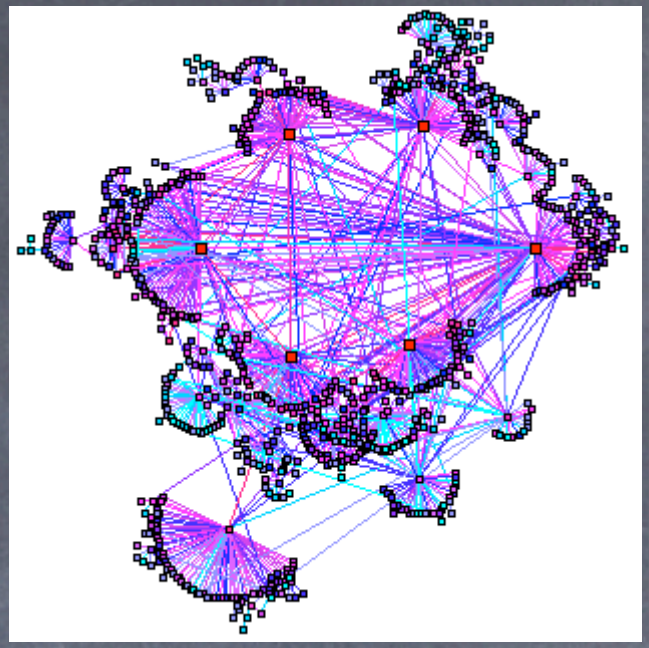


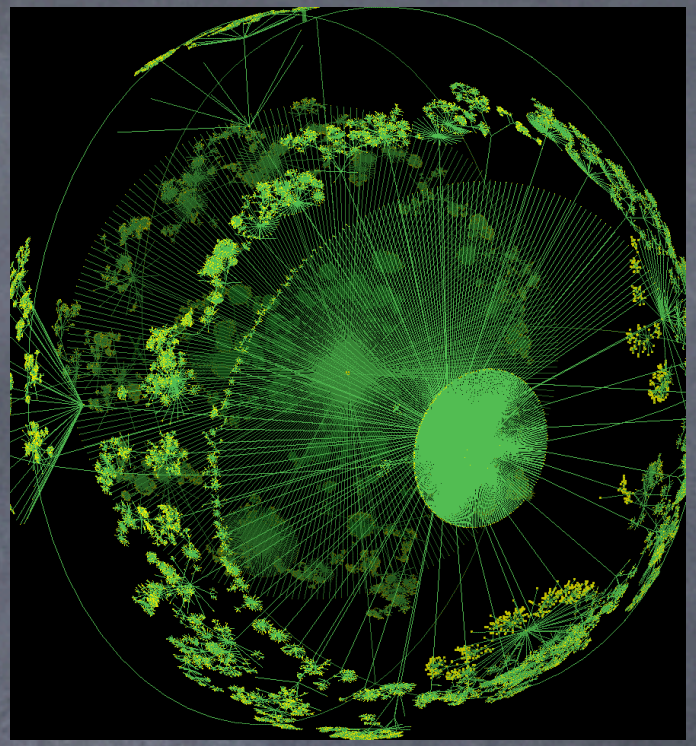
The Web as a text corpus

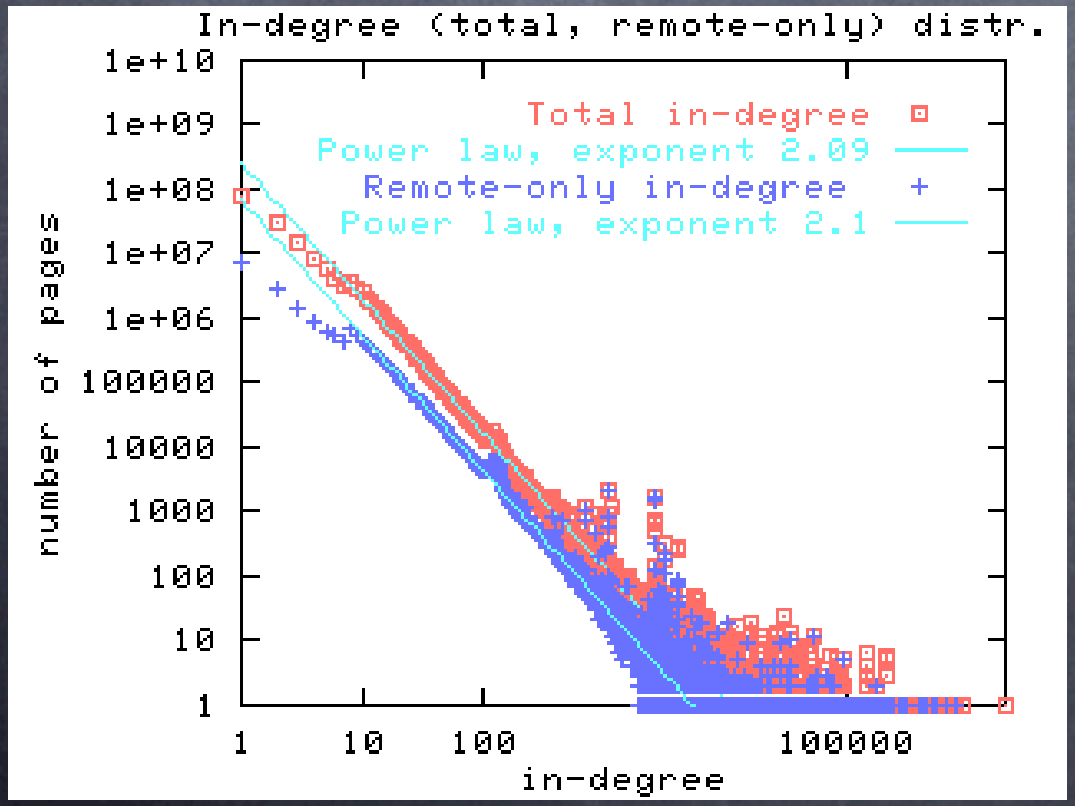
- Pages close in word vector space tend to be related
  - Cluster hypothesis (van Rijsbergen 1979)
  - The WebCrawler (Pinkerton 1994)
  - The whole first generation of search engines

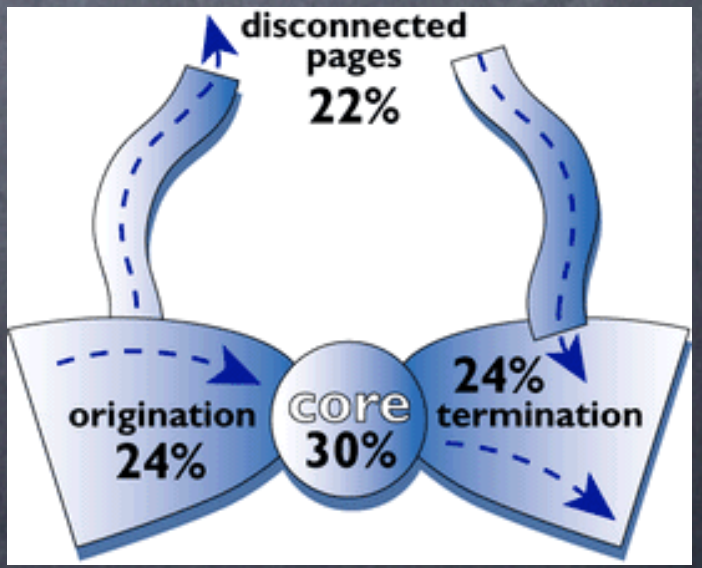


Enter the Web's link structure



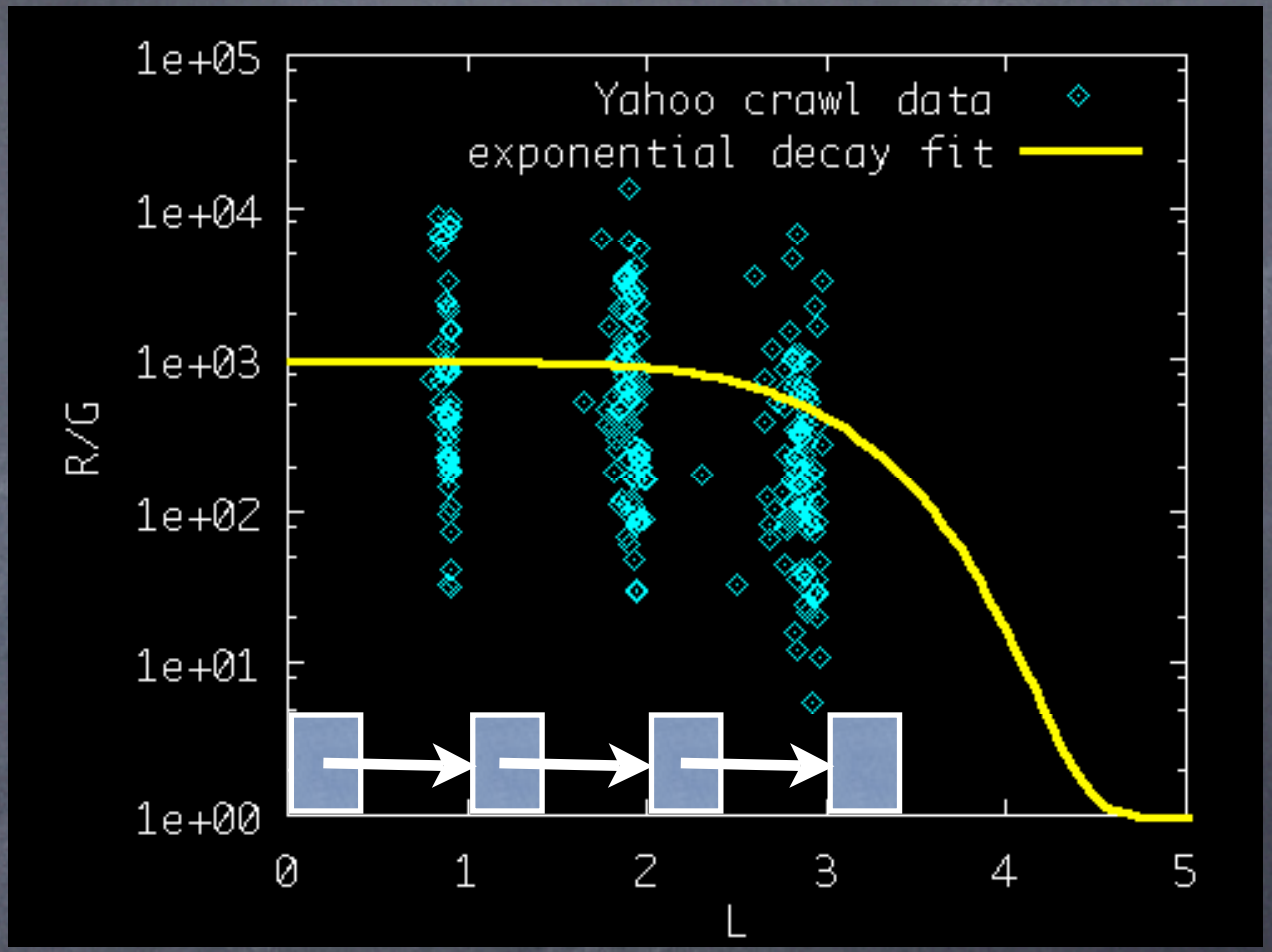






Mining the Web's link cues

- Pages that link to each other tend to be related
  - Link-cluster conjecture (Menczer 1997)
  - Link eigenvector analysis
    - HITS, hubs and authorities (Kleinberg & al 1998, …)
    - Google's PageRank (Brin & Page 1998, …)
- The second generation of search engines



Web growth models

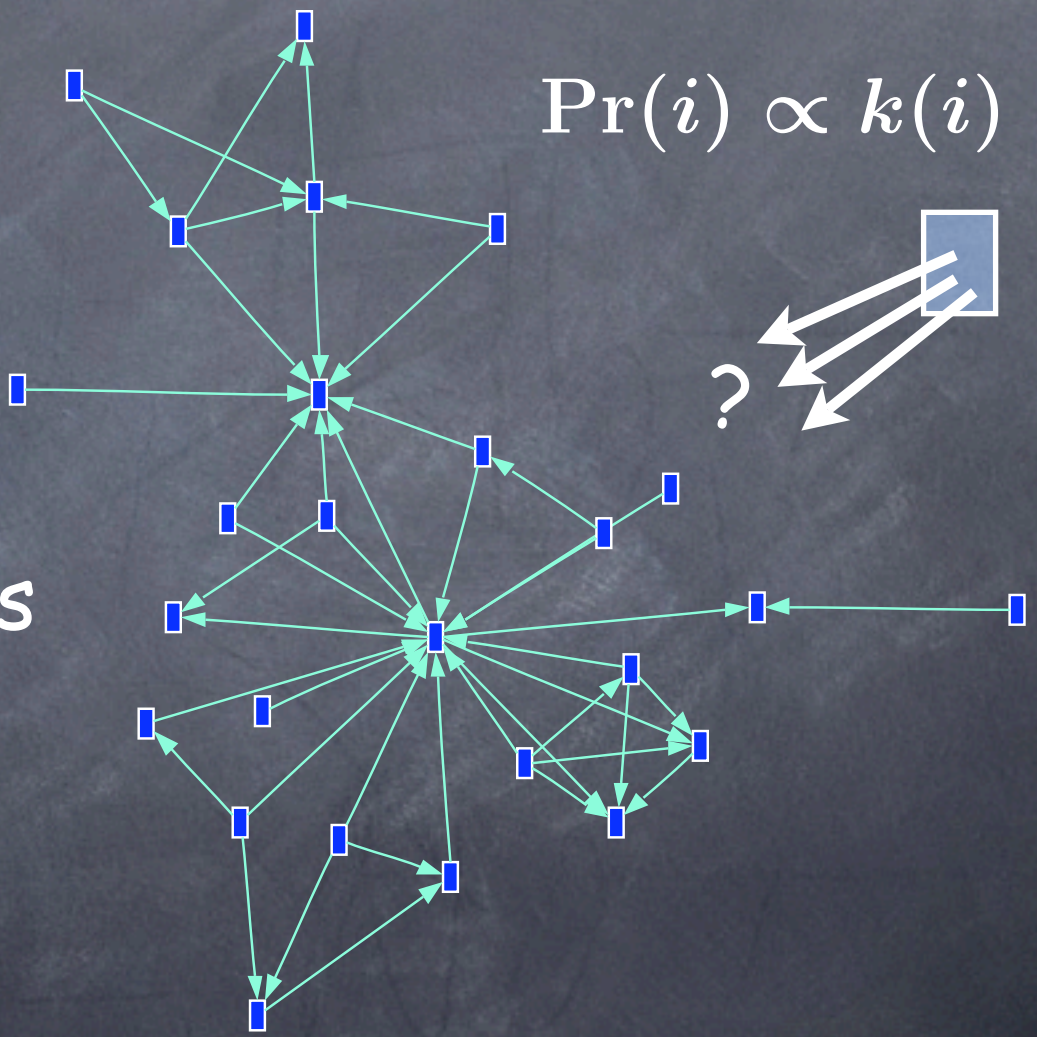
How are links created and why content matters

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



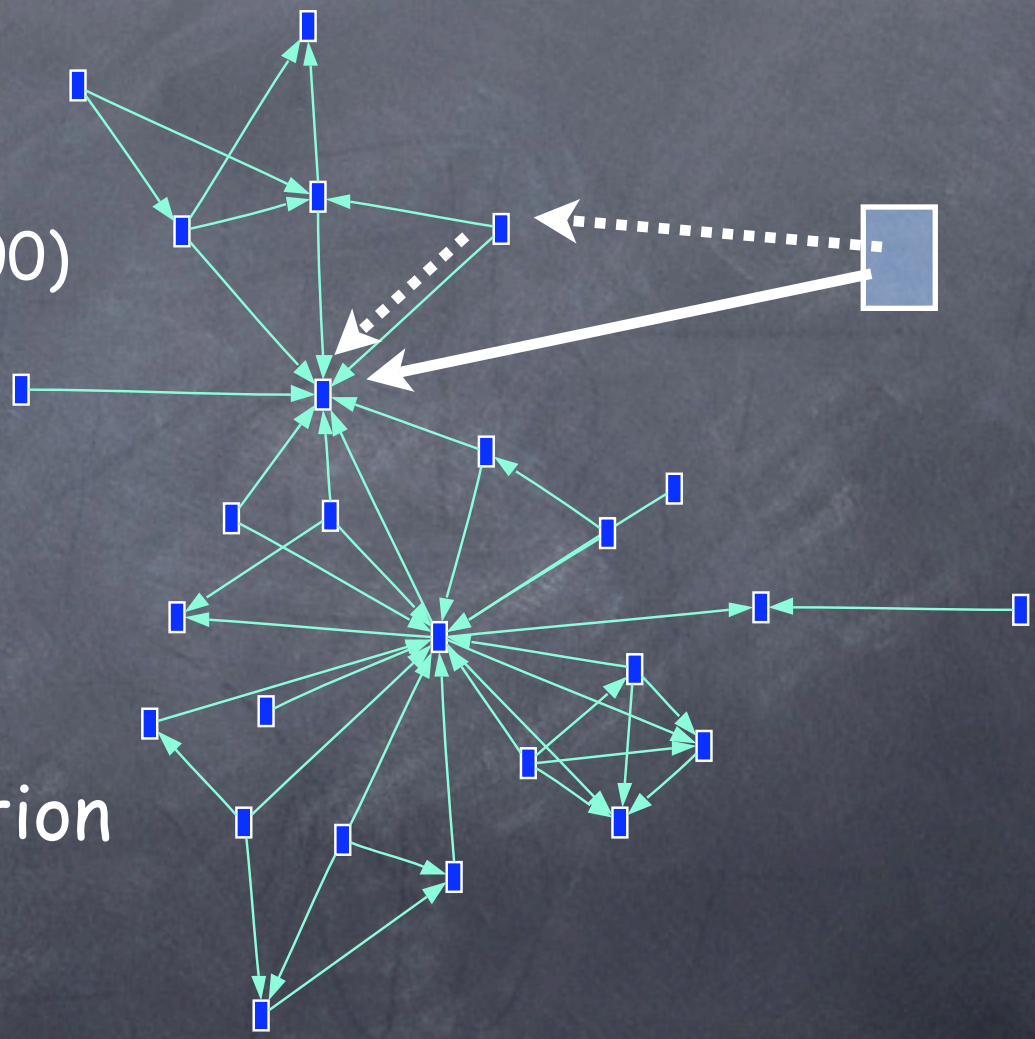
$$\Pr(i) = \frac{k(i)}{mt} \implies \Pr(k) \sim k^{-\gamma}$$

Other growth models

Web copying

(Kleinberg, Kumar & al 1999, 2000)

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



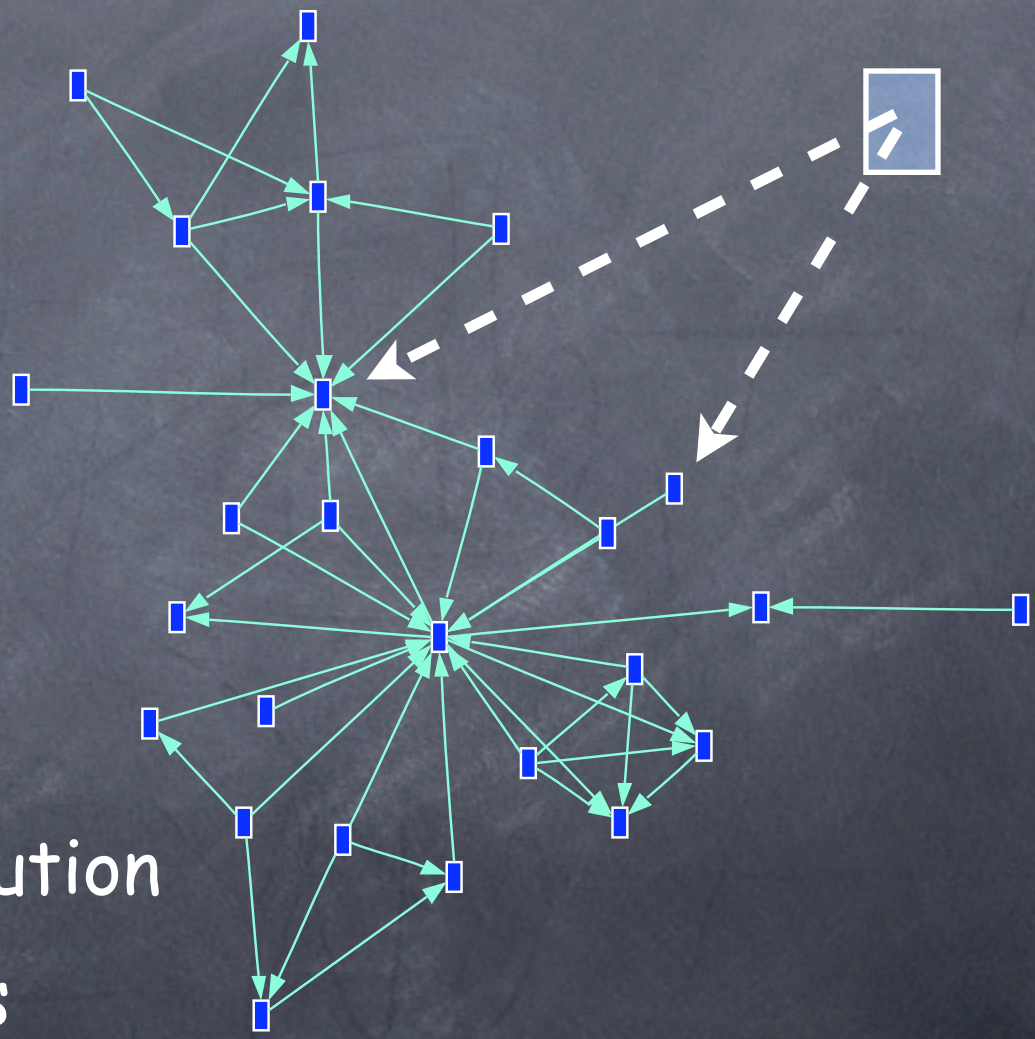
- same indegree distribution
- no need to know degree

Other growth models

Random mixture

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

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



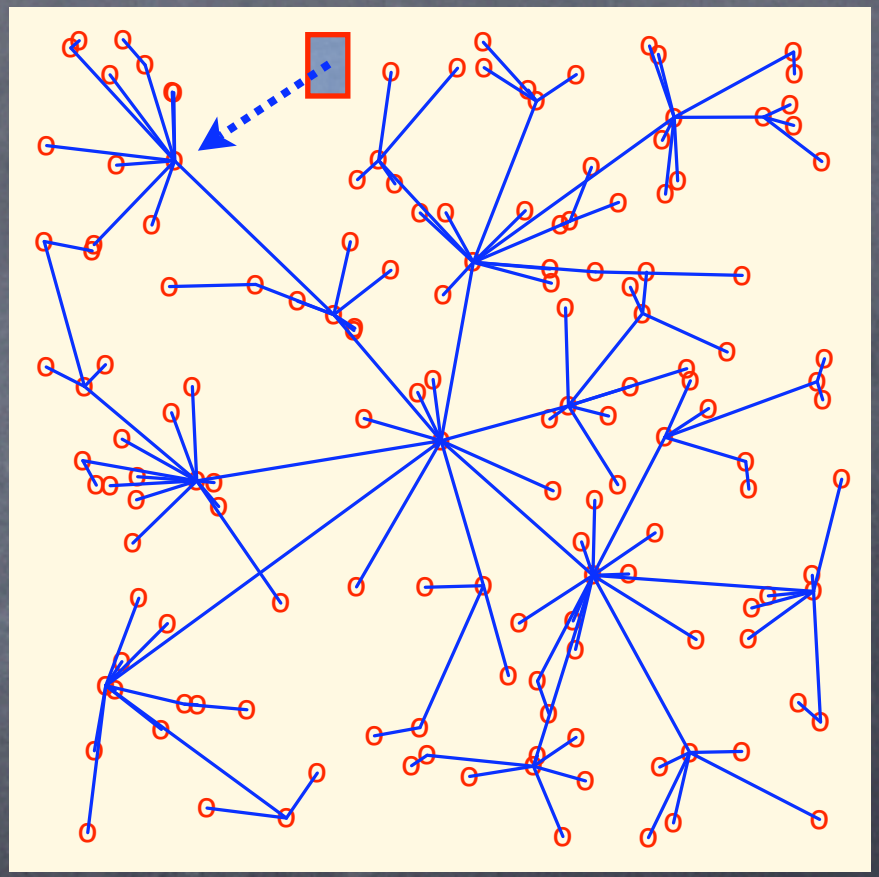
- winners don't take all
- general indegree distribution
- fits non-power-law cases

Other growth models

Mixture with Euclidean distance

(HOT: Fabrikant, Koutsoupias & Papadimitriou 2002)

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



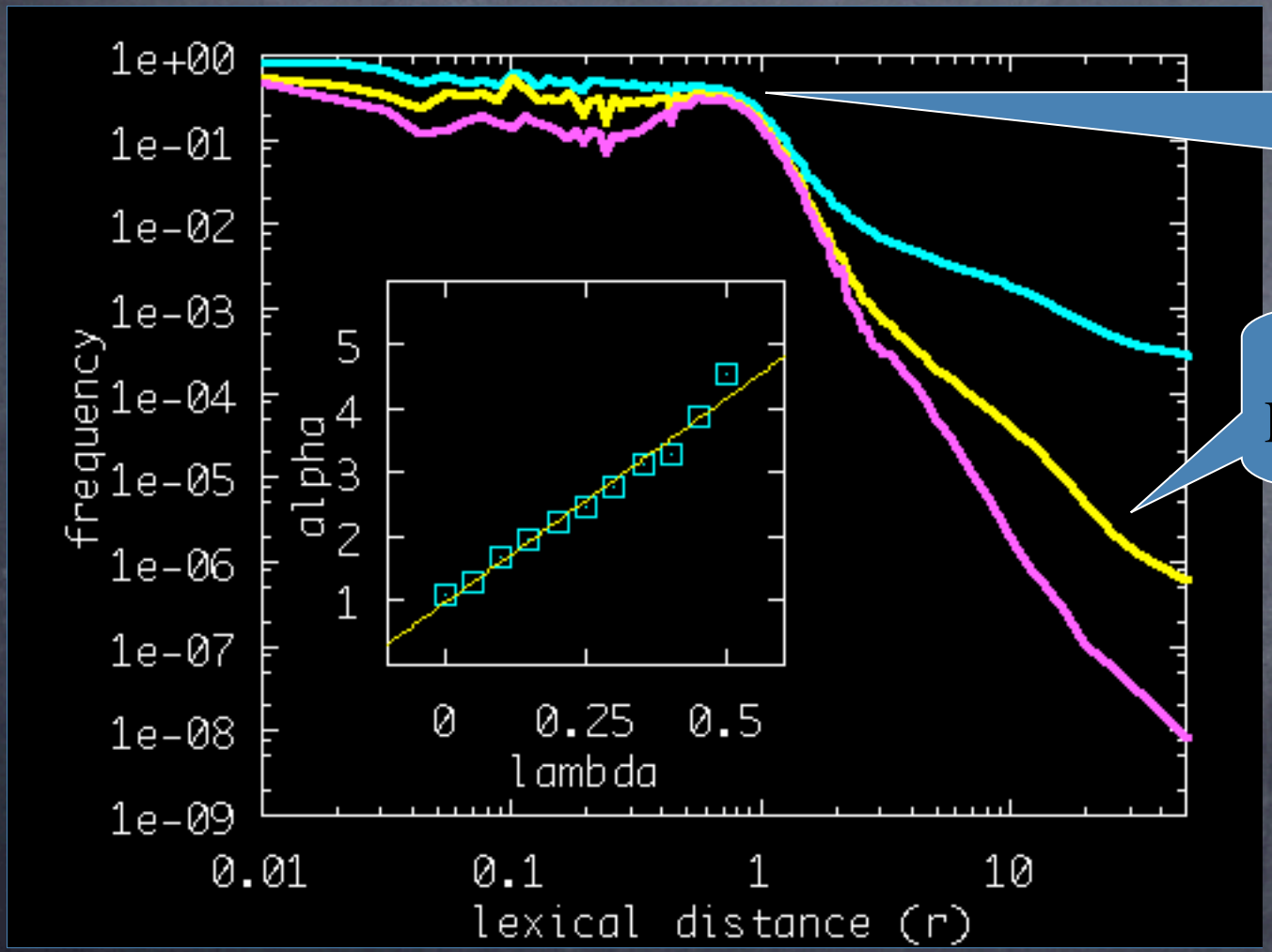
- tradeoff between centrality and geometric locality
- fits power-law in certain critical trade-off regimes

What about content?

Link probability vs lexical distance

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

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



Phase transition ρ^*

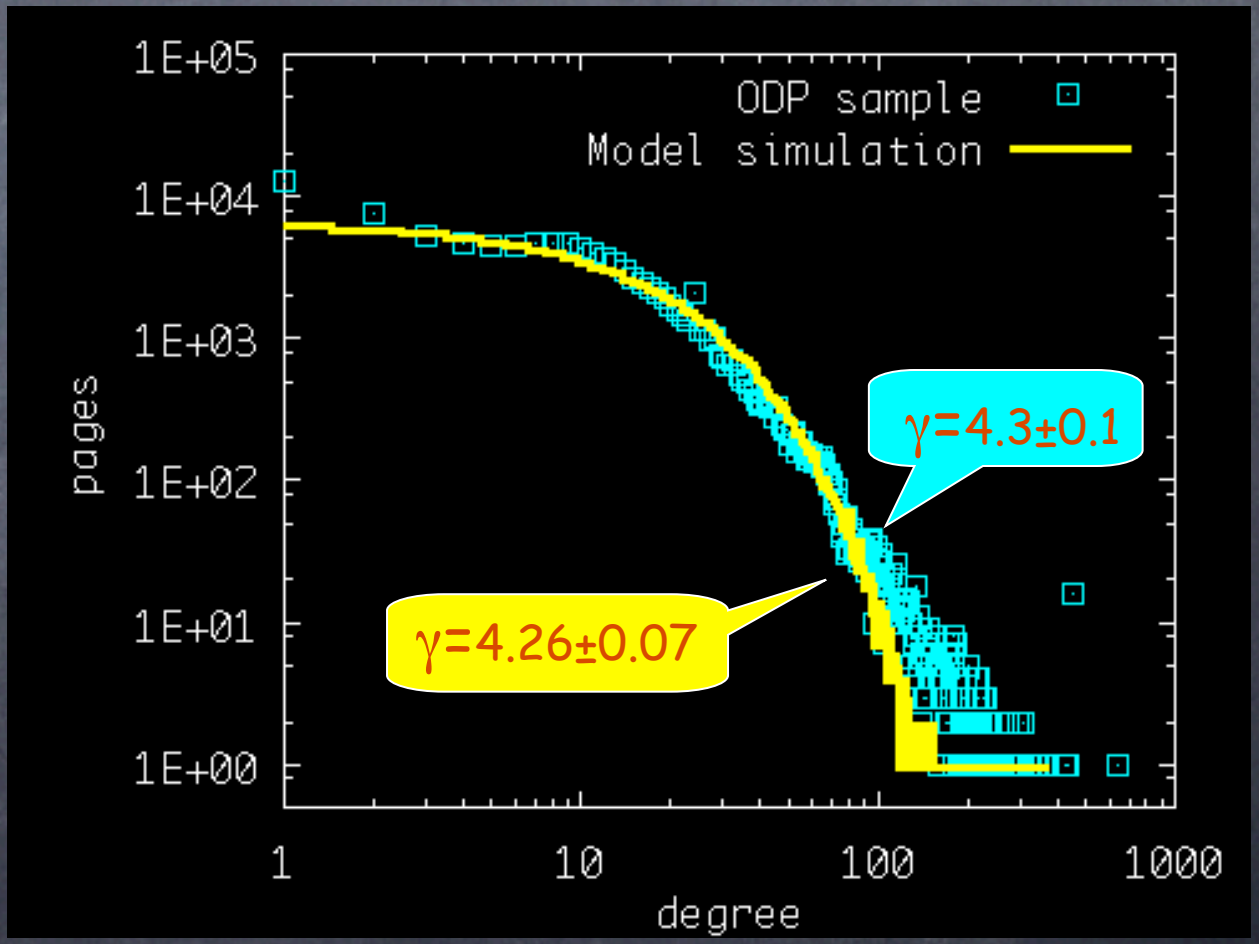
Power law tail $\Pr(\lambda \mid \rho) \sim \rho^{-\alpha(\lambda)}$

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

Local content-based growth model

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

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

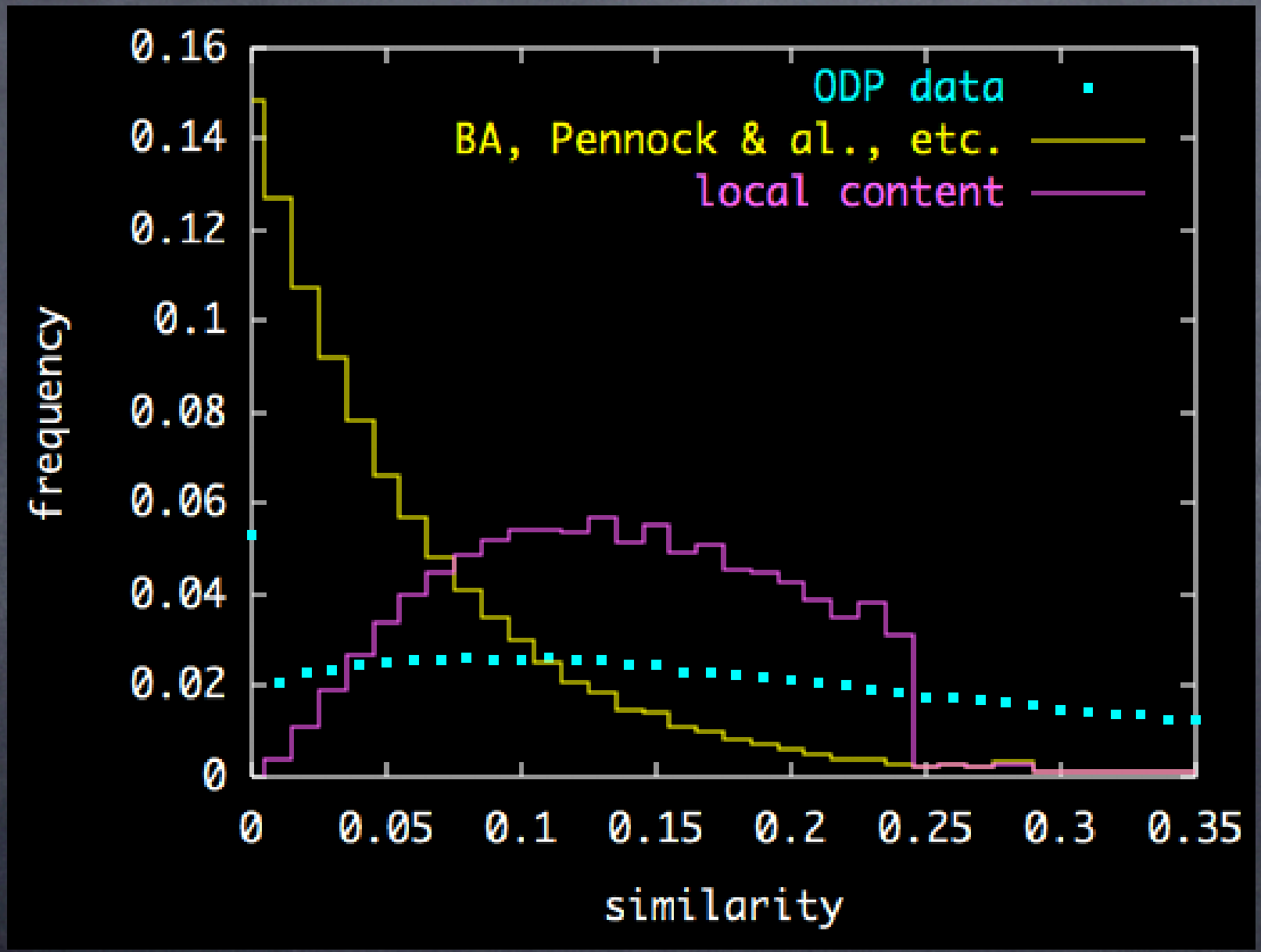


So, many models can predict degree distributions...

- Which is "right" ?
- Need an independent observation (other than degree) to validate models
- Distribution of content similarity across linked pairs
  - Across all pairs: $\Pr(\sigma_c) \sim 10^{-7\sigma_c}$

(Why?!?)

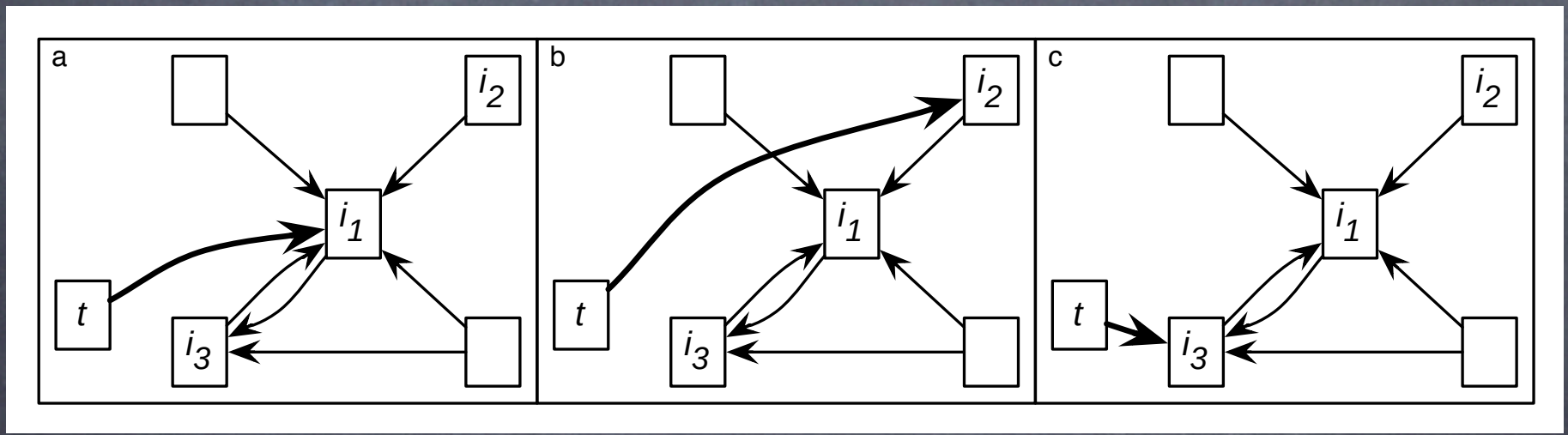
None of these models is right!



Back to the mixture model

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

degree-uniform mixture



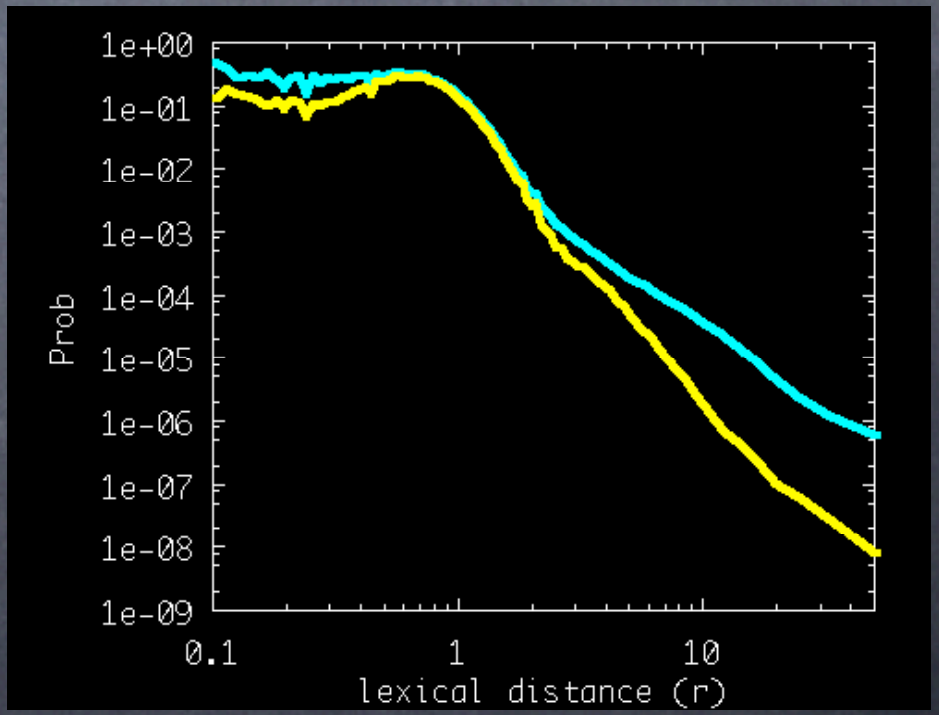
Bias choice by content similarity instead of uniform distribution

Degree-similarity mixture model

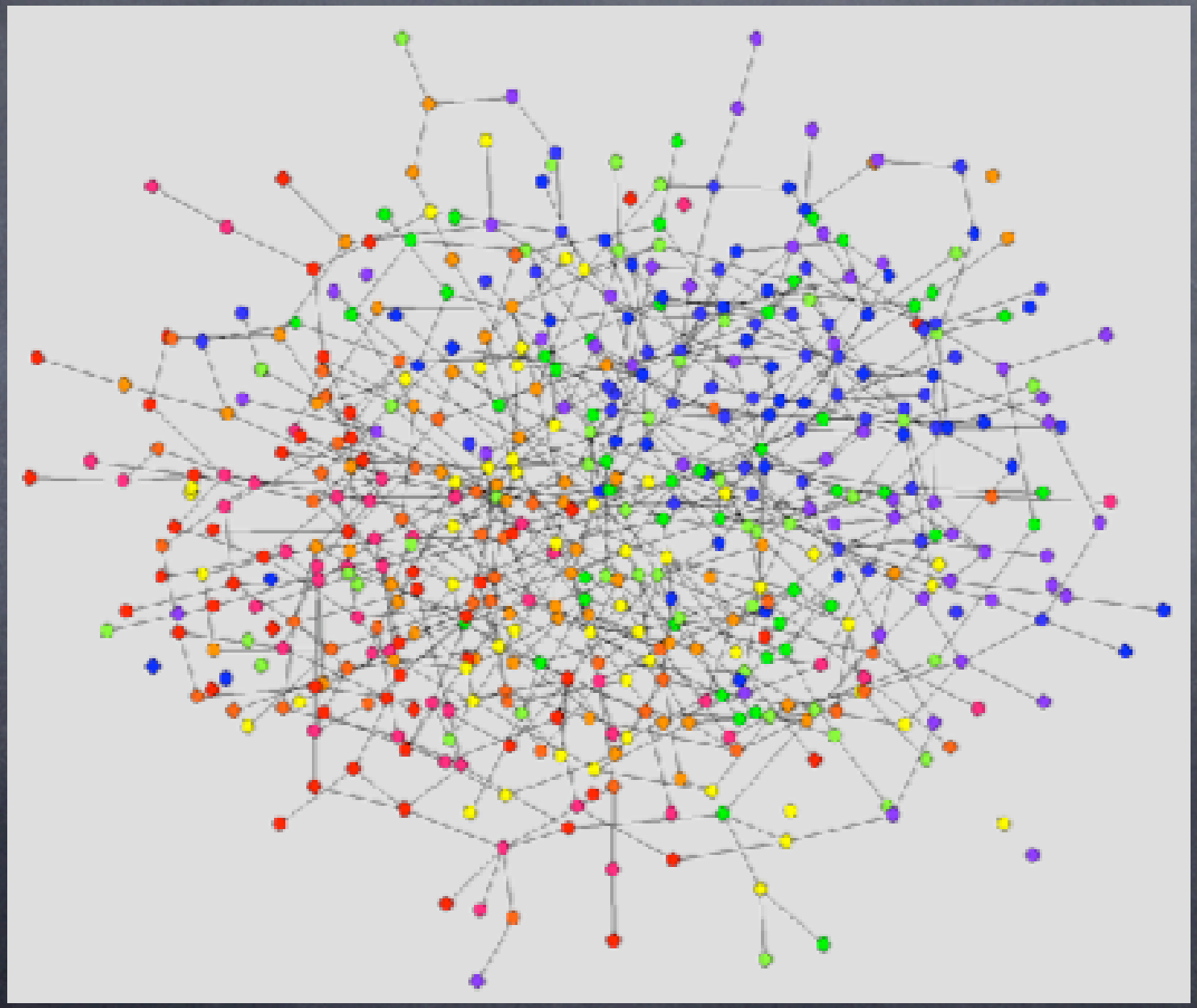
$$\Pr(i) \propto \psi \cdot \hat{\Pr}(i) + (1 - \psi) \cdot \frac{k(i)}{mt}$$

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

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

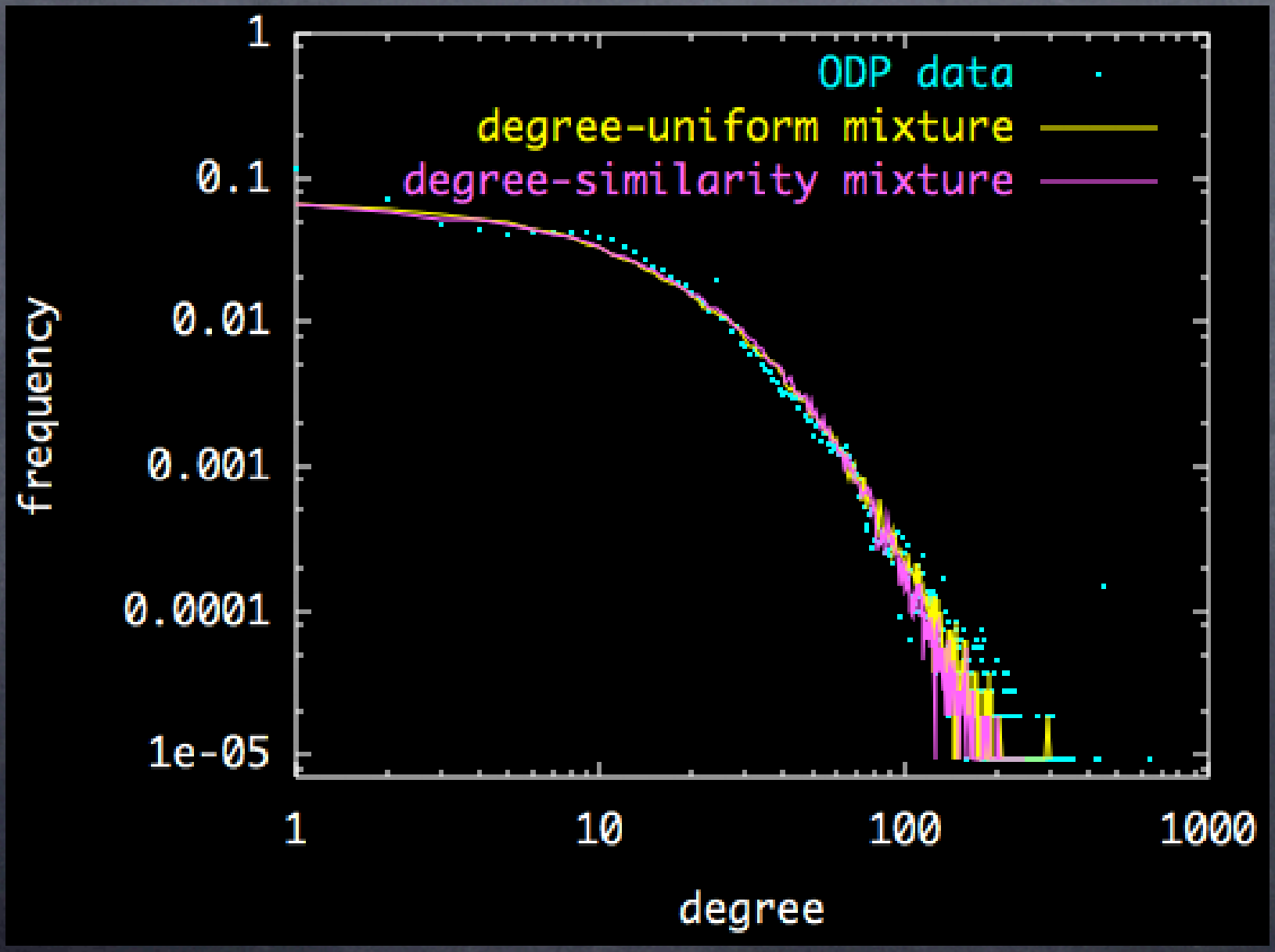


Build it...

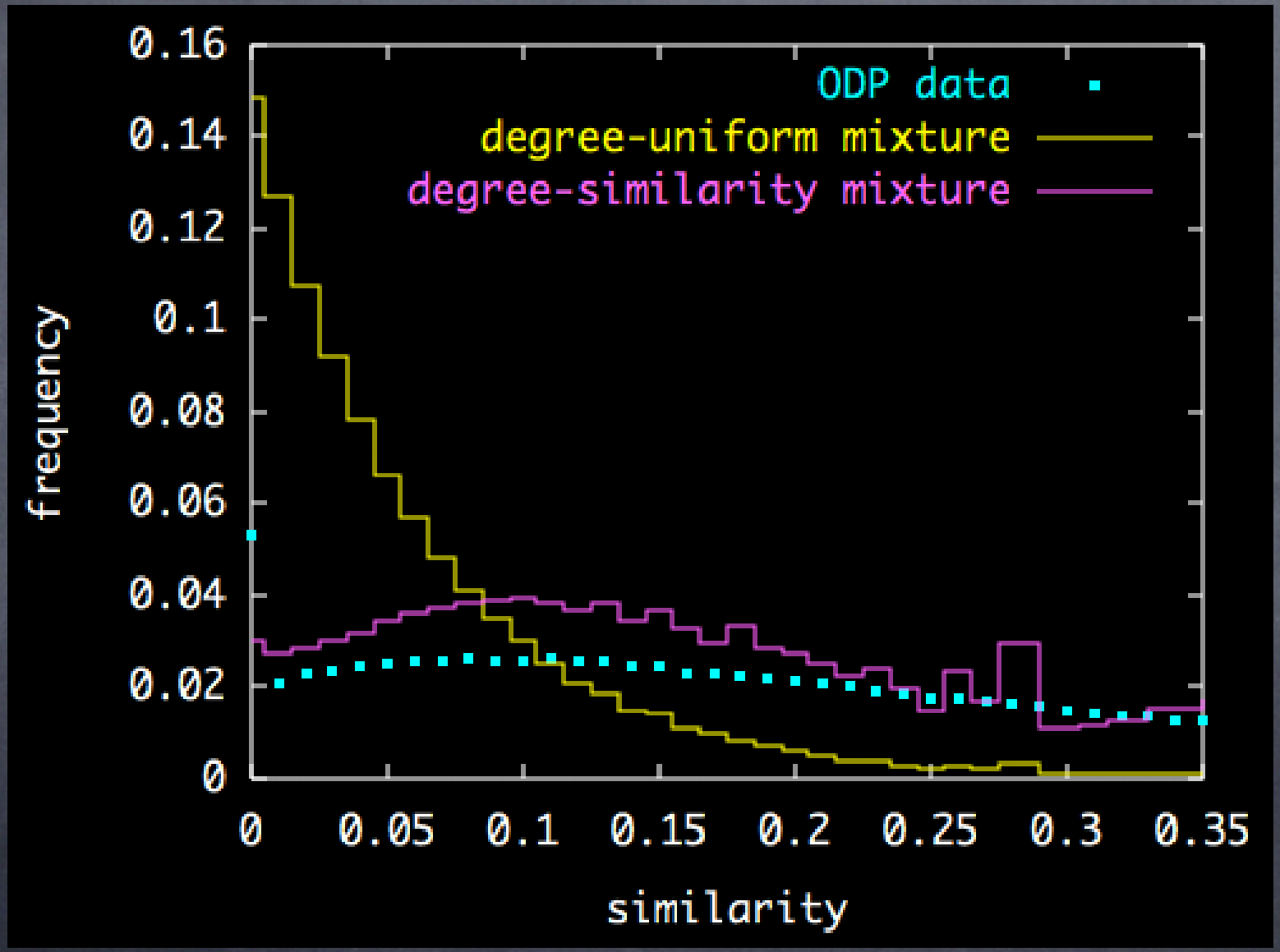


(M.M.)

Both mixture models get the degree distribution right...

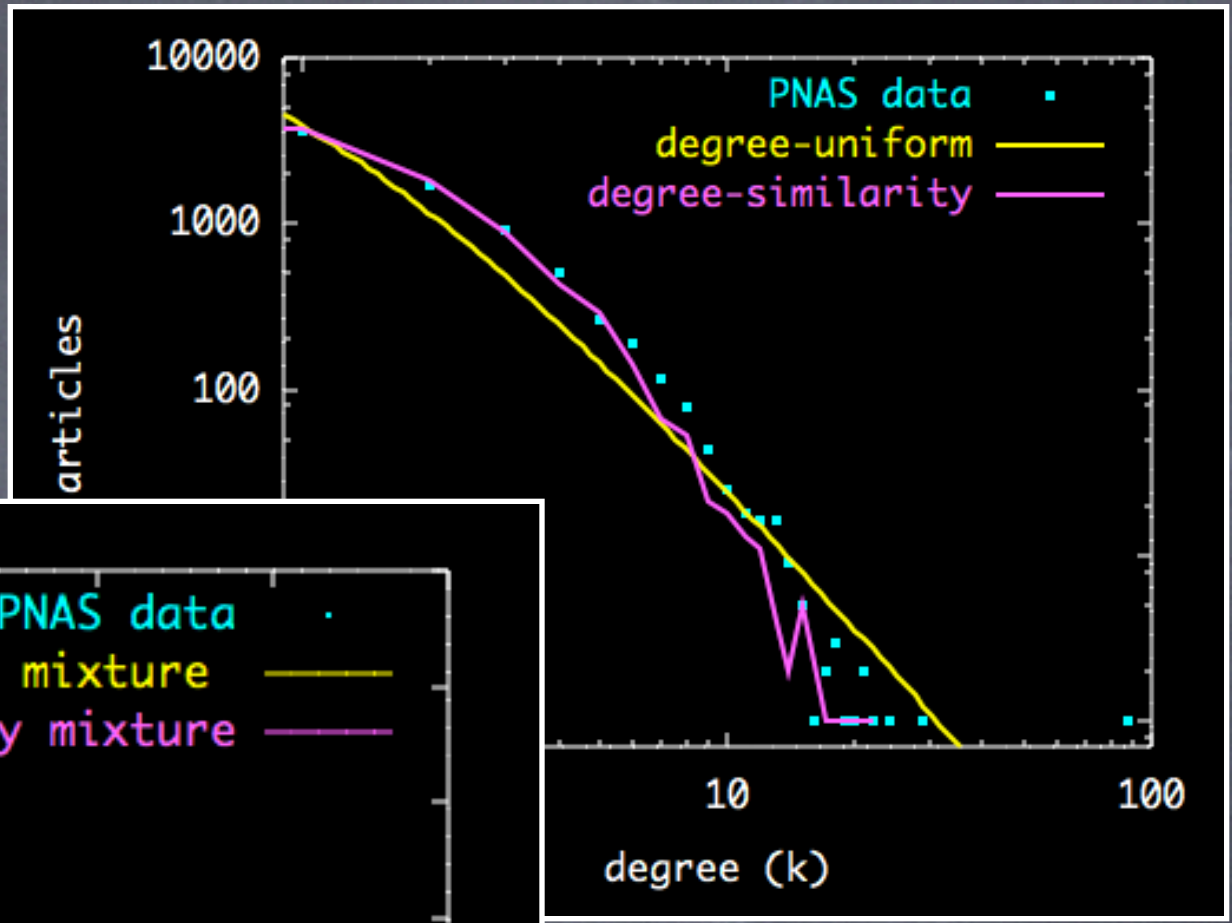


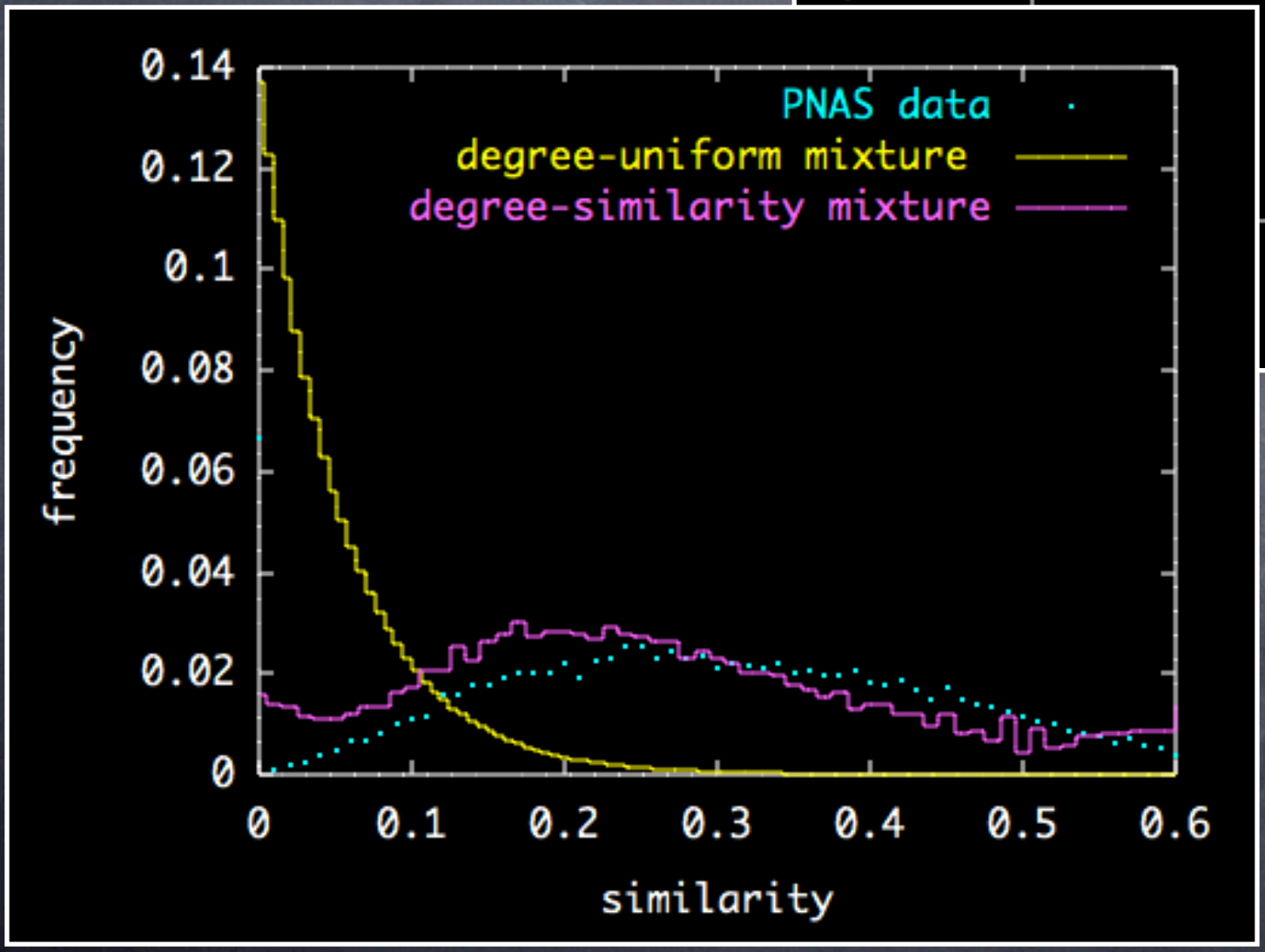
…but the degree-similarity mixture model predicts the similarity distribution better



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

Citation networks





15,785 articles published in PNAS between 1997 and 2002

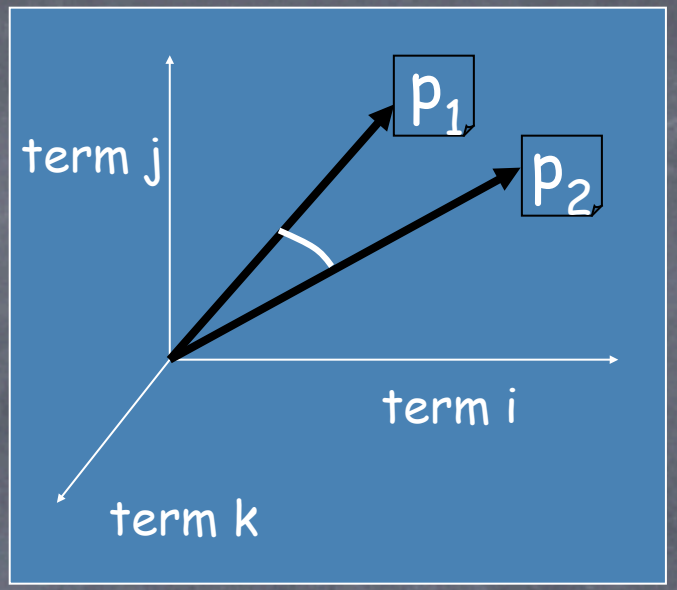
What now?

- Understand exponential distribution of similarity
- Growth model to explain evolution of both link topology and content similarity
  - With Alex Vespignani & Sandro Flammini

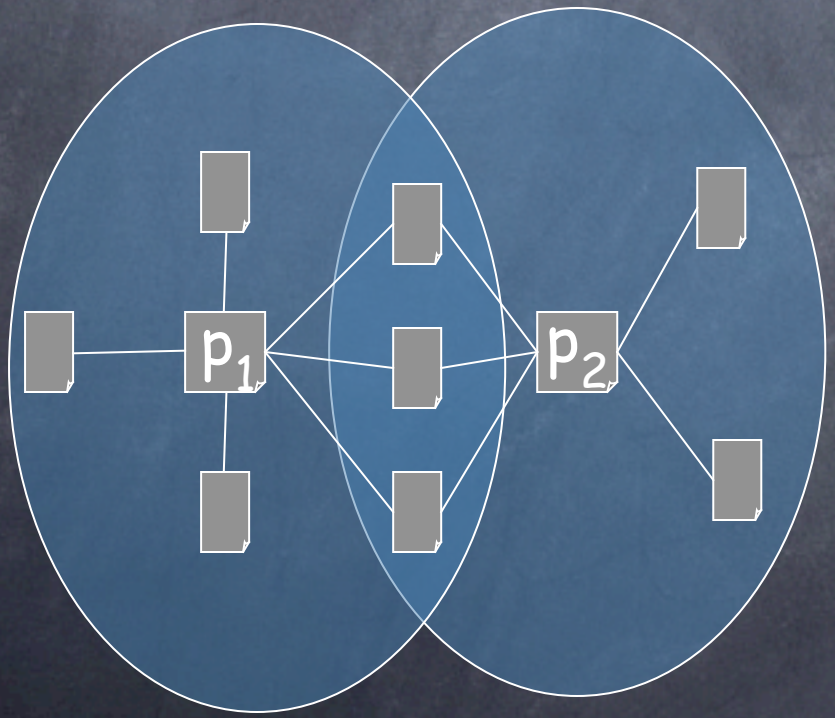
Mapping the relationship between links, content, and semantic topologies

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

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



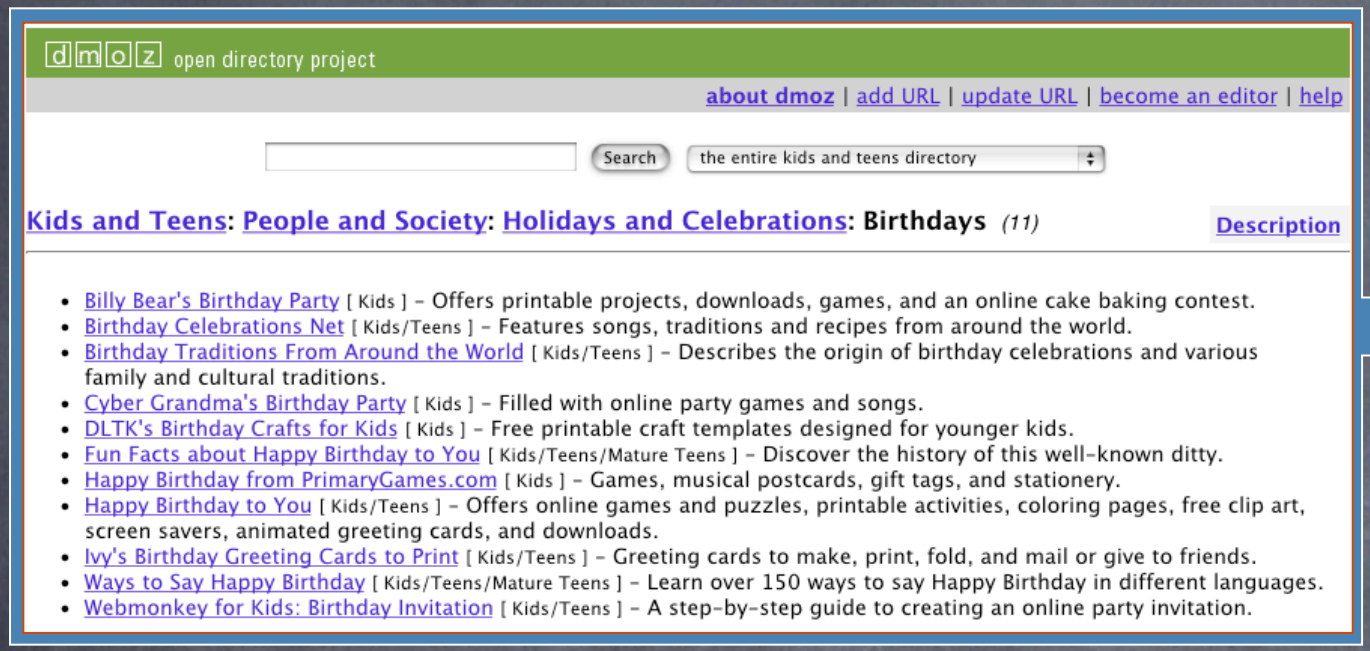
Content similarity

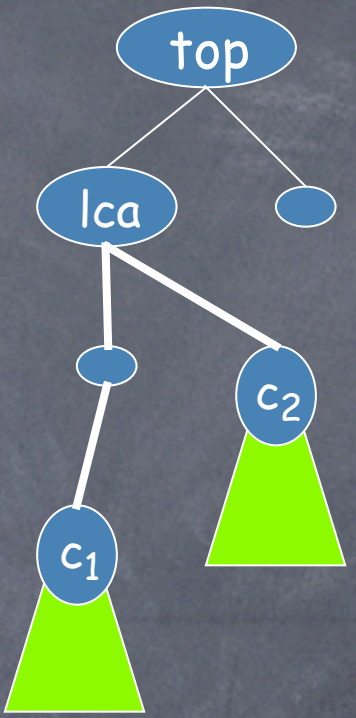


Link similarity

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

Semantic similarity



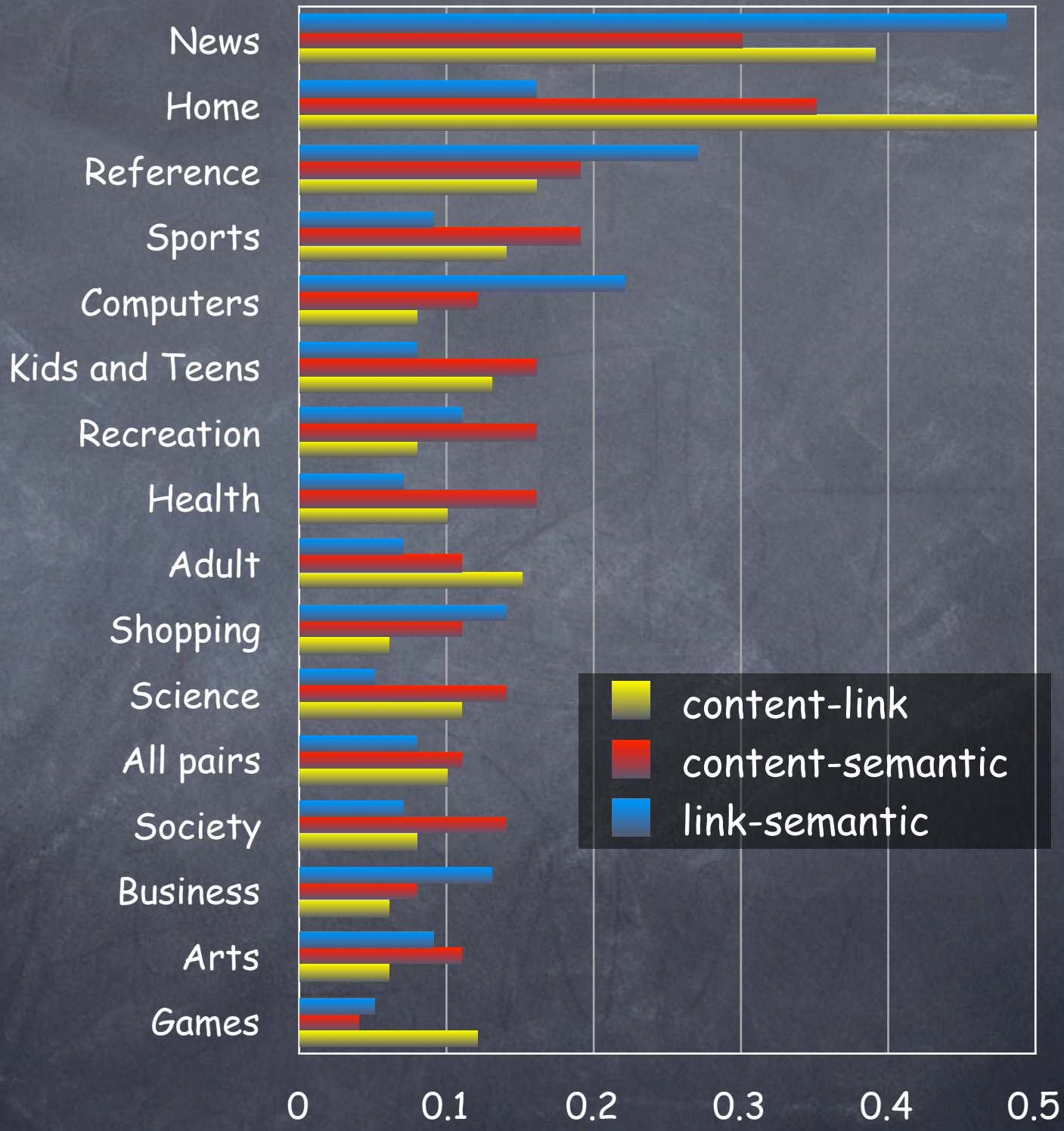


- Information-theoretic measure based on classification tree (Lin 1998)

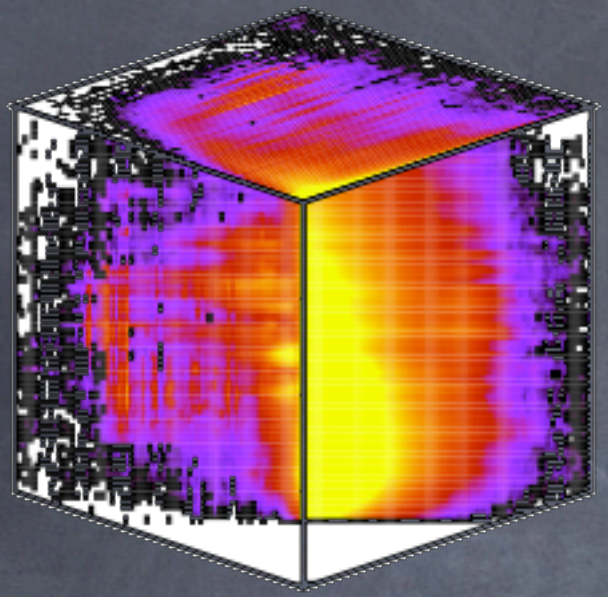
$$\sigma_s(c_1,c_2) = \frac{2\log\Pr[\mathrm{lca}(c_1,c_2)]}{\log\Pr[c_1]+\log\Pr[c_2]}$$

- Classic path distance in special case of balanced tree

Correlations between similarities



European Physical Journal B 38(2): 211-221, 2004



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

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

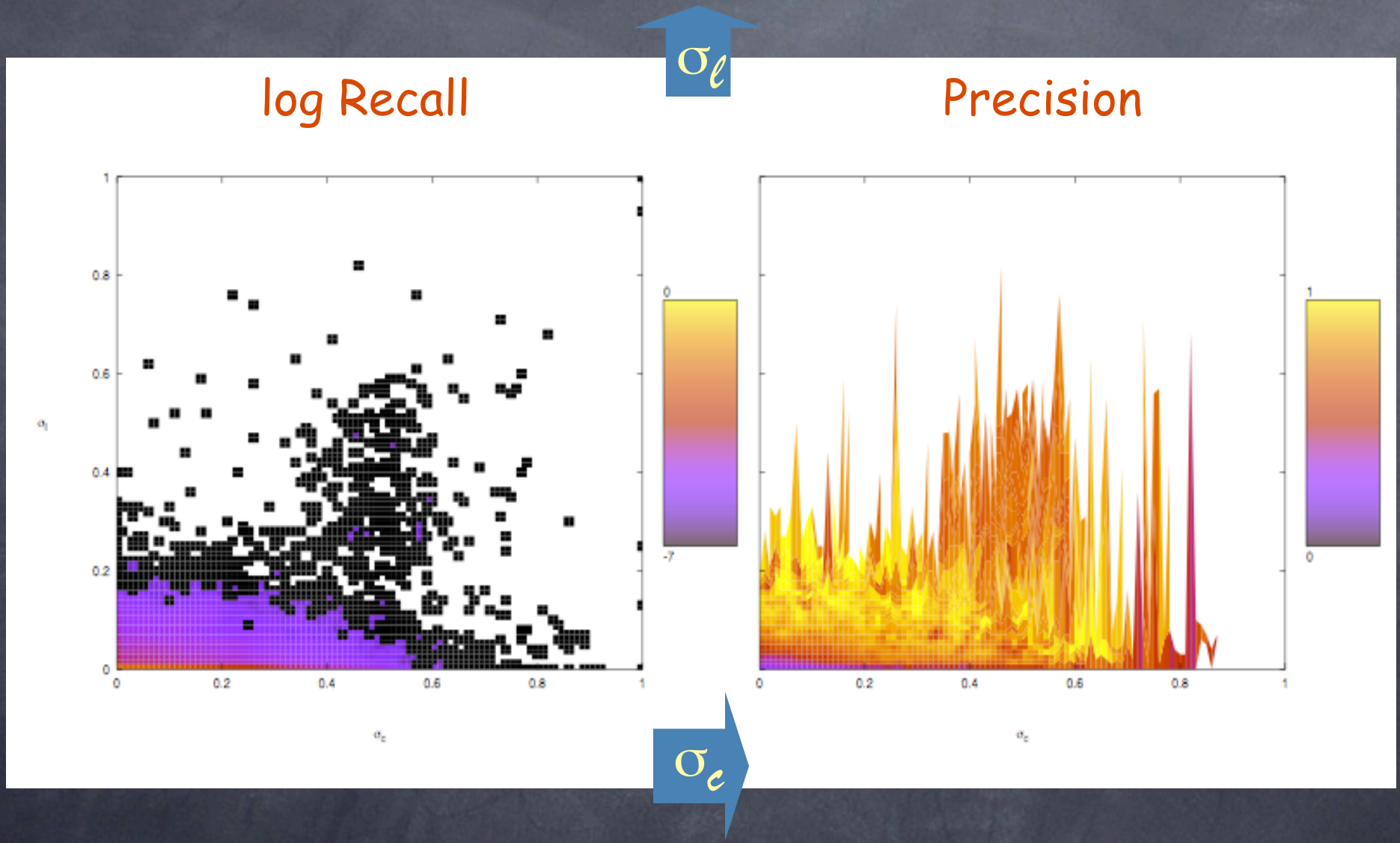
$$P(s_c, s_l) = \frac{\sum_{\{p,q:\sigma_c = s_c, \sigma_l = s_l\}} \sigma_s(p,q)}{\left|\{p,q : \sigma_c = s_c, \sigma_l = s_l\}\right|}$$

Averaging semantic similarity

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

Summing semantic similarity

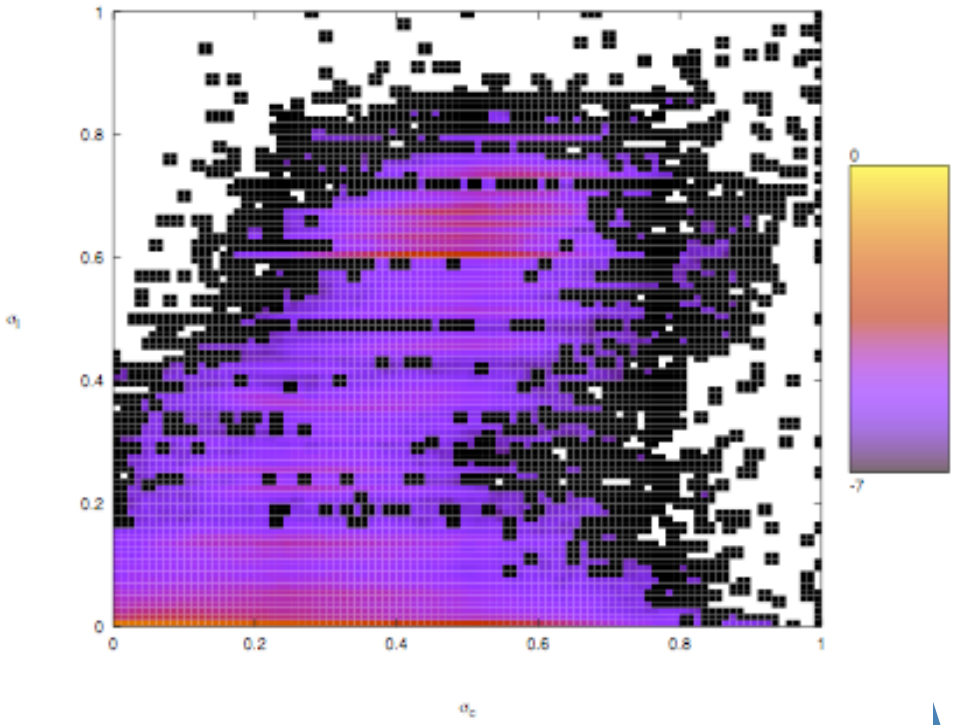
Business


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

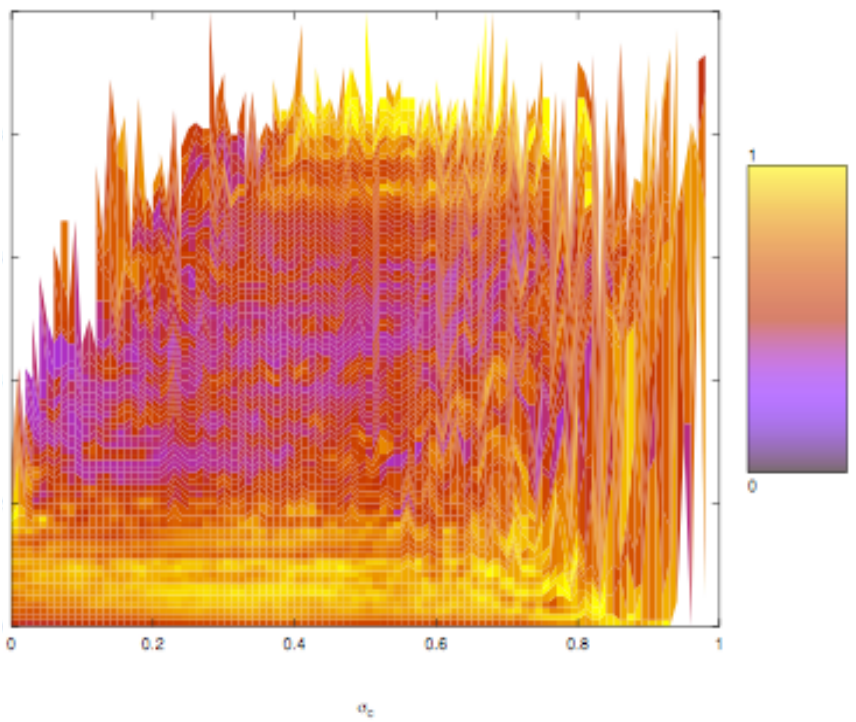
Home

σ_ℓ

log Recall

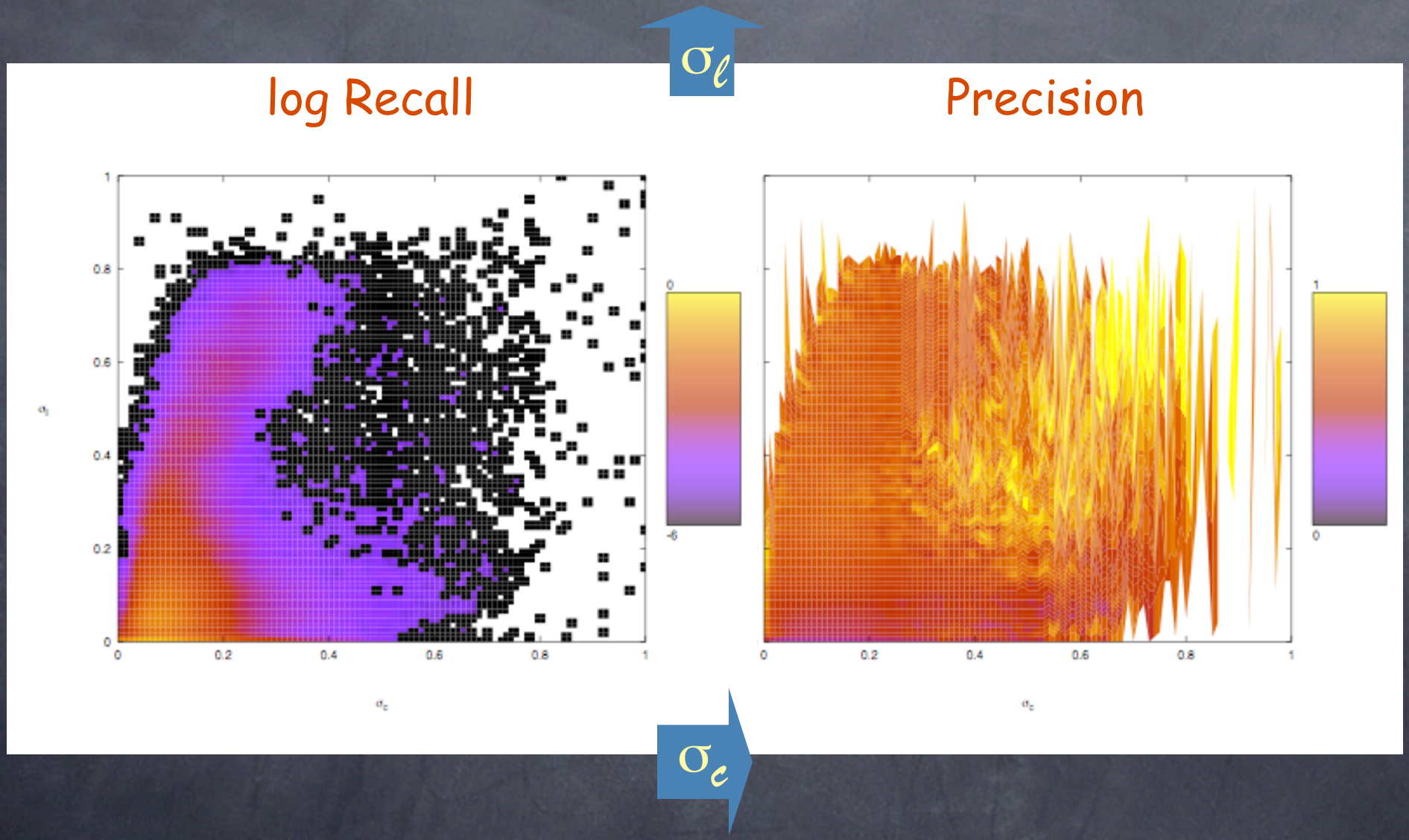


Precision



σ_c

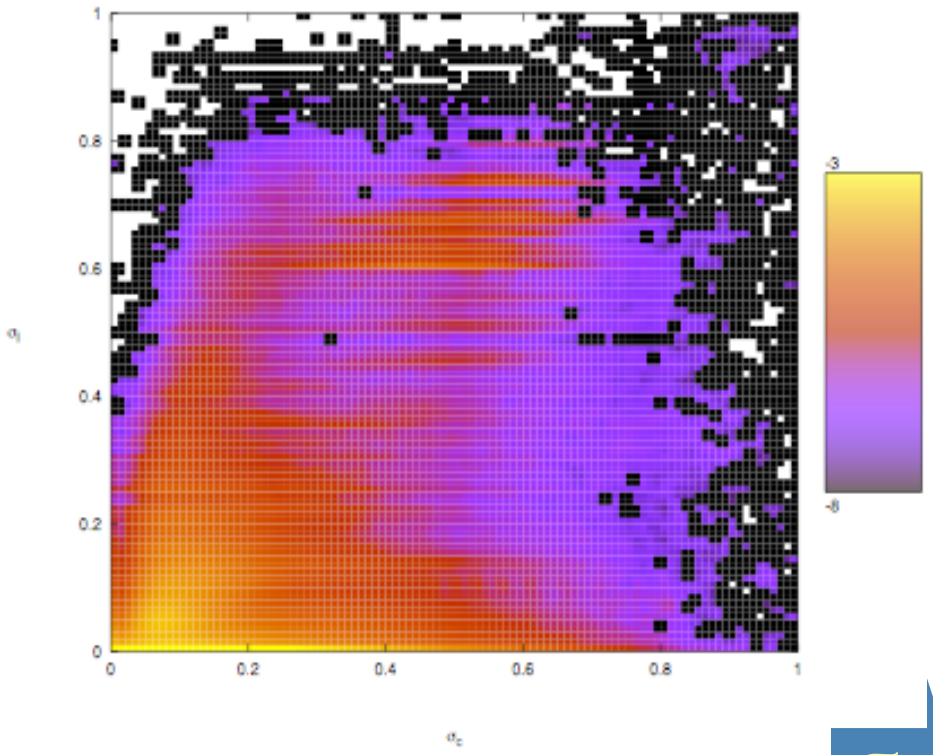
News


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

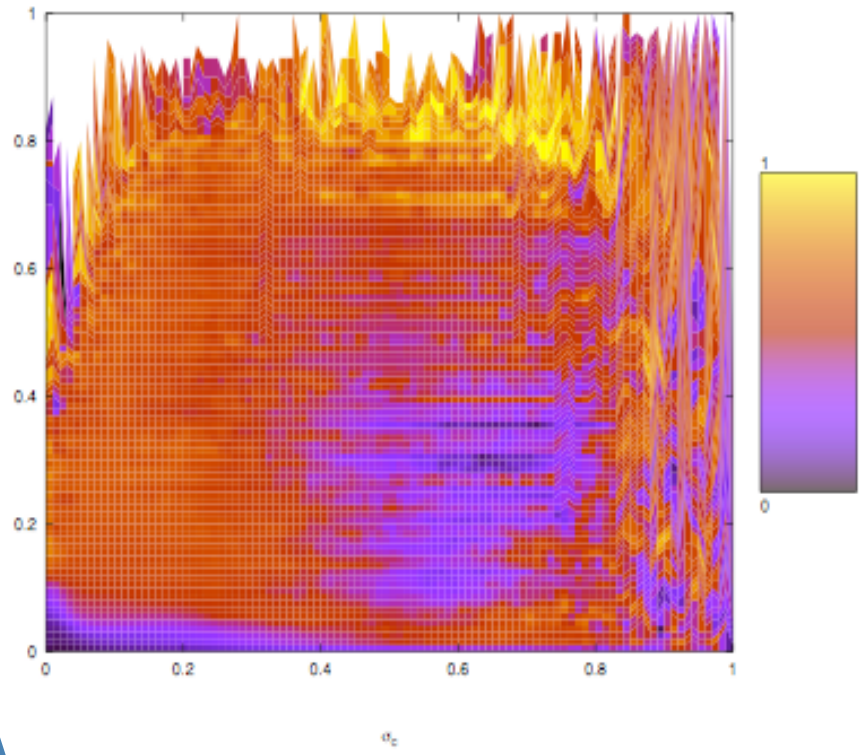
All pairs

σ_ℓ

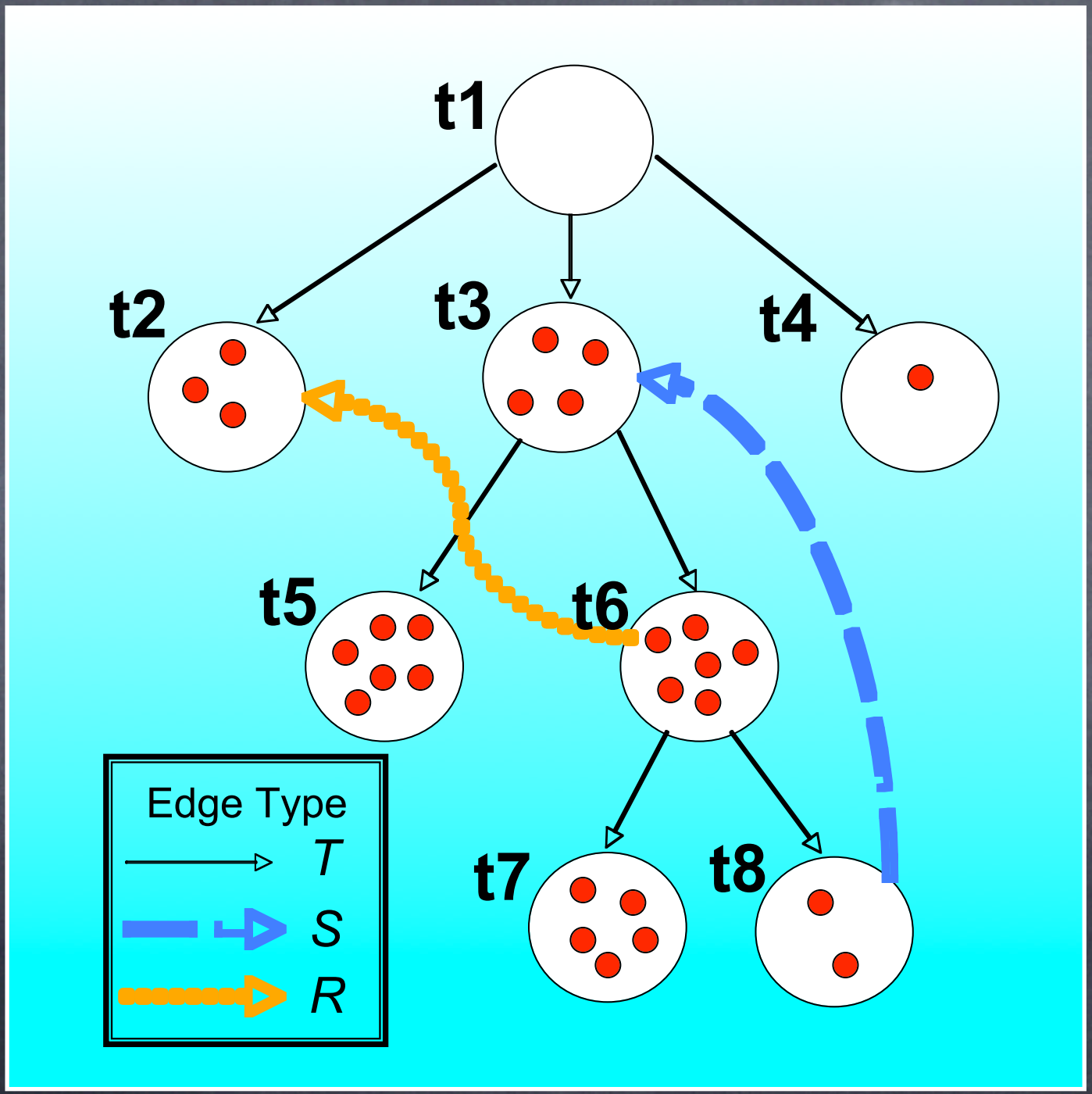
log Recall



Precision



σ_ℓ


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

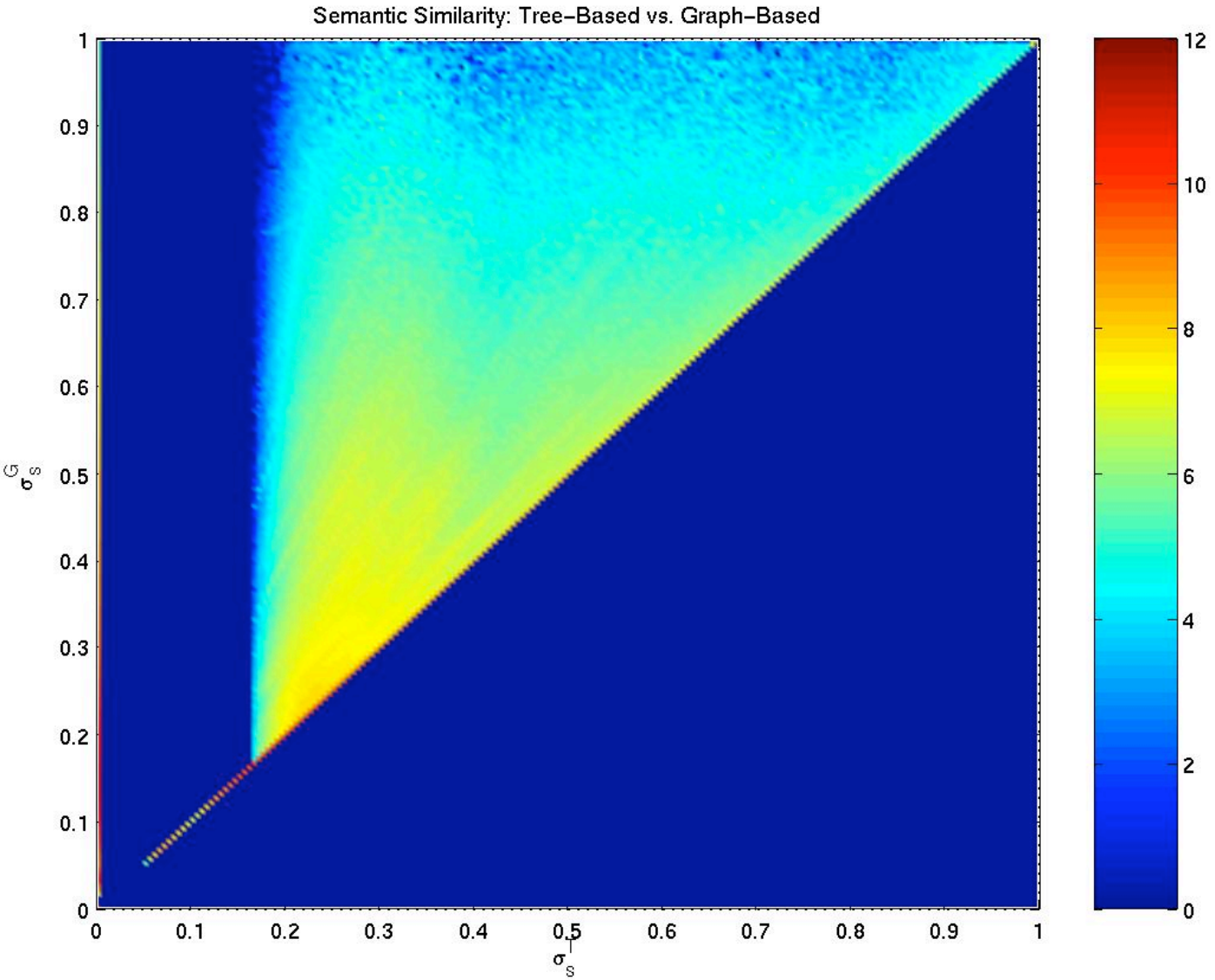
Better semantic similarity measure

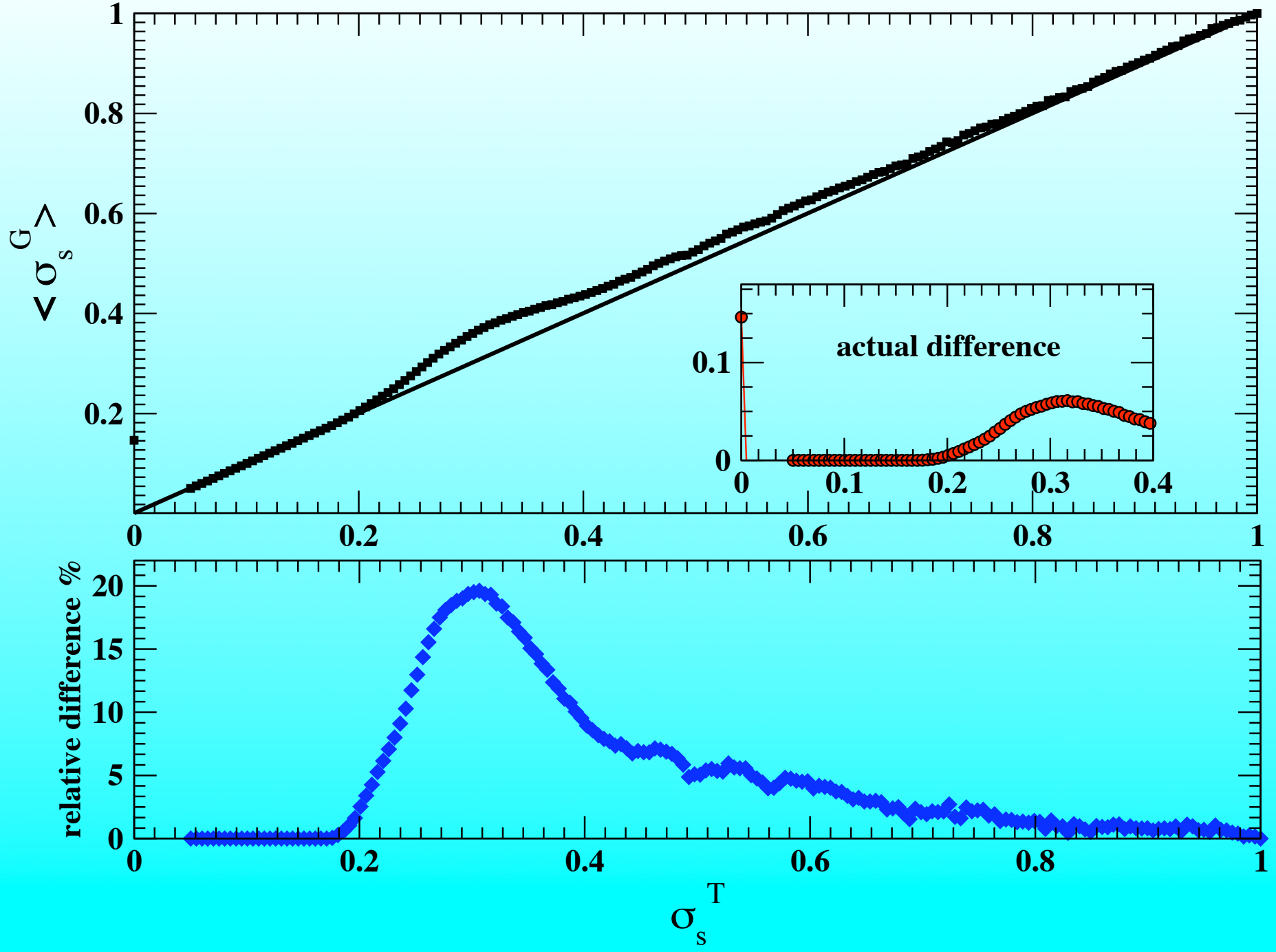
- Work w/Ana Maguitman, Heather Roinestad, Alex Vespignani
  - Include cross-links (symbolic) and see-also links (related)
  - Transitive closure of topic graph
  - Compute entropy based on fuzzy membership matrix

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

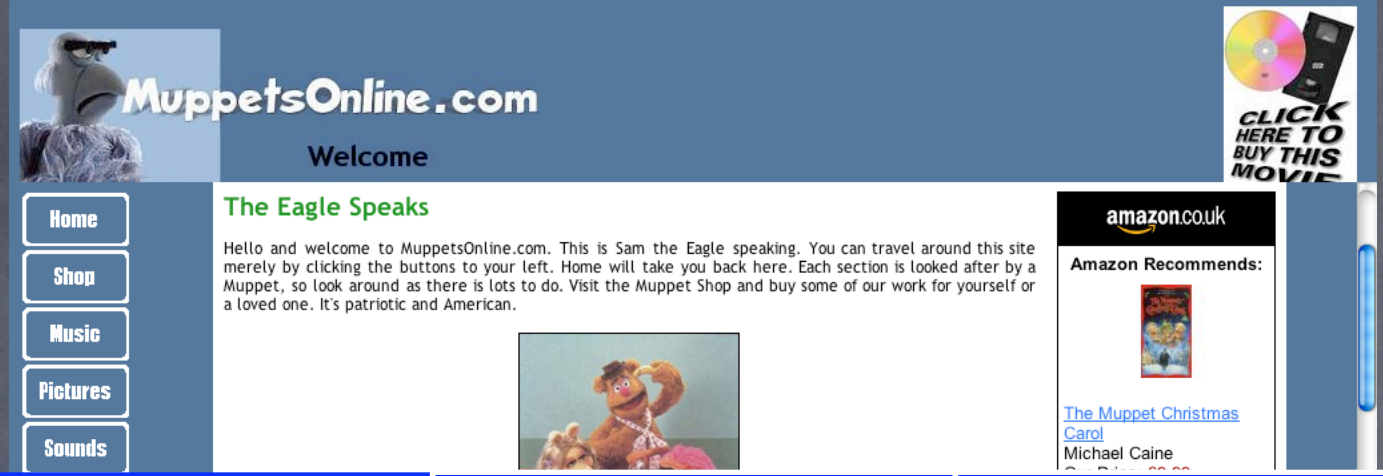
$$\sigma_s(t_1, t_2) = \max_k \frac{2 \cdot \min(\mathrm{W}_{k1}, \mathrm{W}_{k2}) \cdot \log \Pr[t_k]}{\log(\Pr[t_1|t_k] \cdot \Pr[t_k]) + \log(\Pr[t_2|t_k] \cdot \Pr[t_k])}$$

Differences

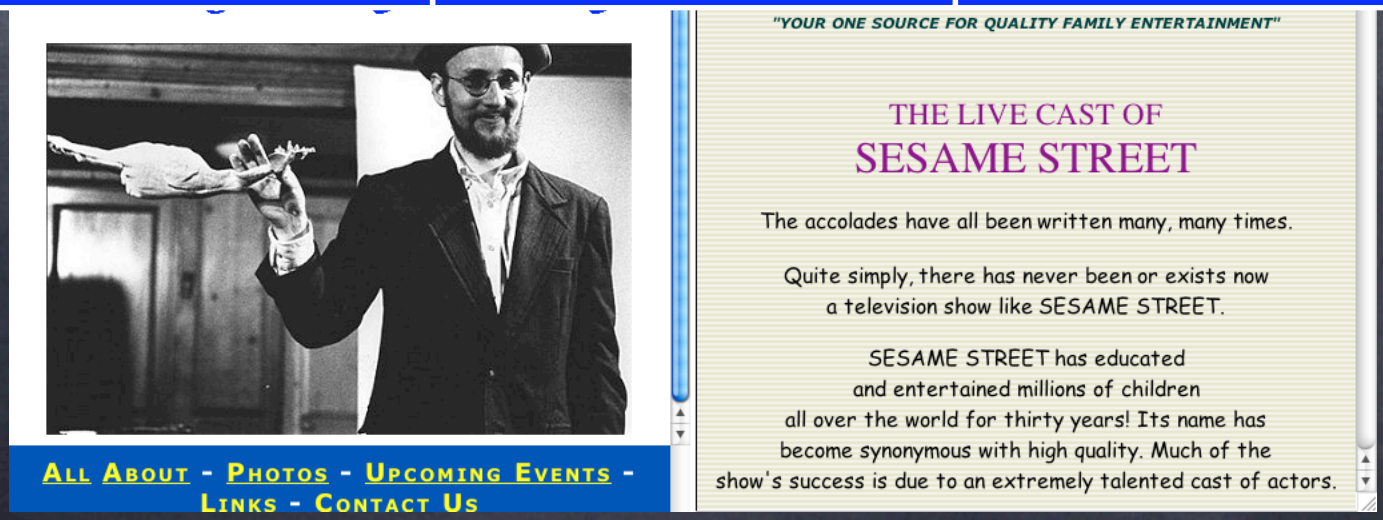



1
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1
$\langle \sigma_s^G \rangle$
actual difference
0.1
0
0.1
0.2
0.3
0.4
relative difference %
20
15
10
5
0
σ_s^T

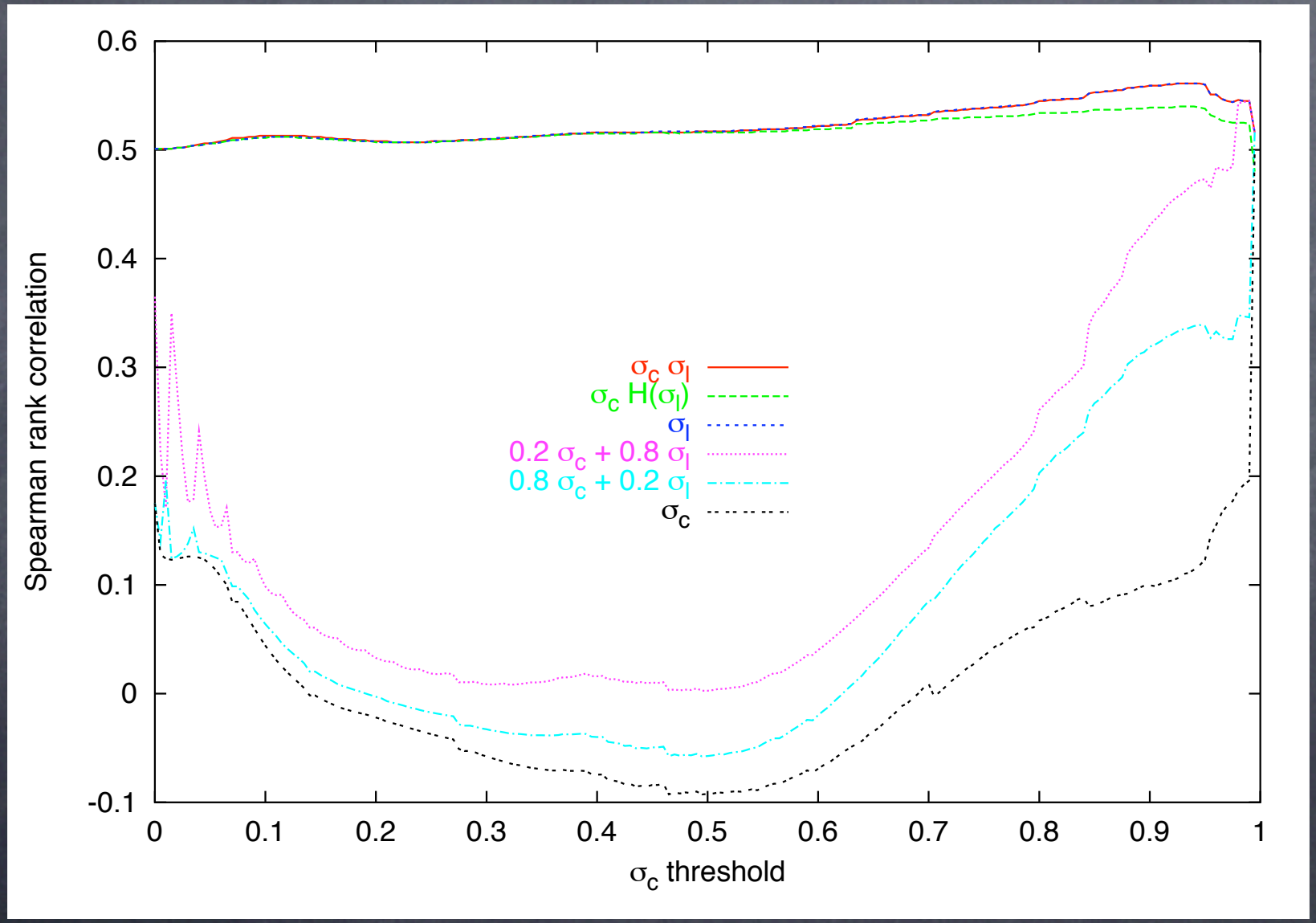
Web Page Relatedness



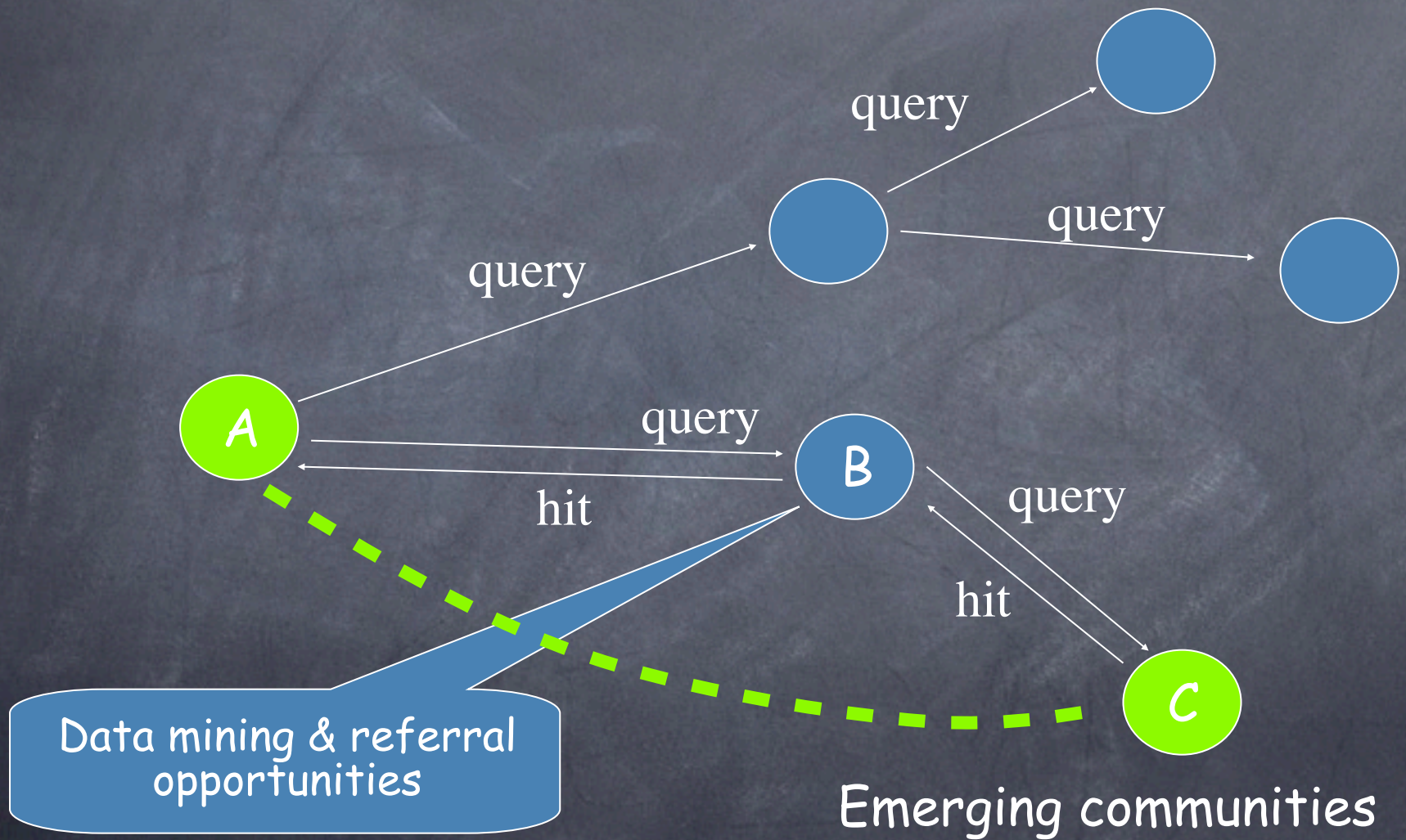
	mean	stderr
tree	5.7%	0.8%
graph	84.7%	1.8%

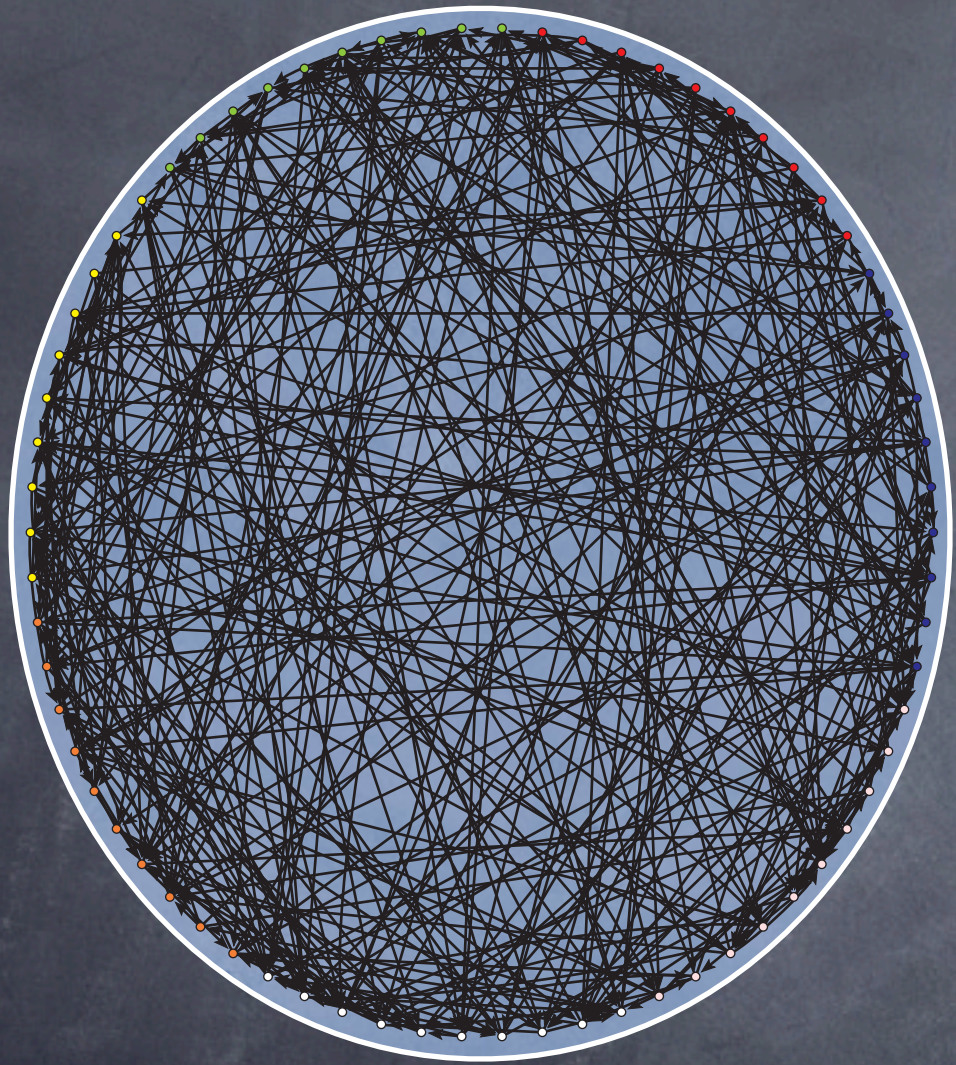


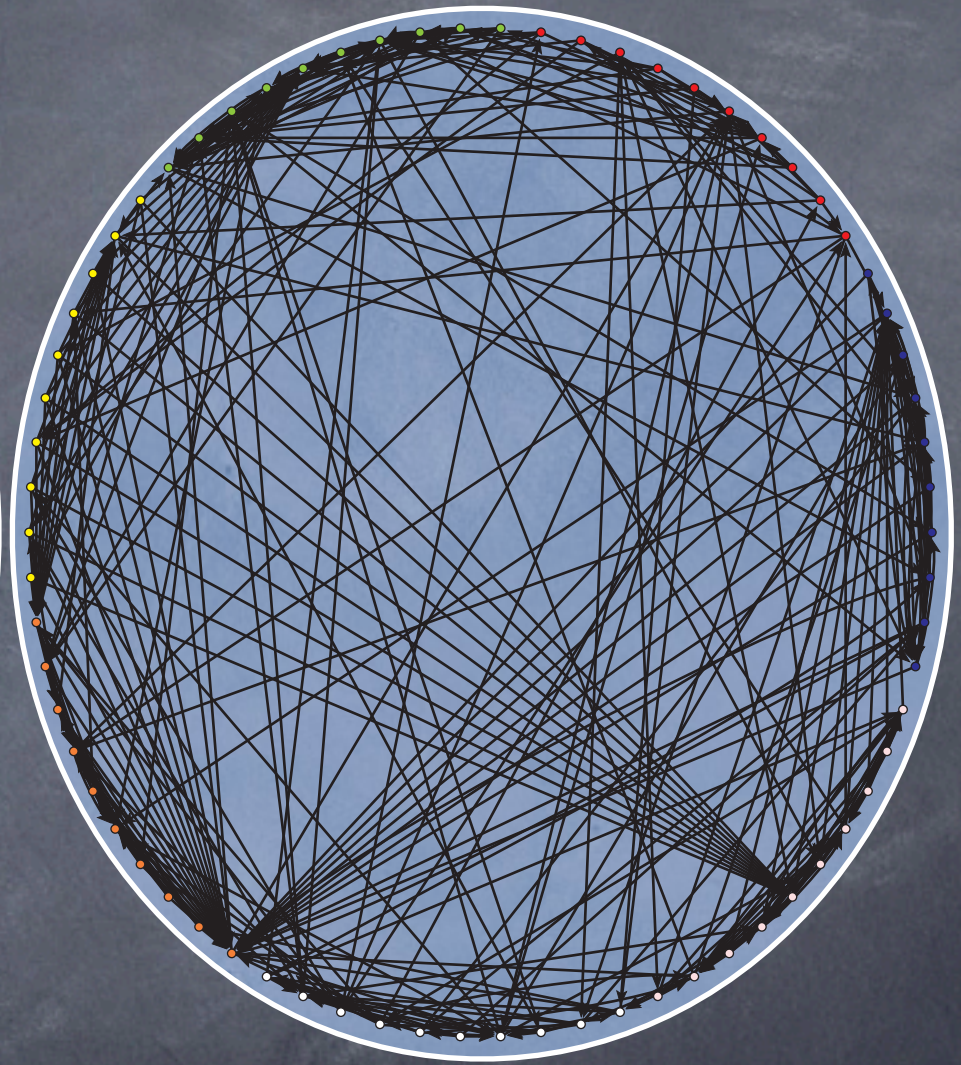
Combining content & links



6S: peer distributed crawling and collaborative searching

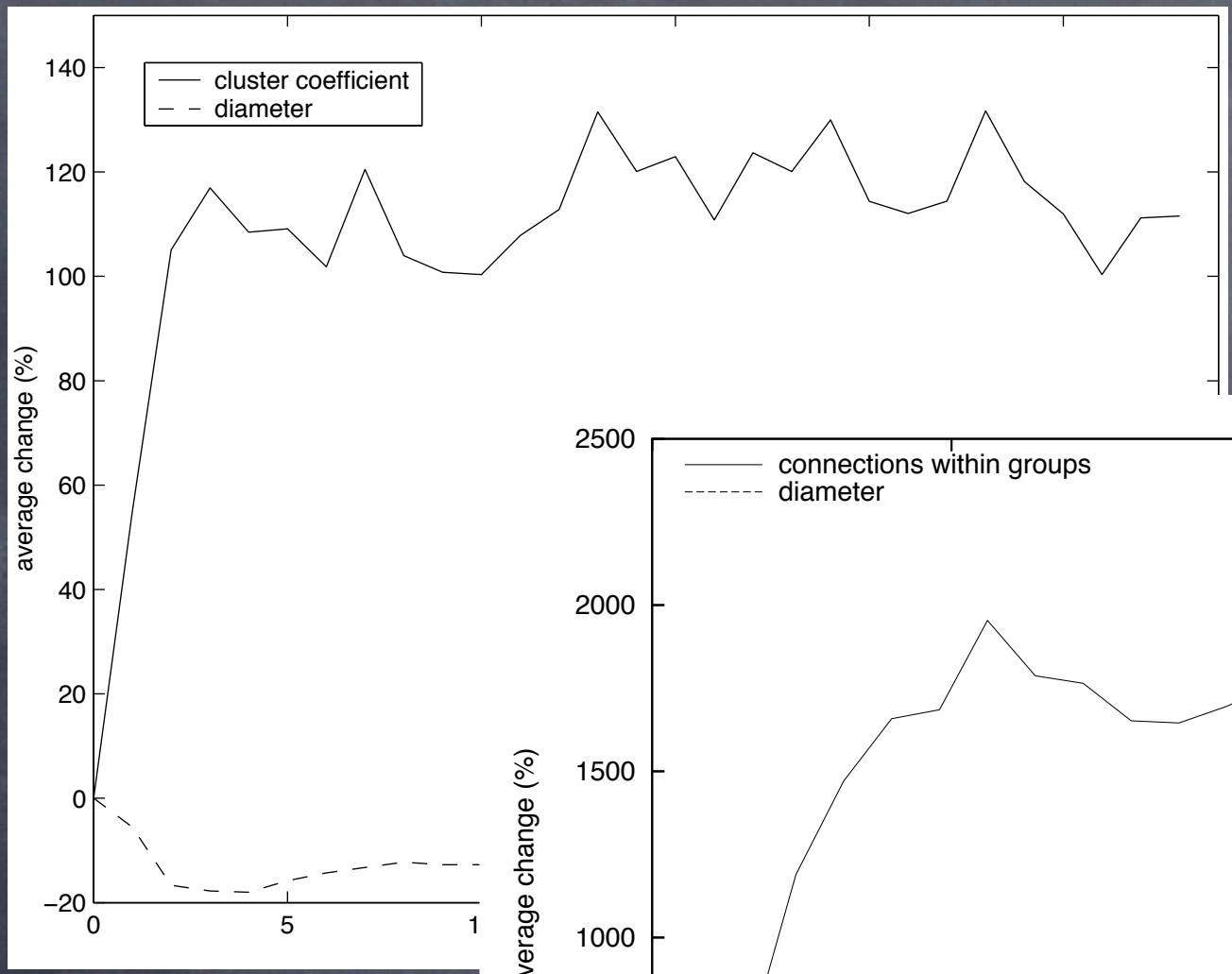




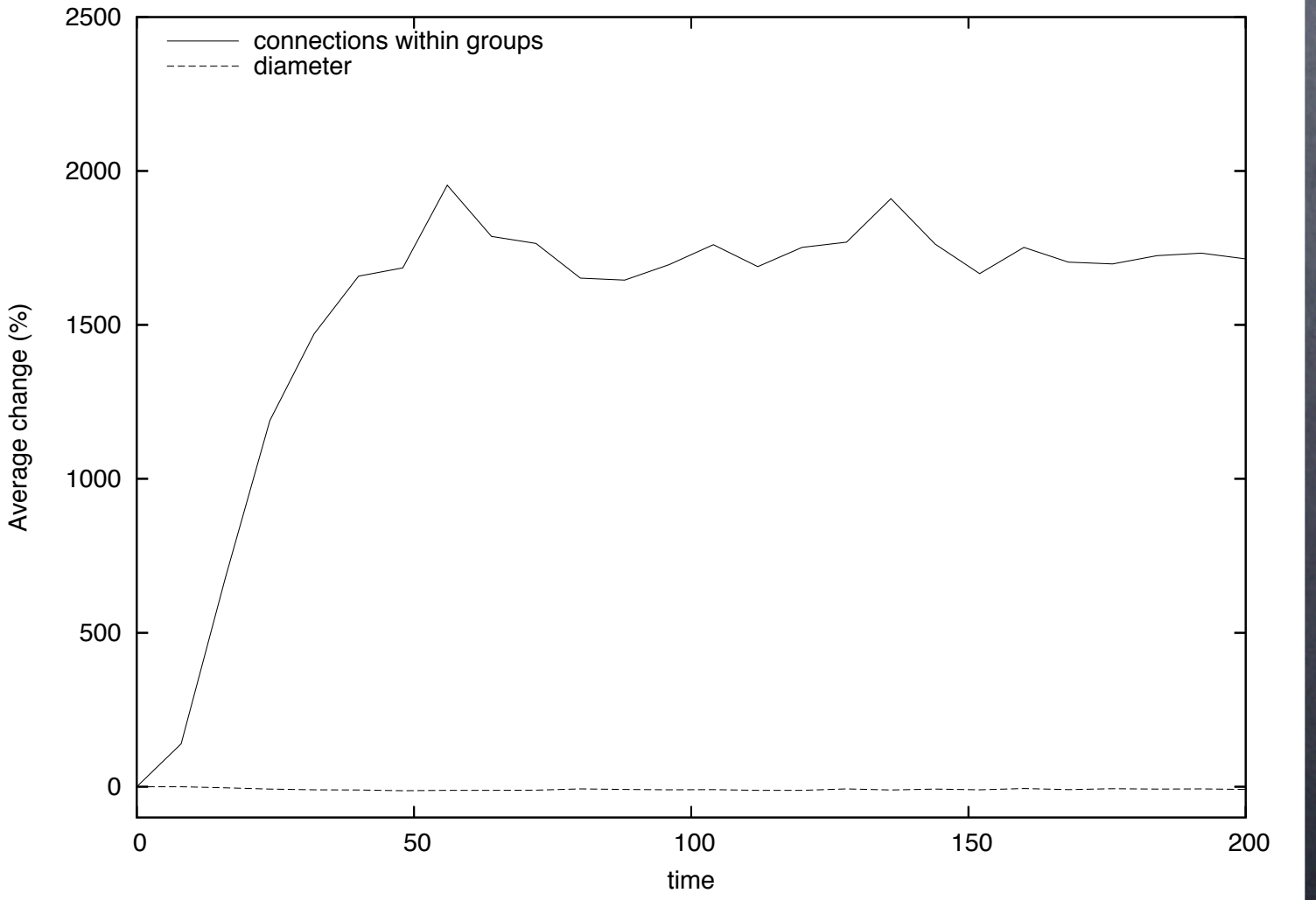


Work with Le-Shin Wu & Ruj Akavipat

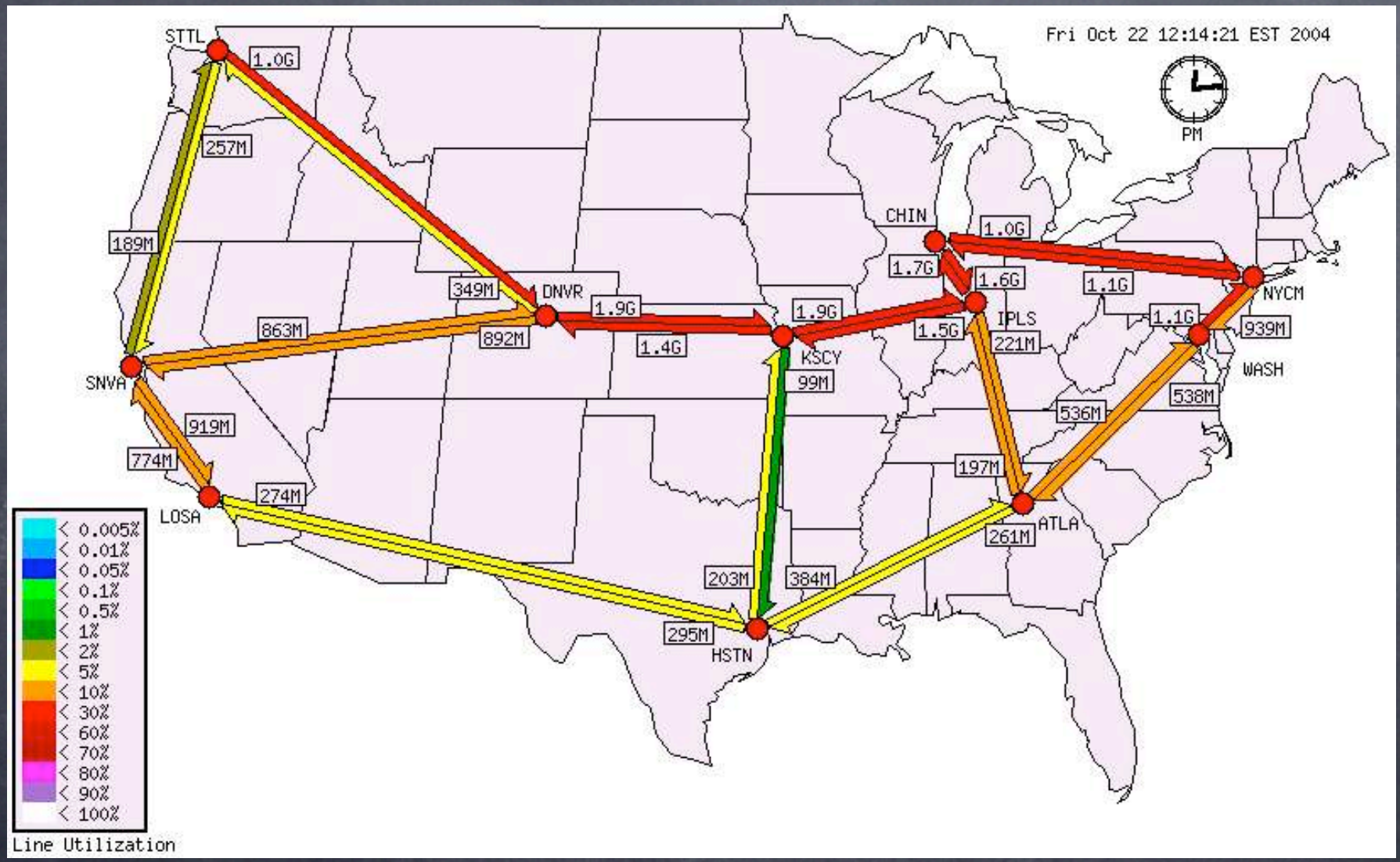
70 peers



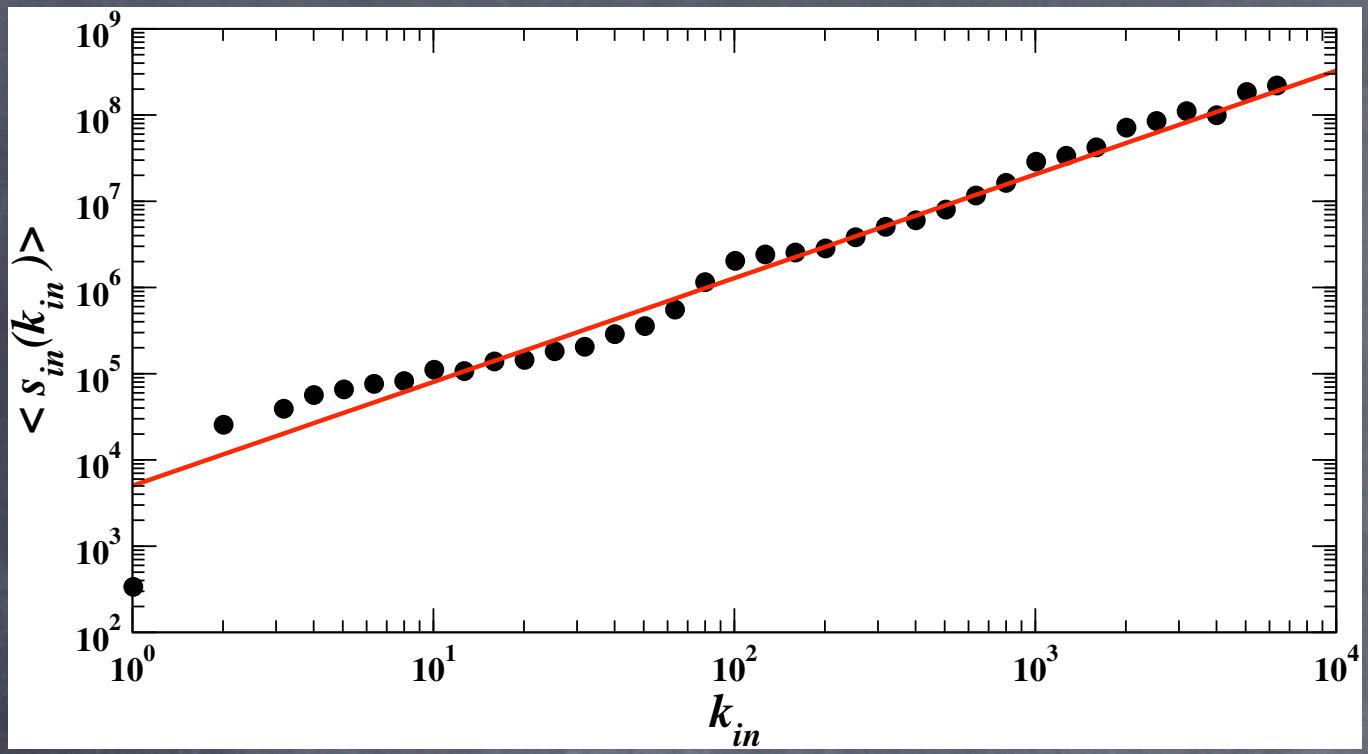
500 peers

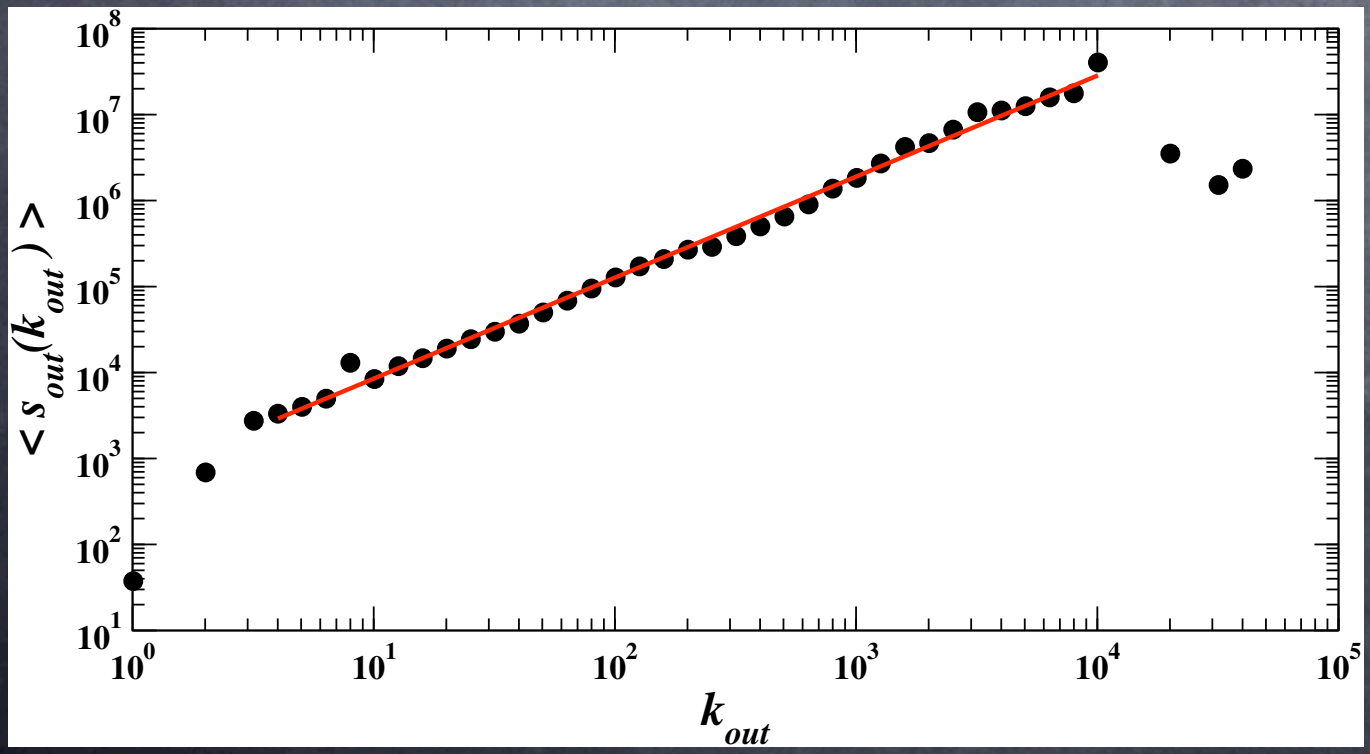


WWW traffic network on Internet2



Work with Mark Meiss & Alex Vespignani

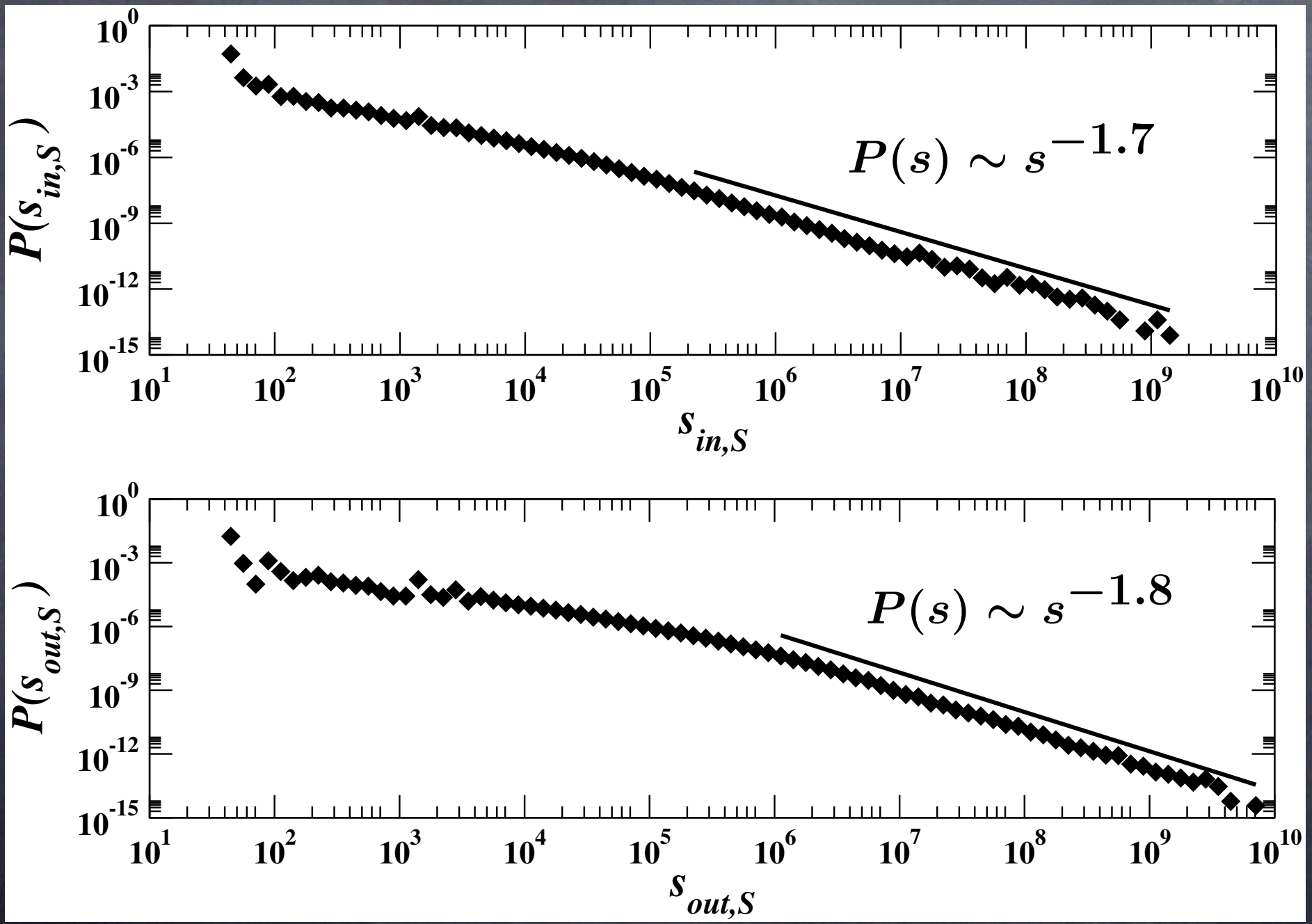




Web clients: super-linear growth of traffic handled as a function of number of connections

$$s \sim k^{1.2}$$

Web servers: no typical traffic (diverging average)



Thank you!
Questions?

http://informatics.indiana.edu/fil

Research supported by NSF
CAREER Award IIS-0348940