

Web Networks

Filippo Menczer

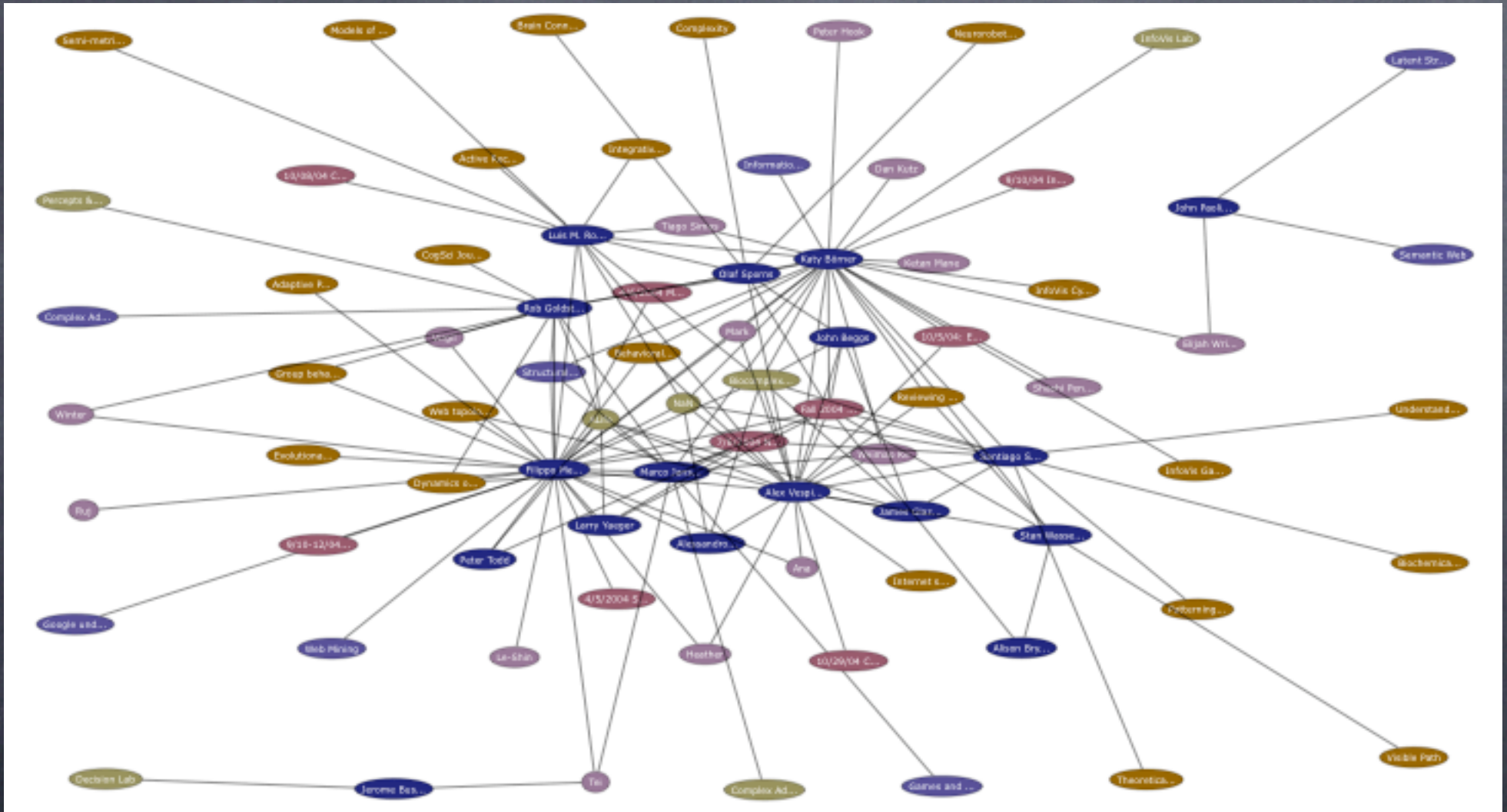
Department of Computer Science
School of Informatics

Indiana University



Research supported by NSF
CAREER Award IIS-0348940

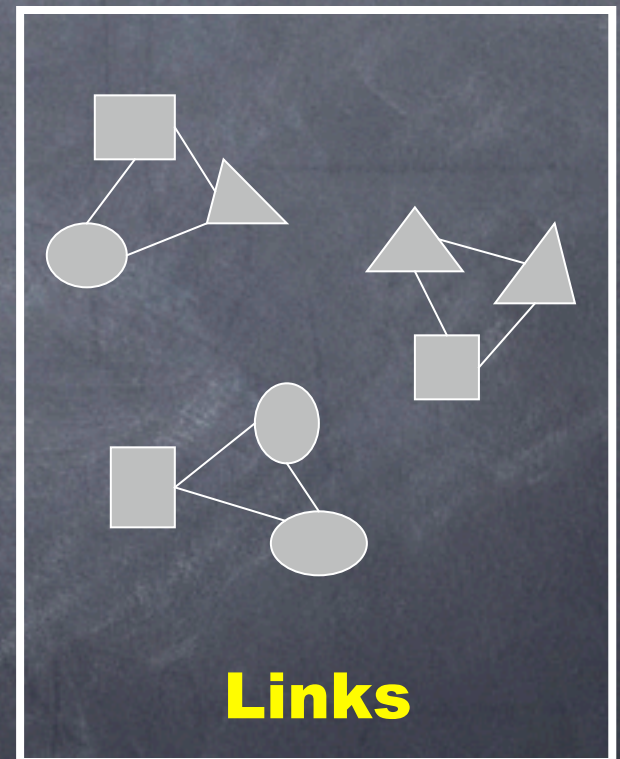
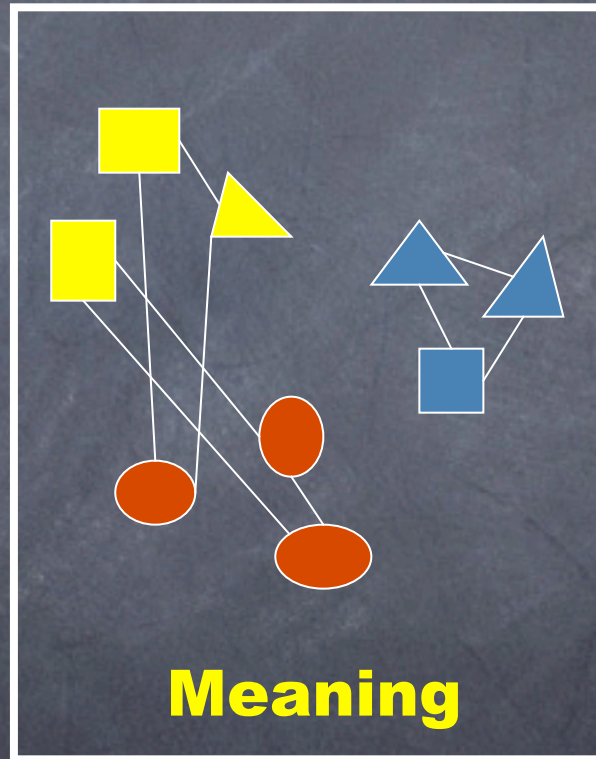
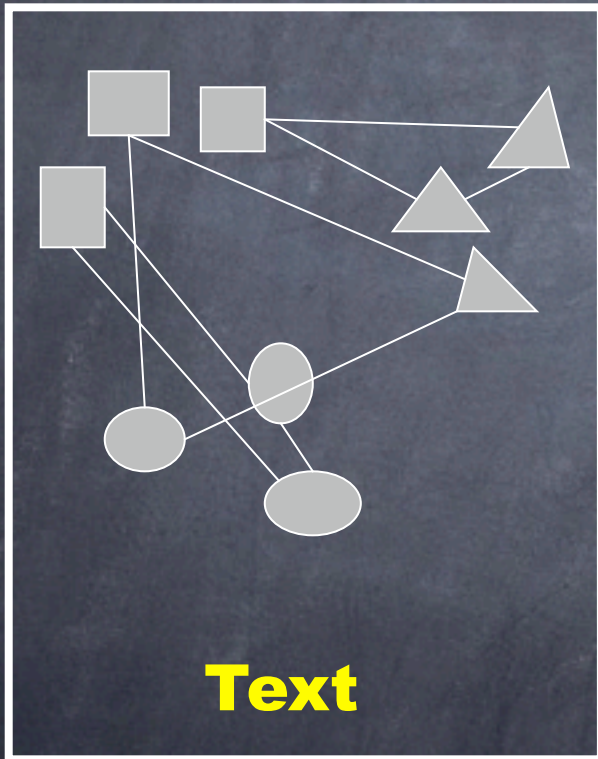
csn.indiana.edu



Outline

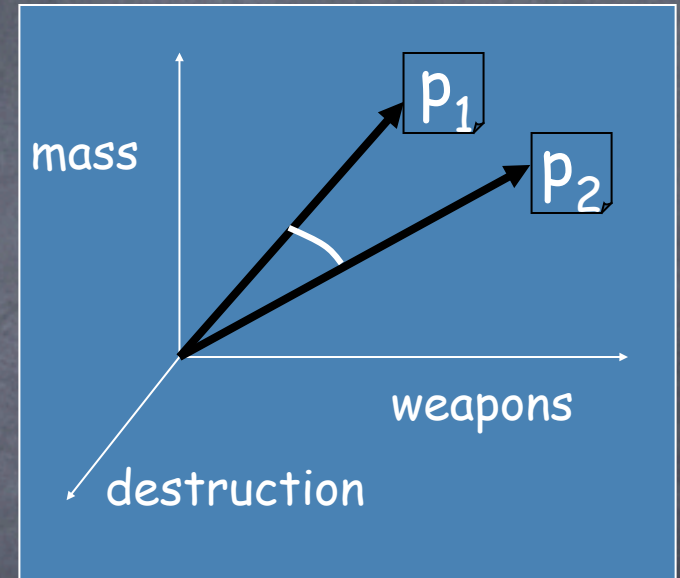
- ✓ Link network
- ✓ Lexical network
- ✓ Growth models
- ⦿ Semantic network
- ⦿ Peer search network
- ⦿ Traffic network

Three network topologies

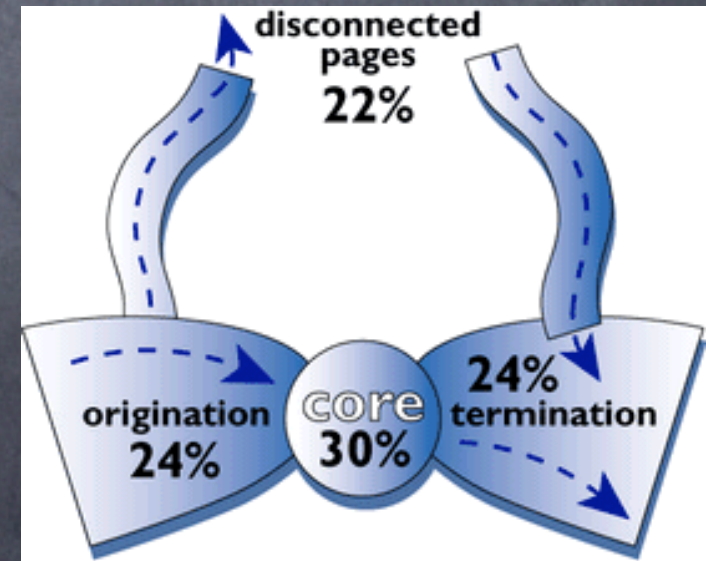
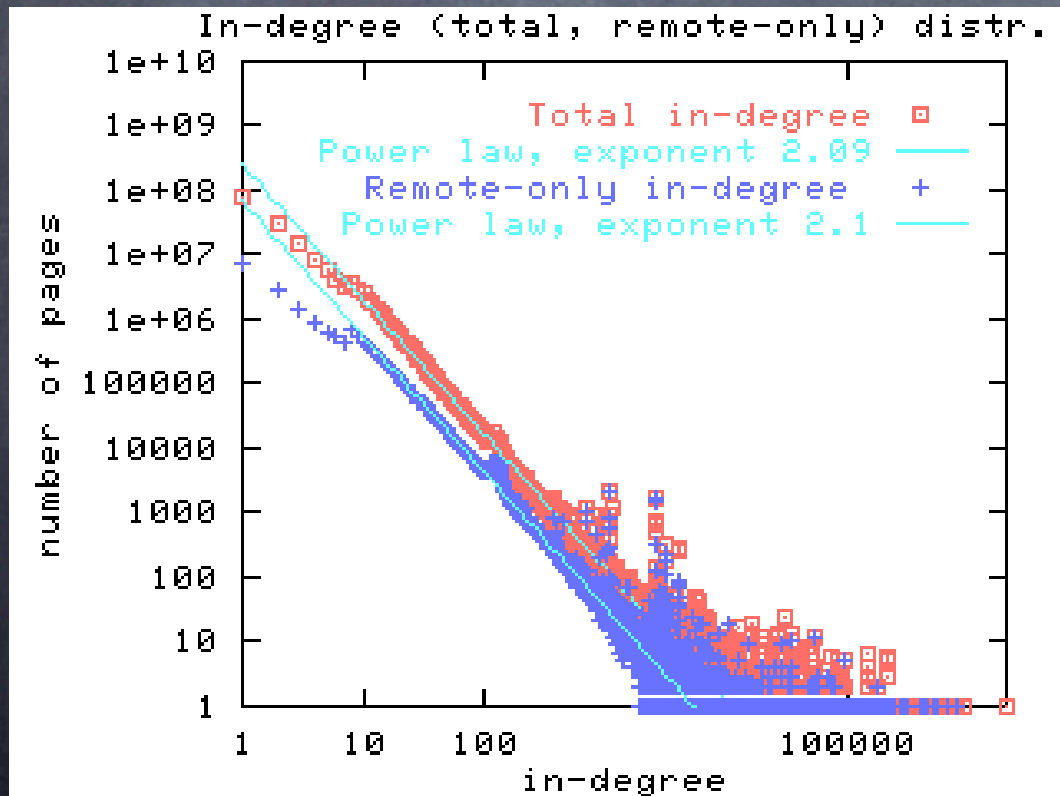
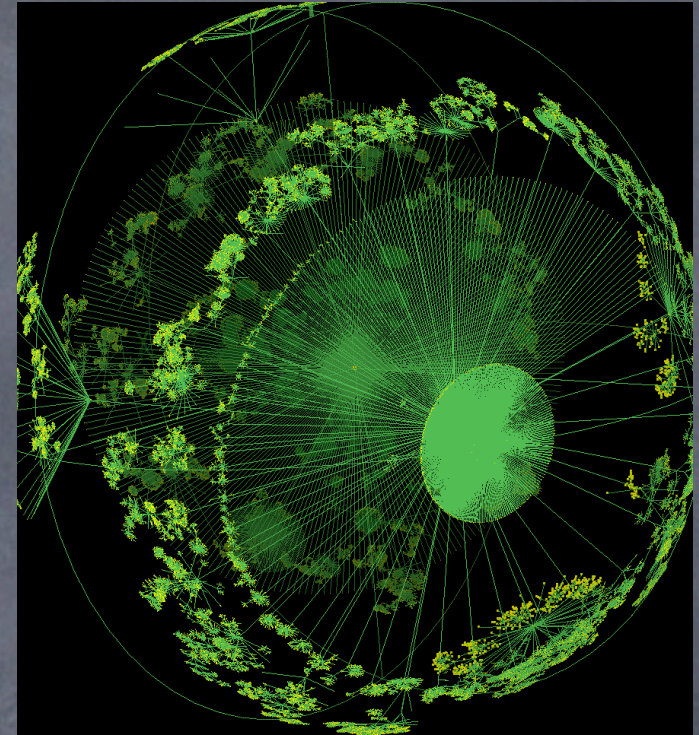
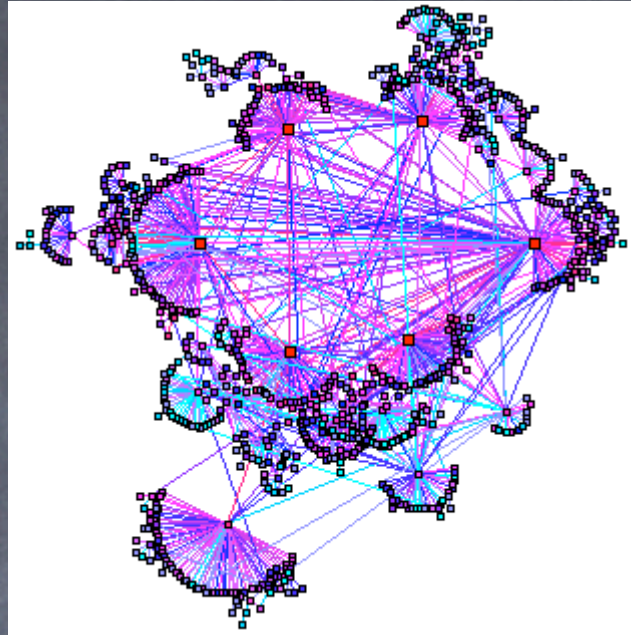


The Web as a text corpus

- 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

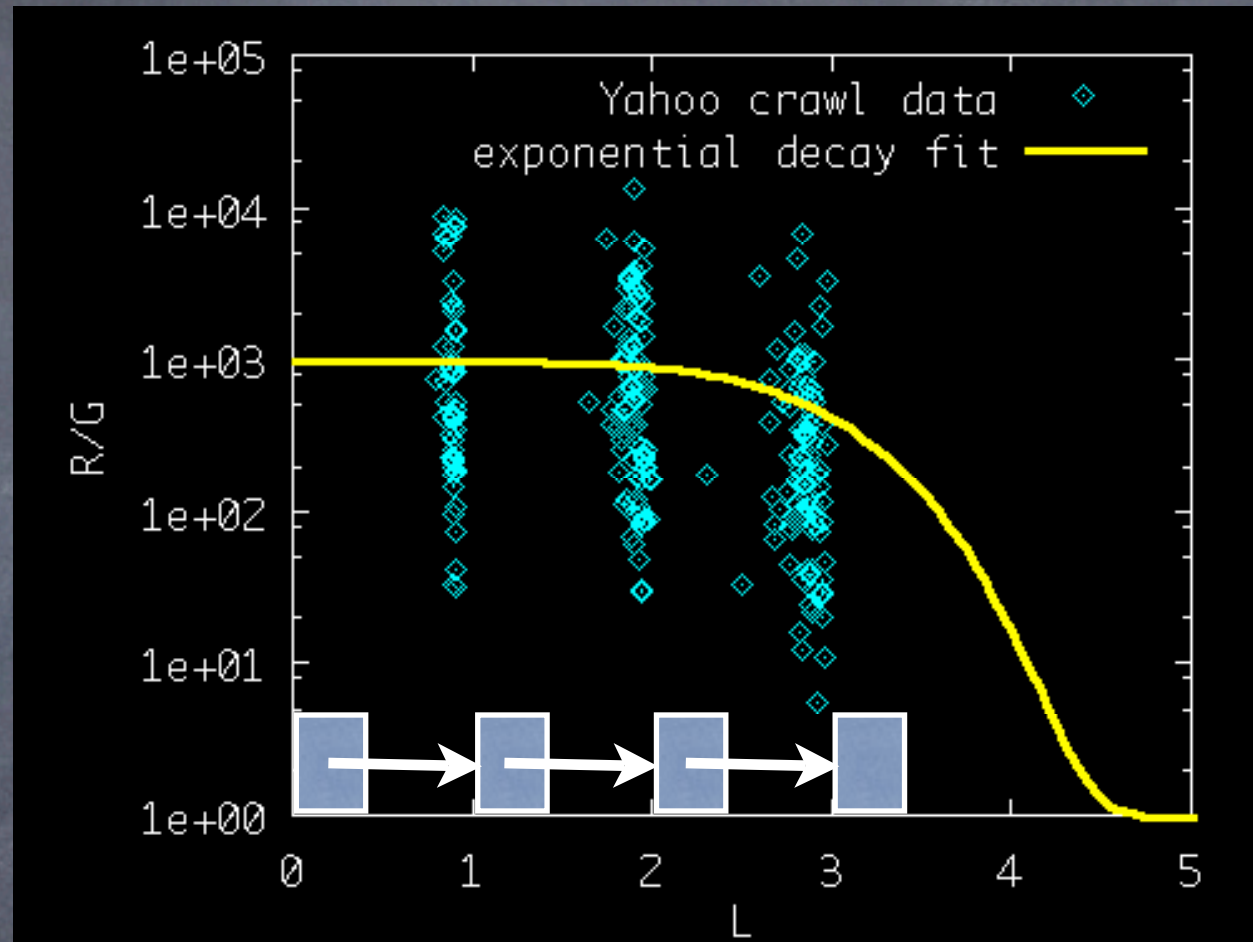
- Link-cluster conjecture** (Menczer 1997)

- 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

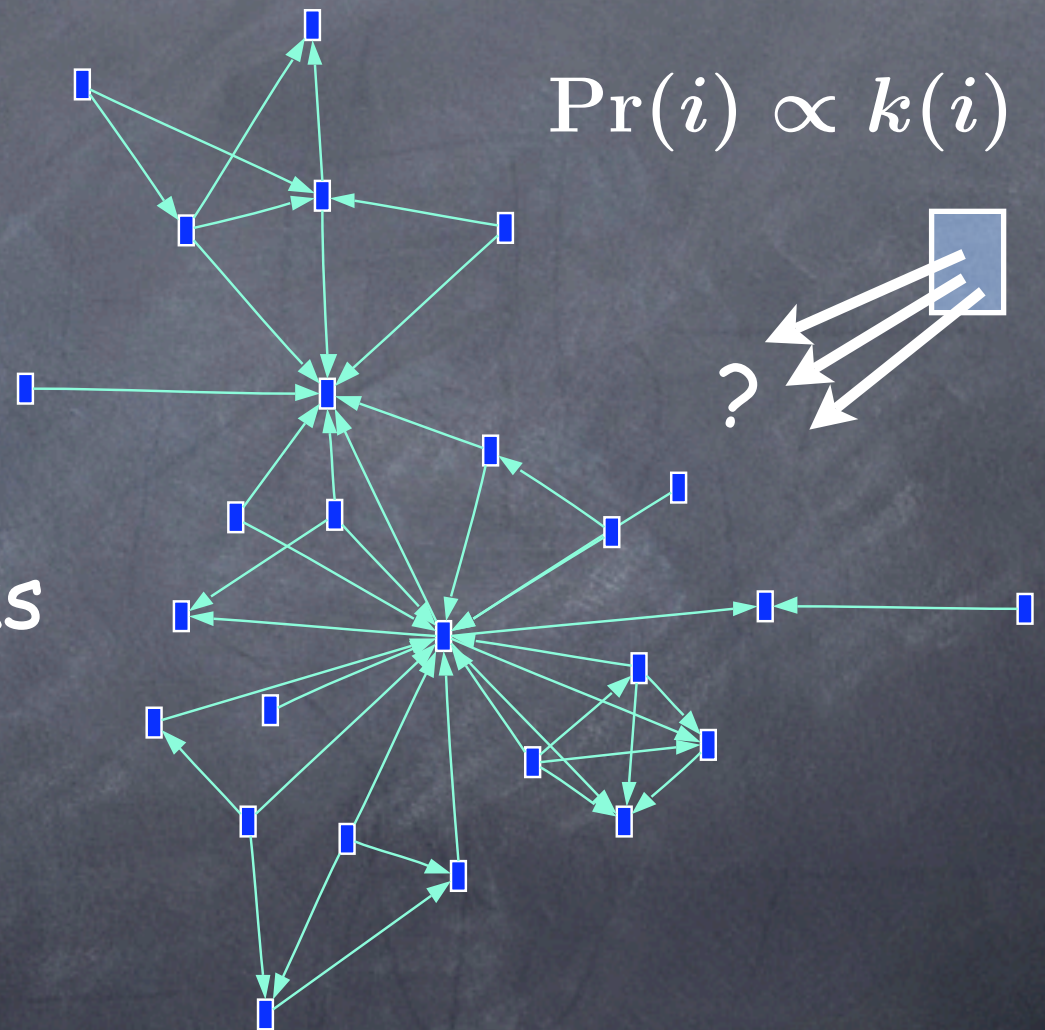
“BA” model

(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

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



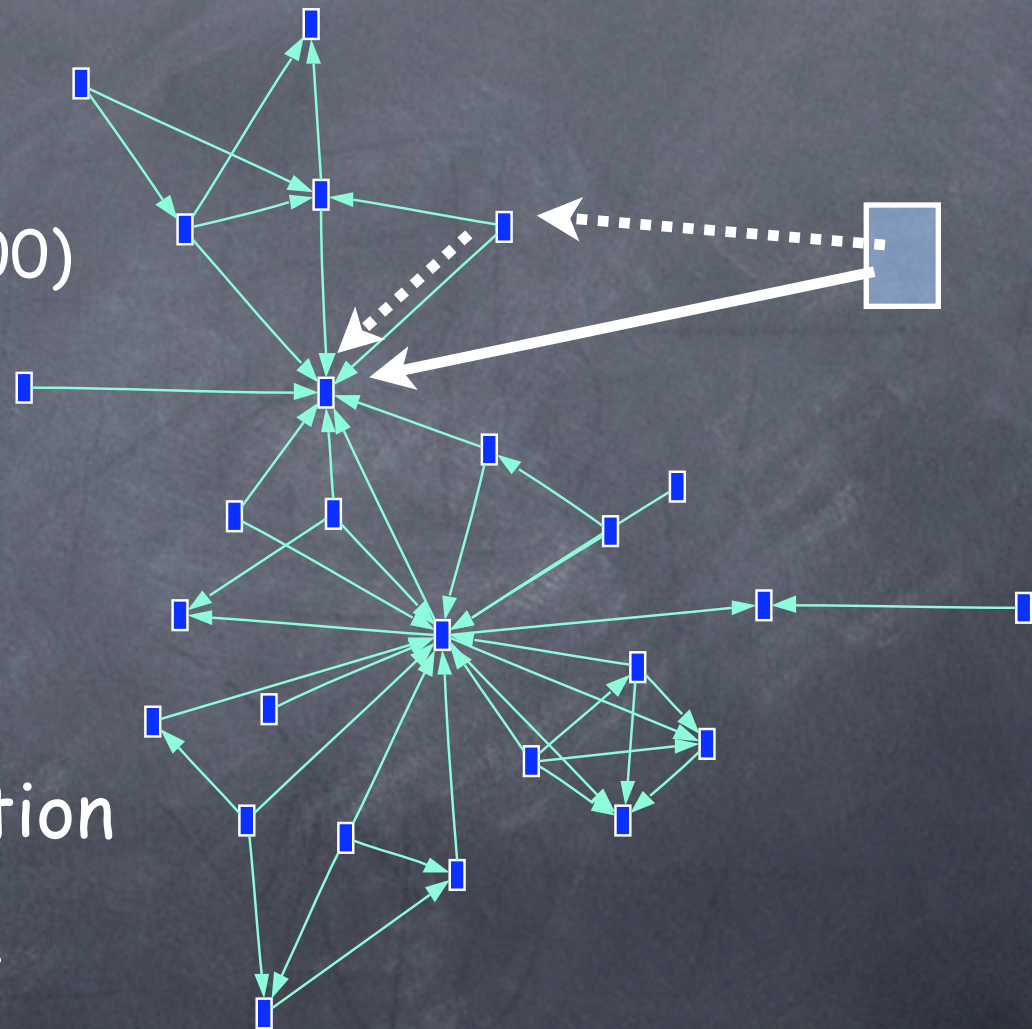
Other growth models

Web copying

(Kleinberg, Kumar & al 1999, 2000)

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

- same indegree distribution
- no need to know degree



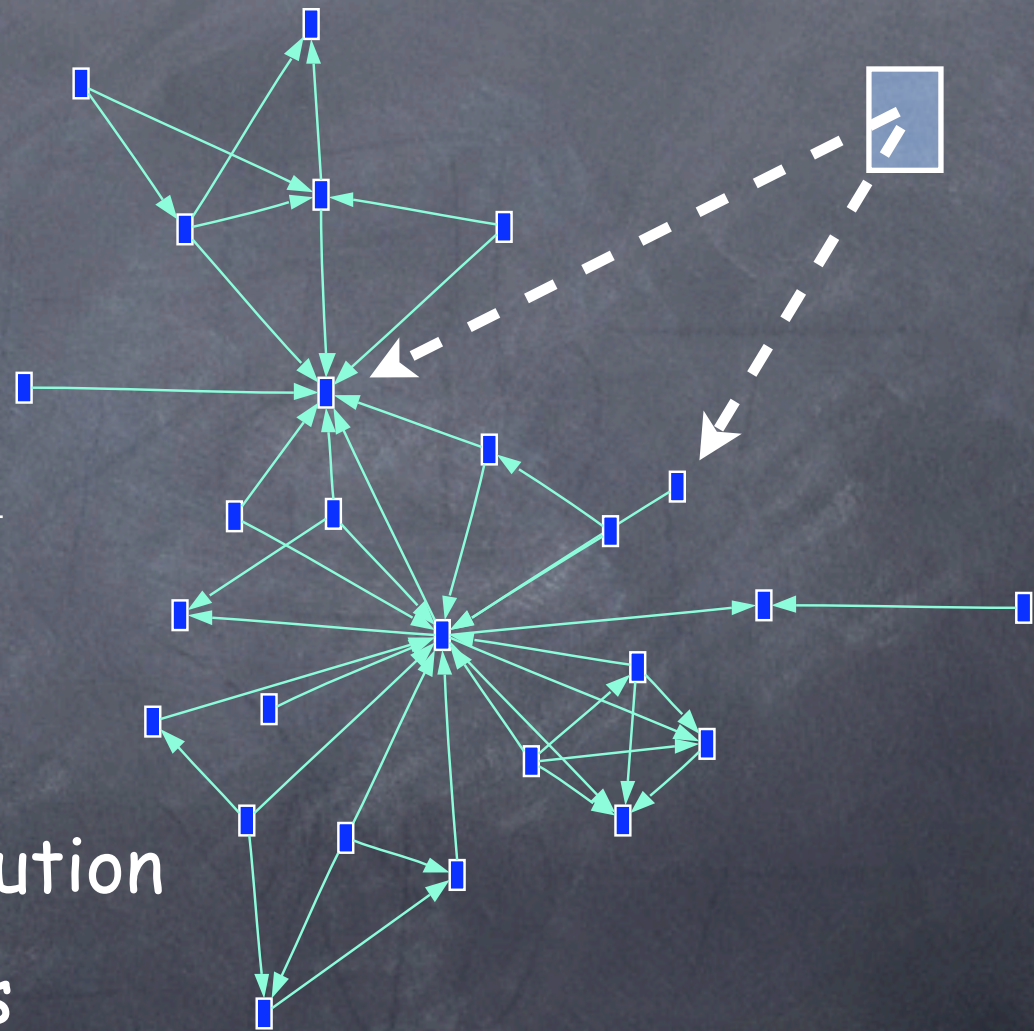
Other growth models

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



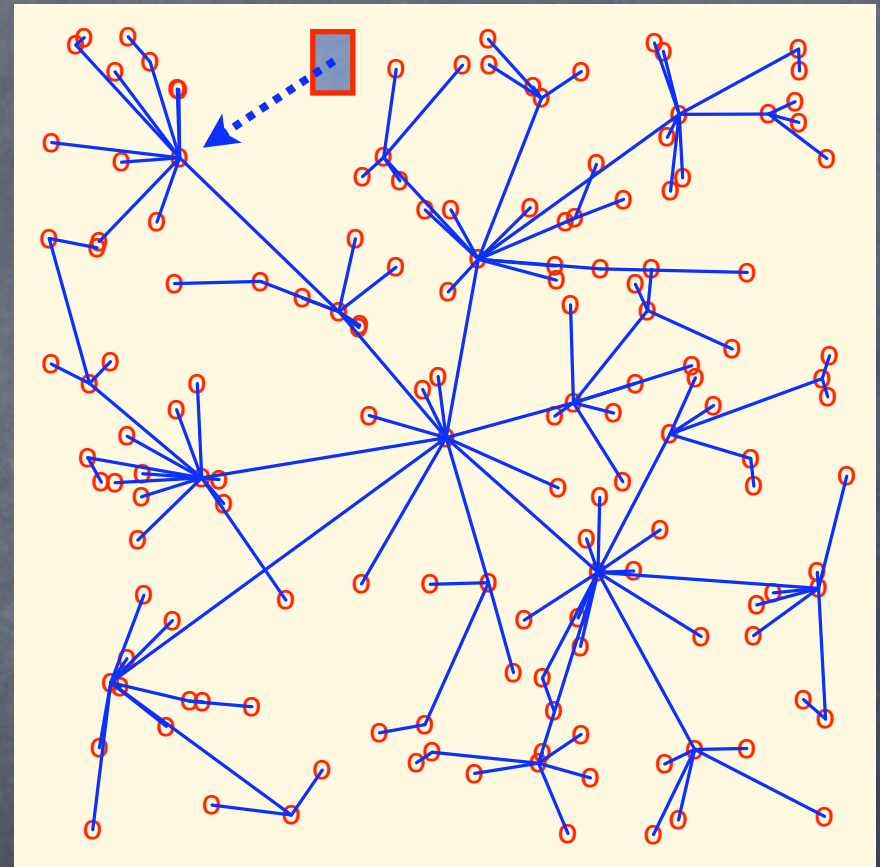
Other growth models

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

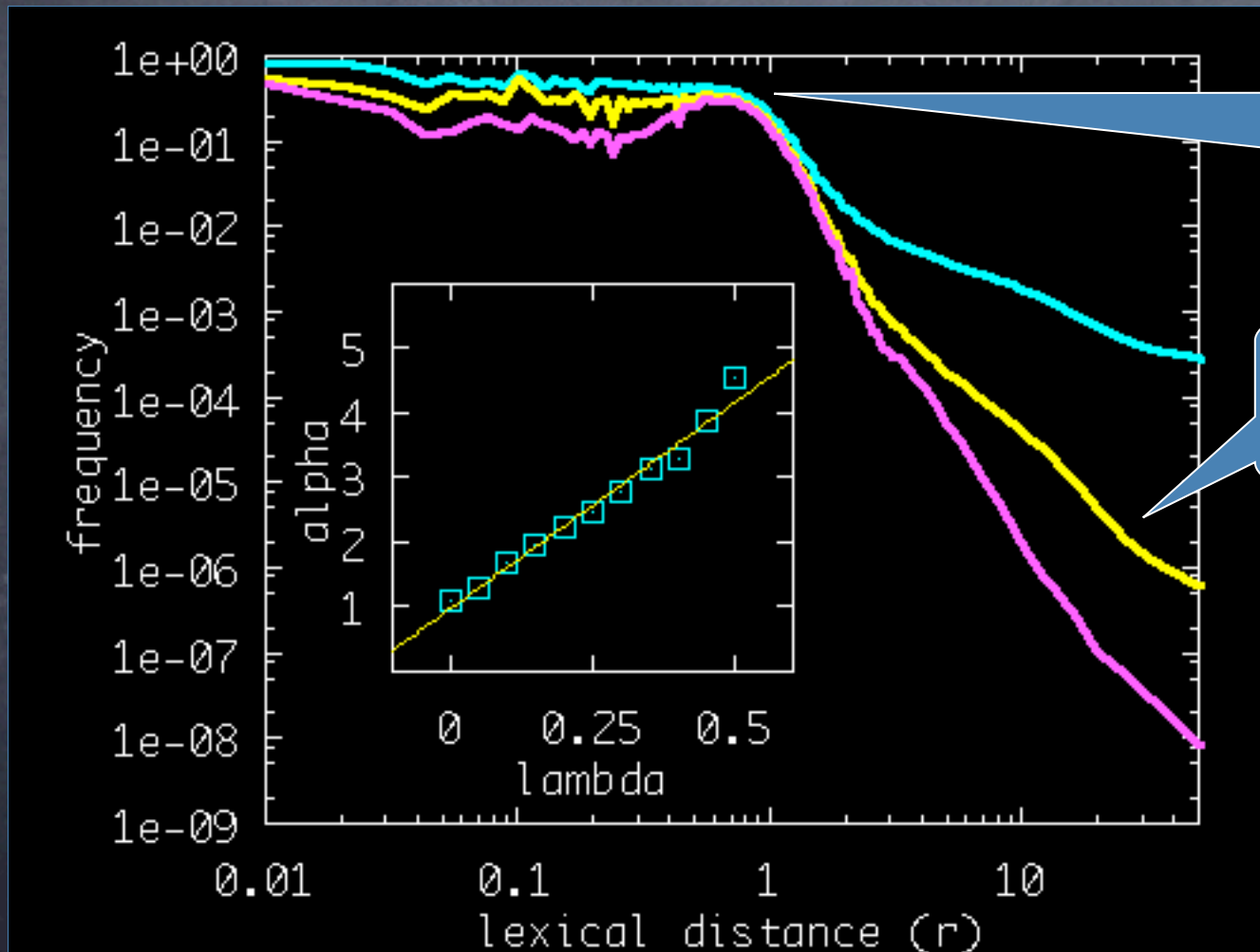


What about content?

Link probability vs lexical distance

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

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



Phase
transition
 ρ^*

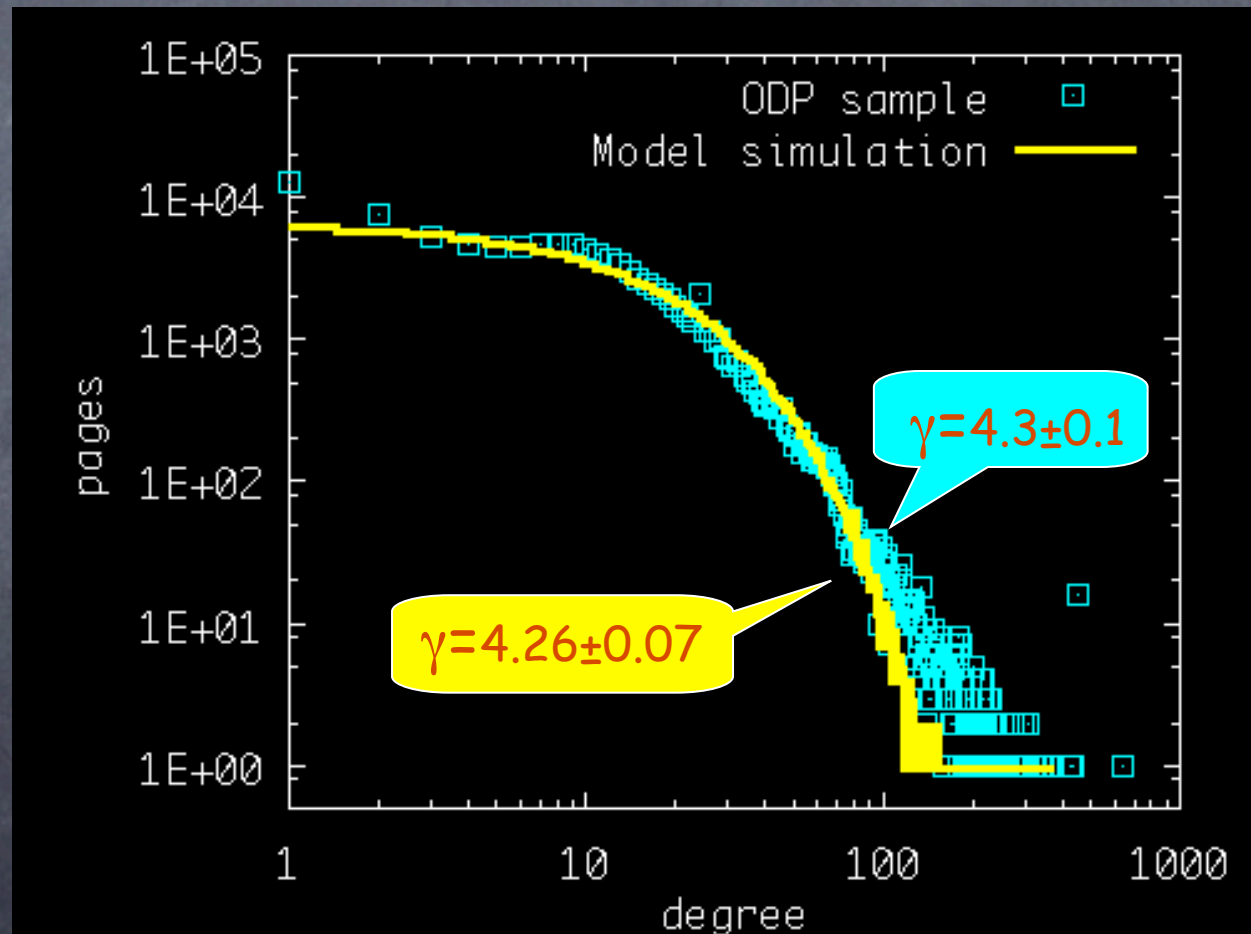
Power law tail
 $\Pr(\lambda | \rho) \sim \rho^{-\alpha(\lambda)}$

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

Local content-based growth model

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

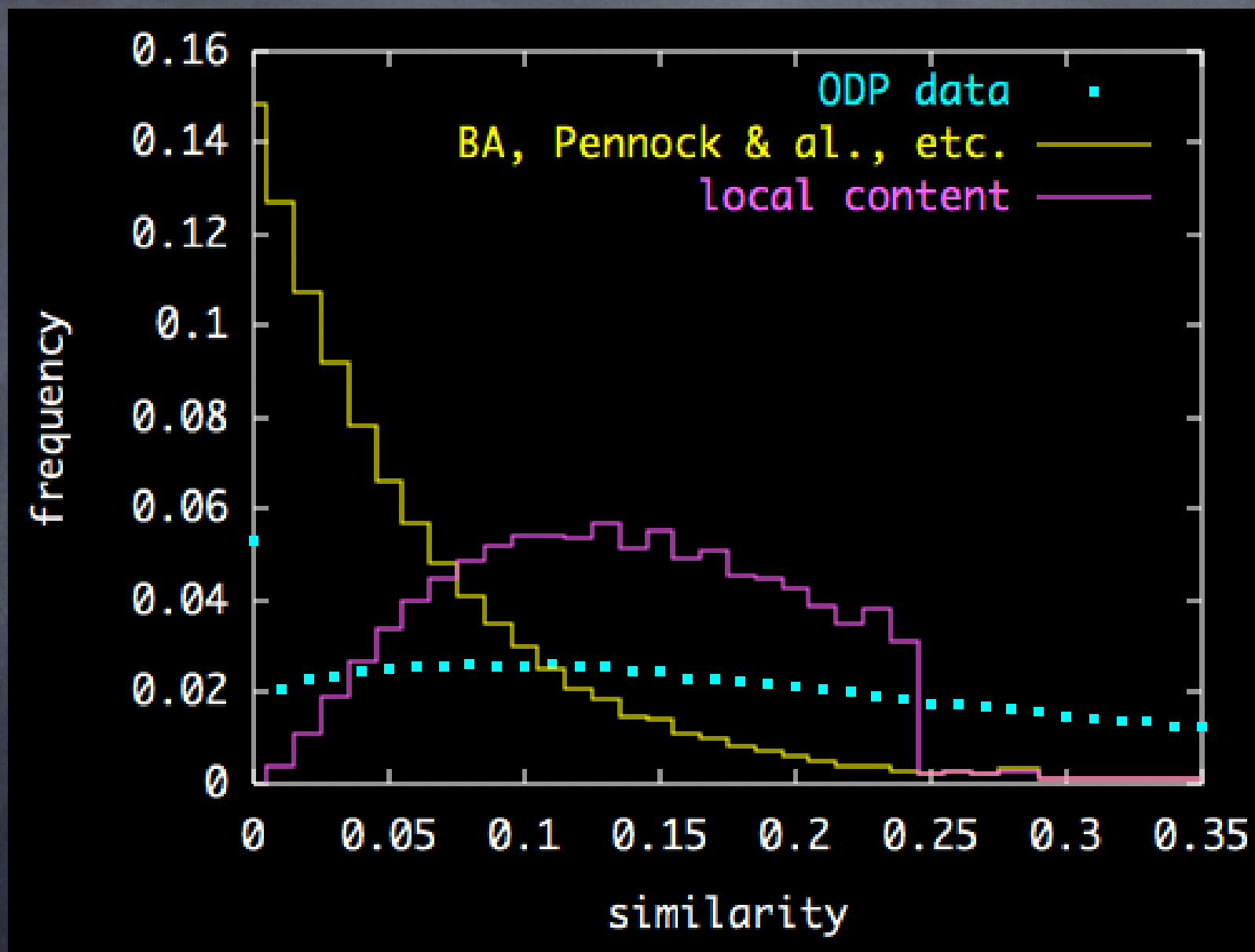
- Similar to preferential attachment (BA)
- Use degree info (popularity/ importance) only for nearby (similar/ related) pages



So, many models can predict degree distributions...

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

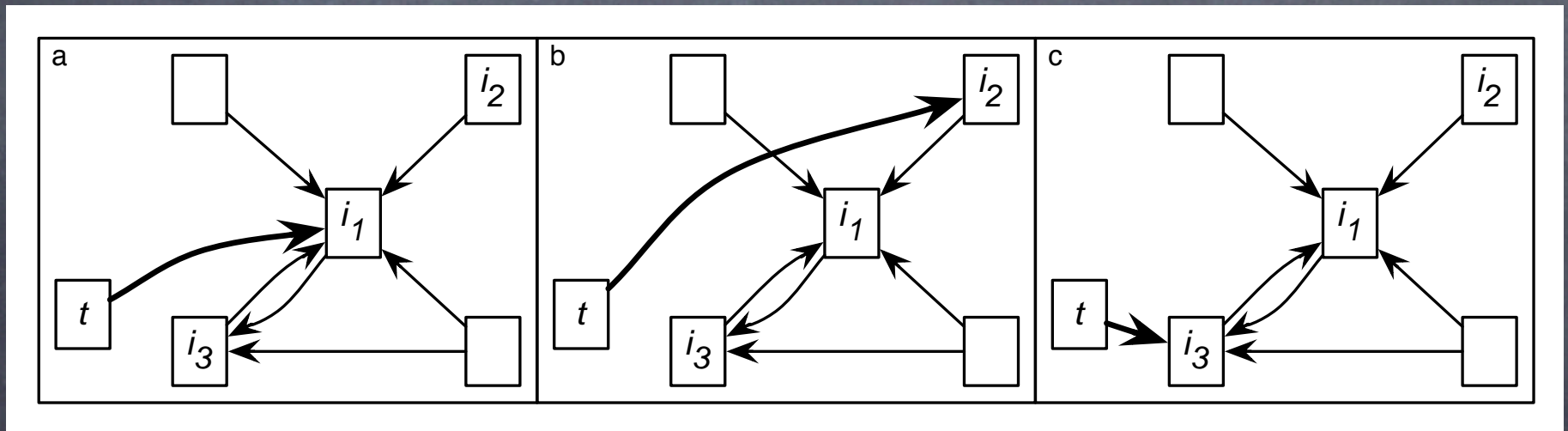
None of these models is right!



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

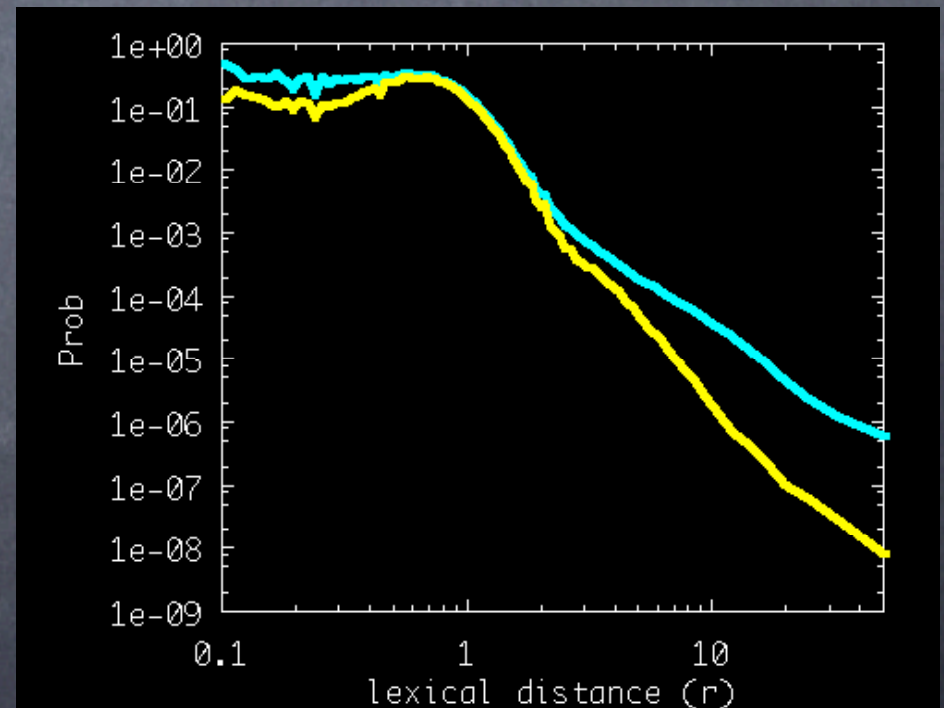
Degree-similarity mixture model

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

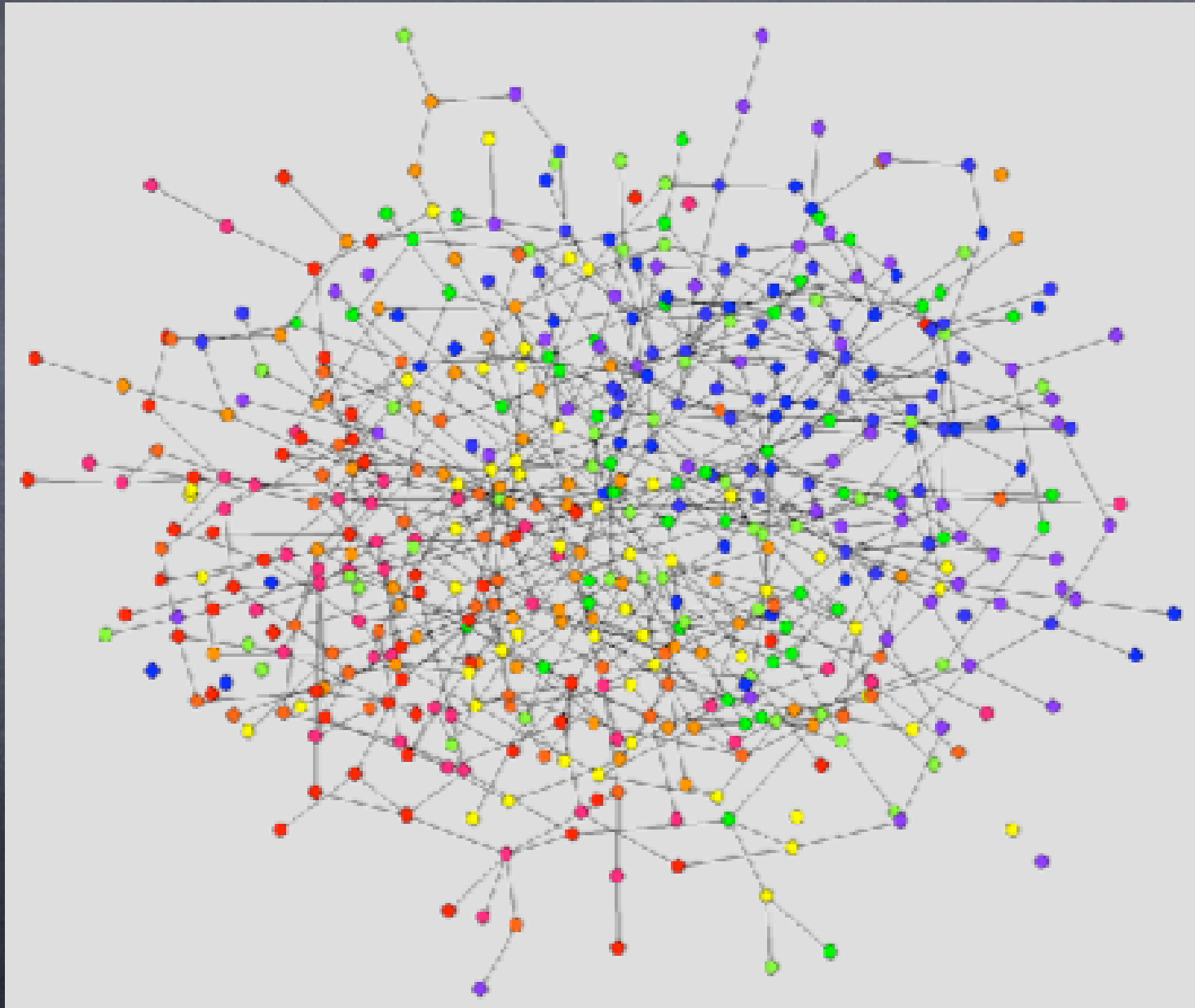


$$\hat{\text{Pr}}(i) \propto [r(i, t)]^{-\alpha}$$

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

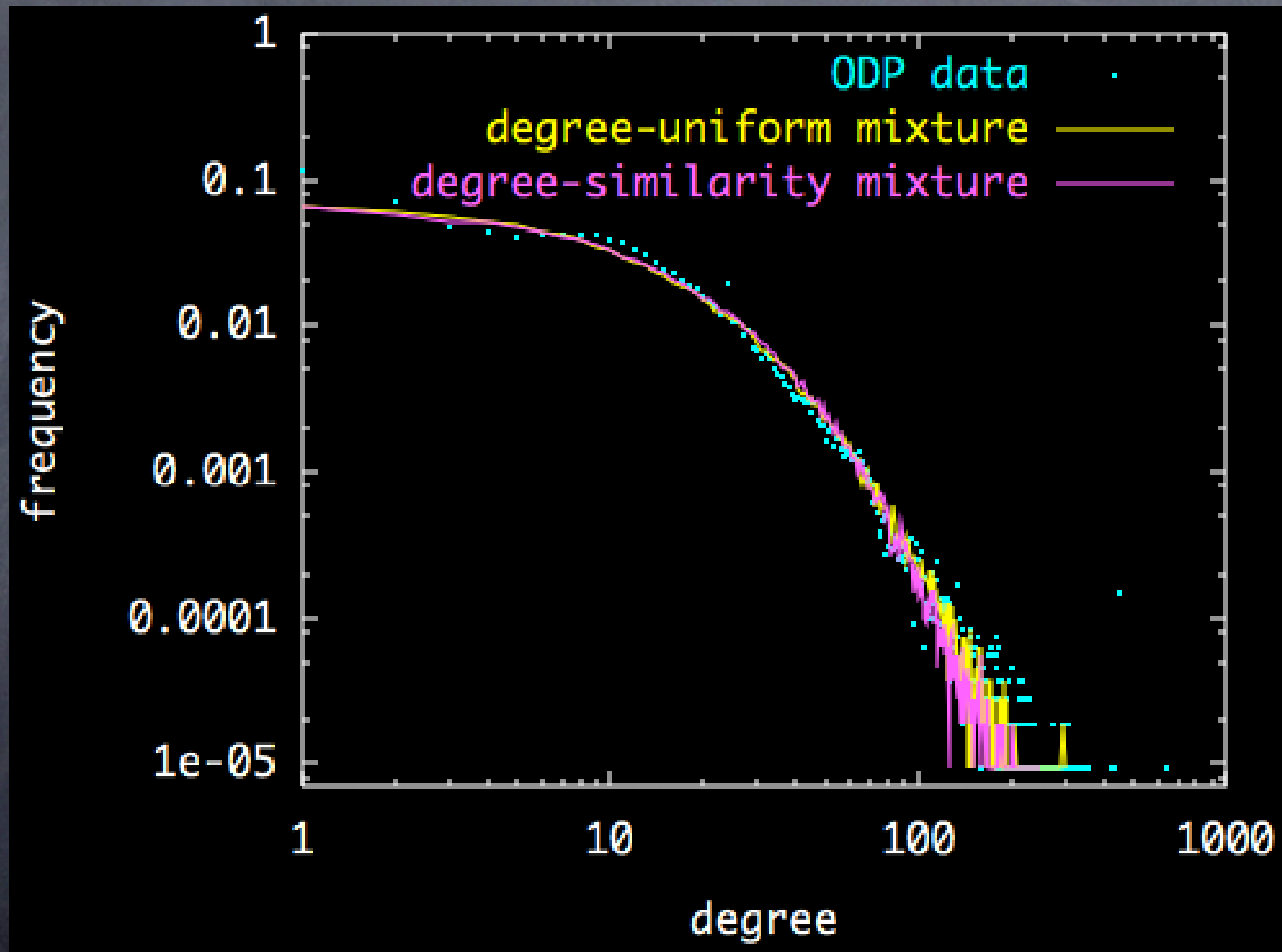


Build it...

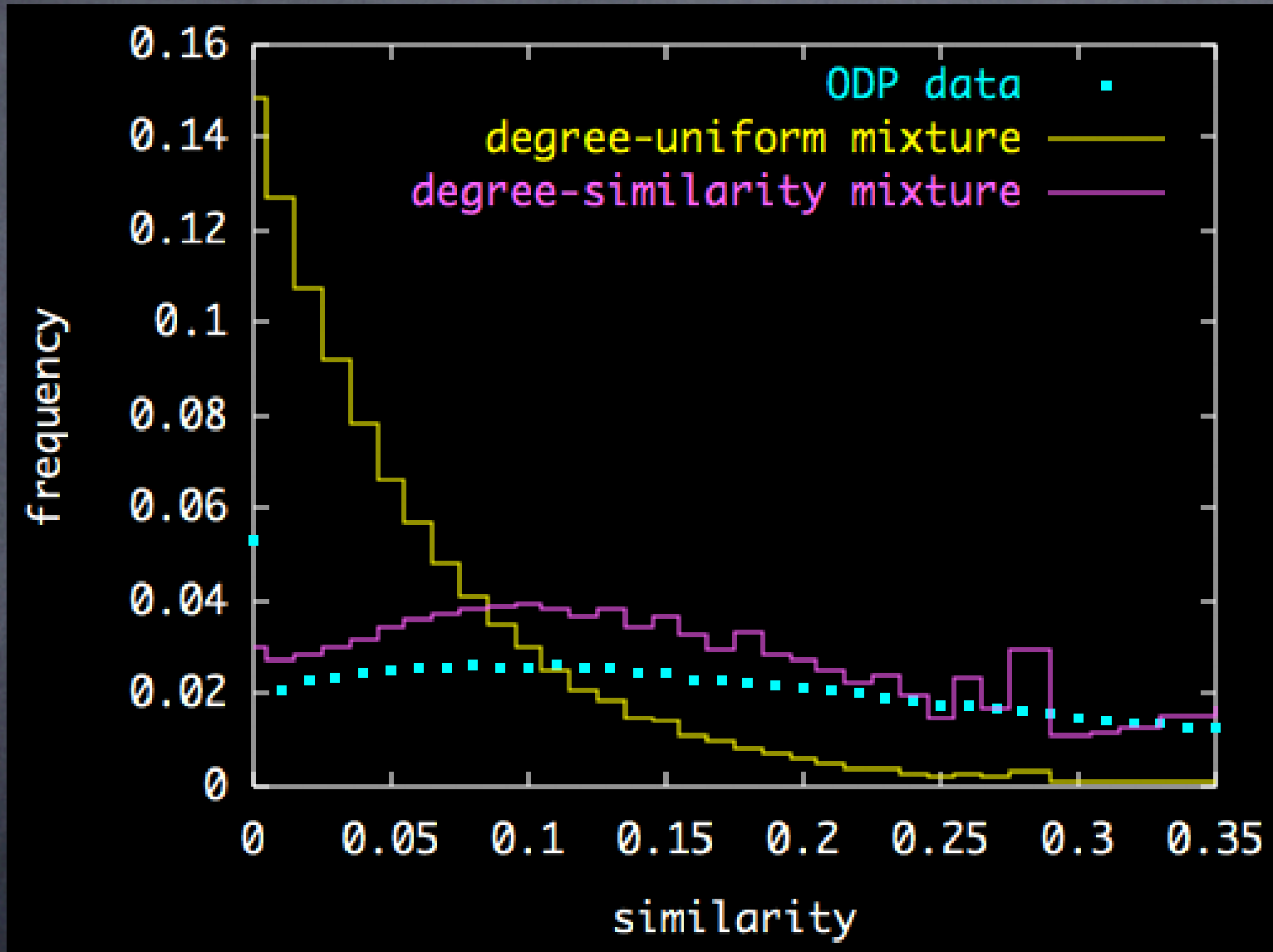


(M.M.)

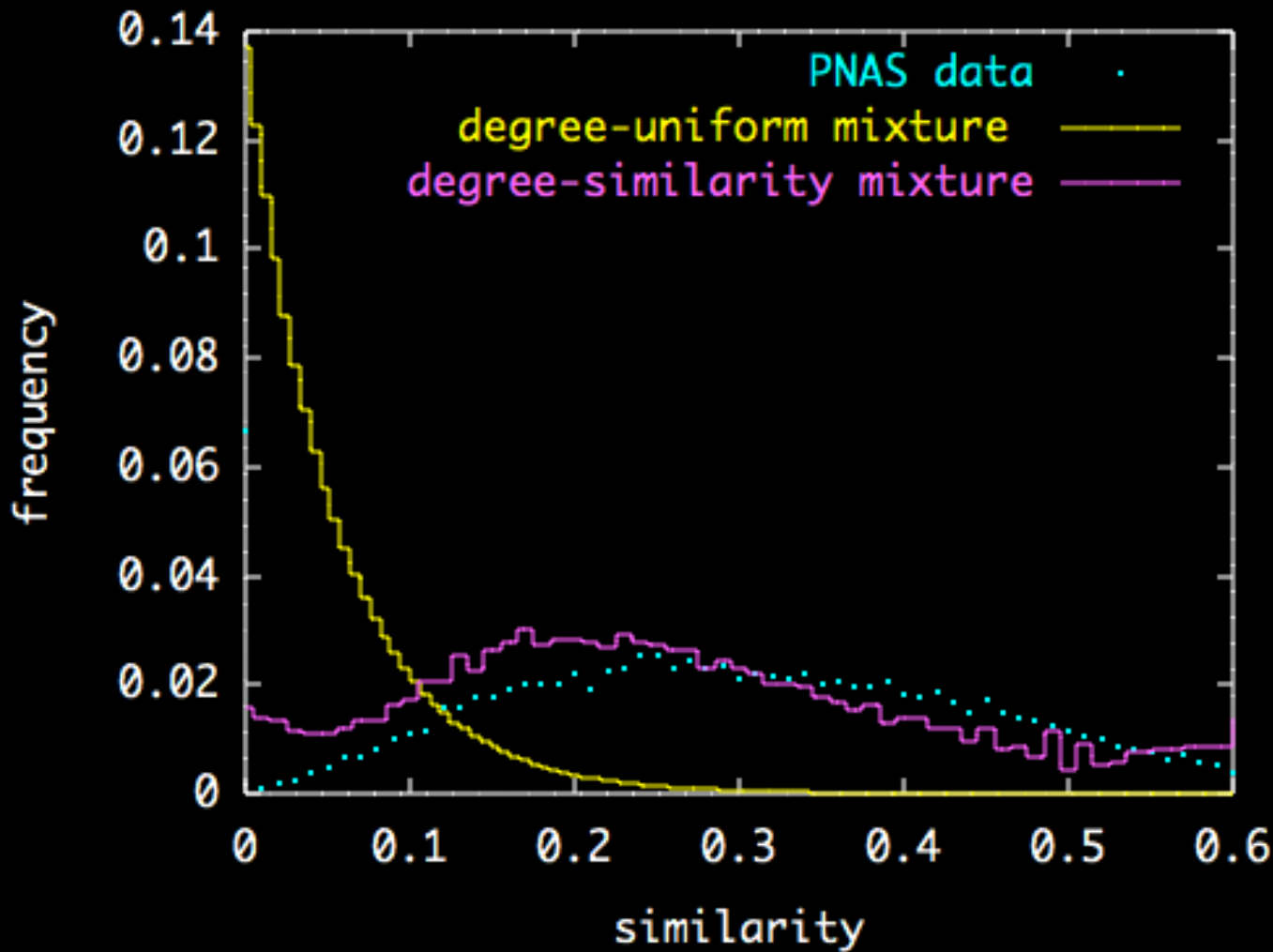
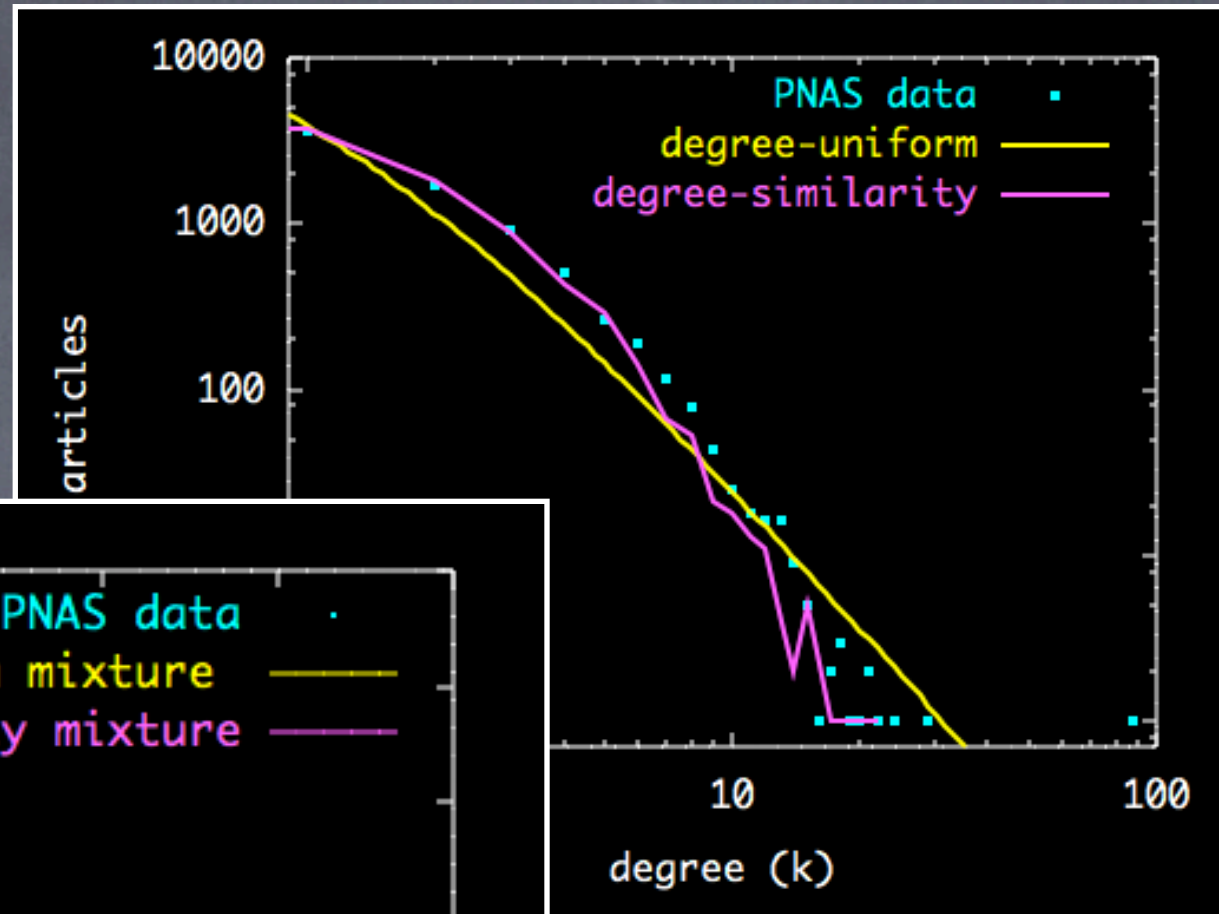
Both mixture models get the degree distribution right...



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



Citation networks



15,785 articles
published in
PNAS between
1997 and 2002

What now?

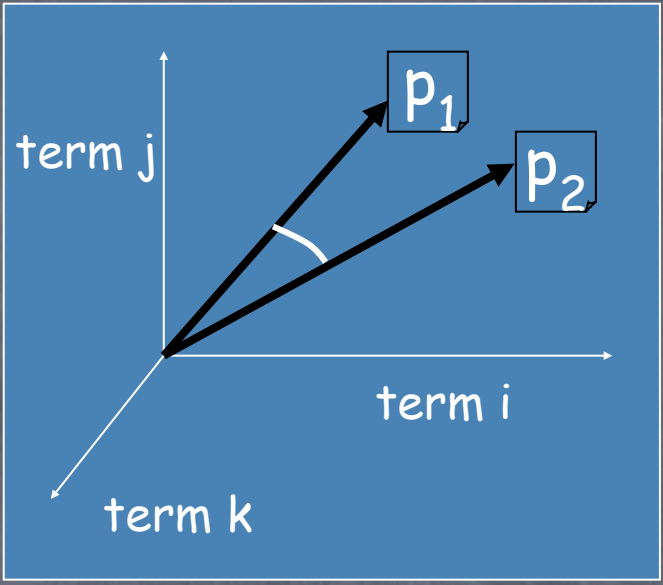
- Understand exponential distribution of similarity
- Growth model to explain evolution of both link topology and content similarity
 - With Alex Vespignani & Sandro Flammini

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!

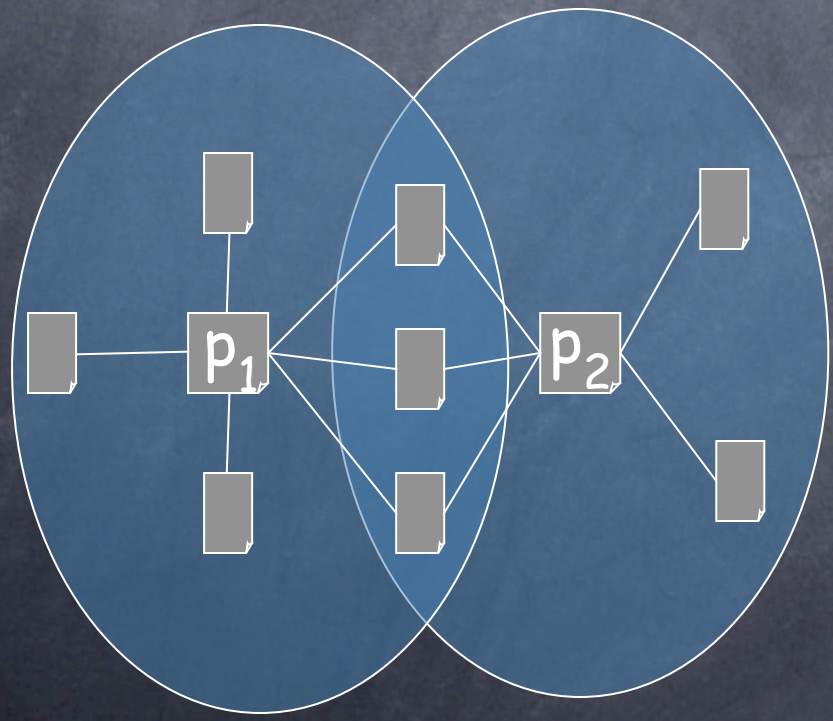
$$\sigma_c(\vec{p}_1, \vec{p}_2) = \frac{\vec{p}_1 \cdot \vec{p}_2}{\|\vec{p}_1\| \cdot \|\vec{p}_2\|}$$

Content similarity



Link similarity

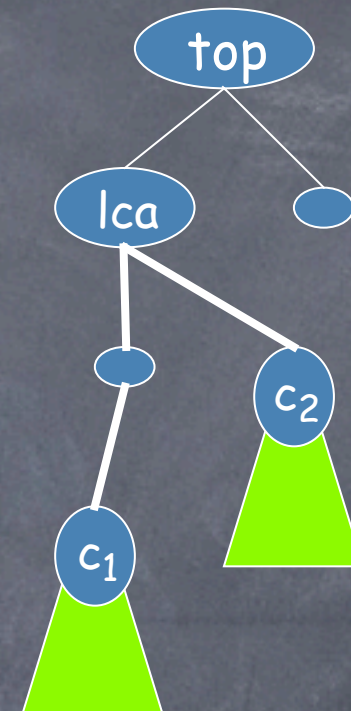
$$\sigma_l(p_1, p_2) = \frac{|U_{p_1} \cap U_{p_2}|}{|U_{p_1} \cup U_{p_2}|}$$



Semantic similarity

The screenshot shows the DMOZ directory page for 'Kids and Teens: People and Society: Holidays and Celebrations: Birthdays'. The page features a search bar, a list of 11 items, and a 'Description' link. The items listed are:

- [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.

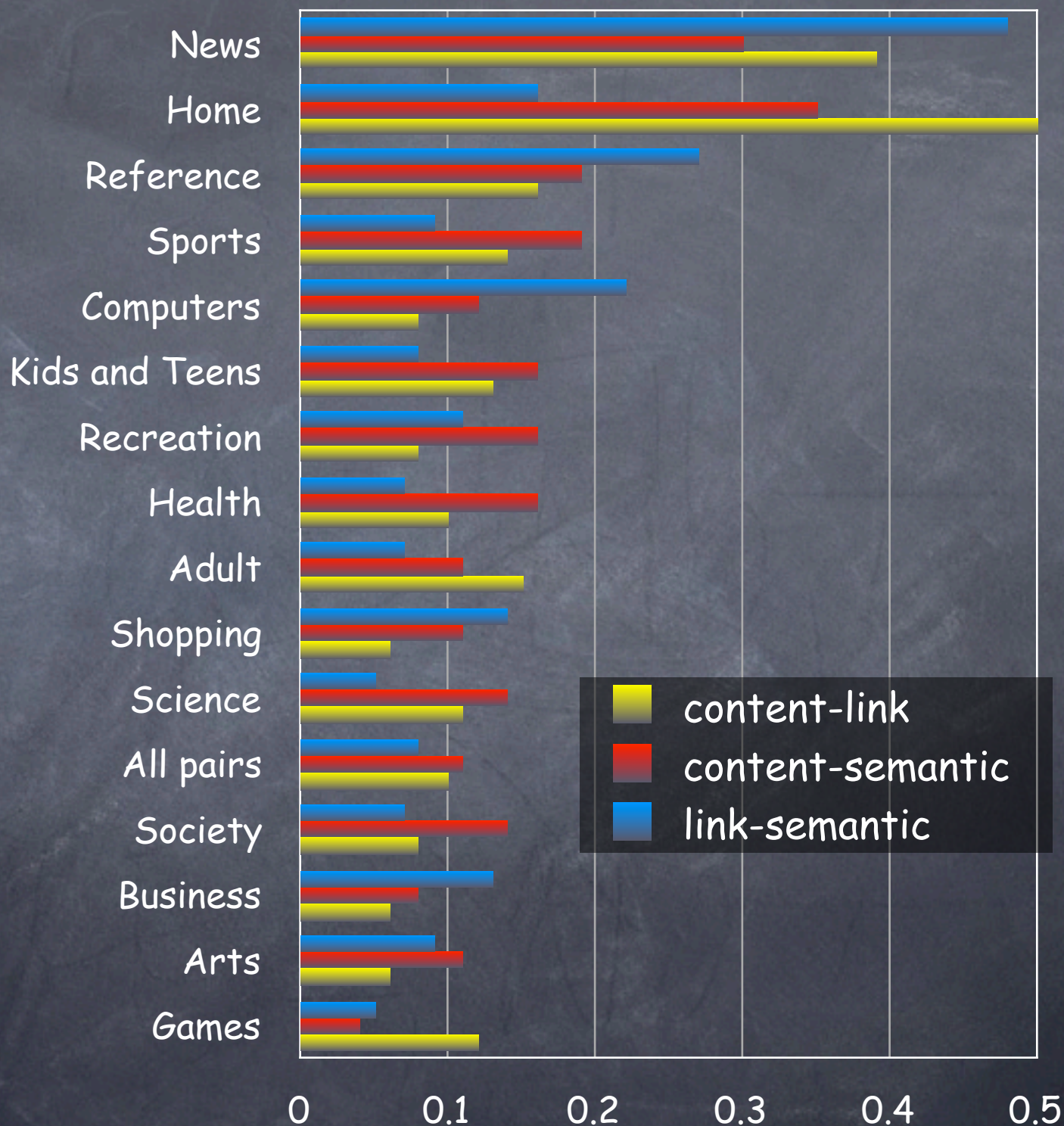


- Information-theoretic measure based on classification tree (Lin 1998)

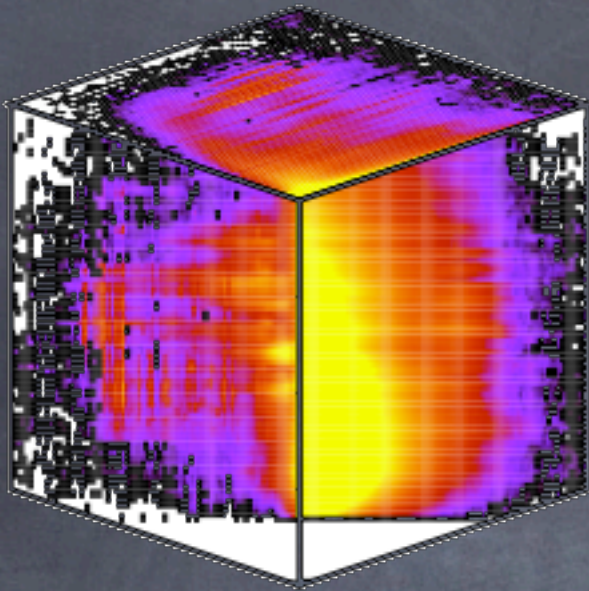
$$\sigma_s(c_1, c_2) = \frac{2 \log \Pr[lca(c_1, c_2)]}{\log \Pr[c_1] + \log \Pr[c_2]}$$

- Classic path distance in special case of balanced tree

Correlations between similarities



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



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

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

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

Averaging
semantic
similarity

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

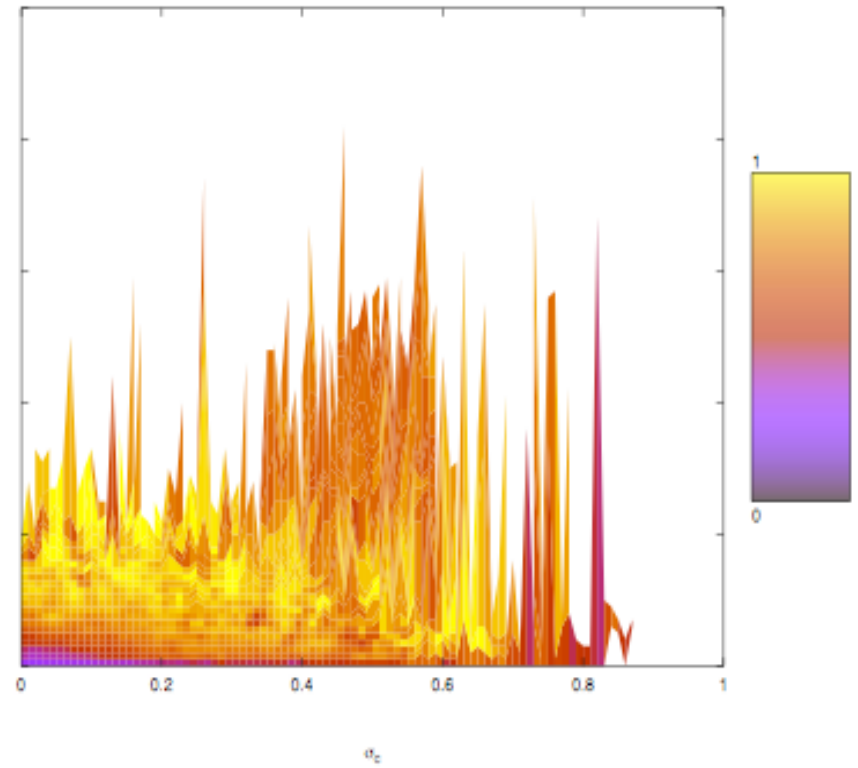
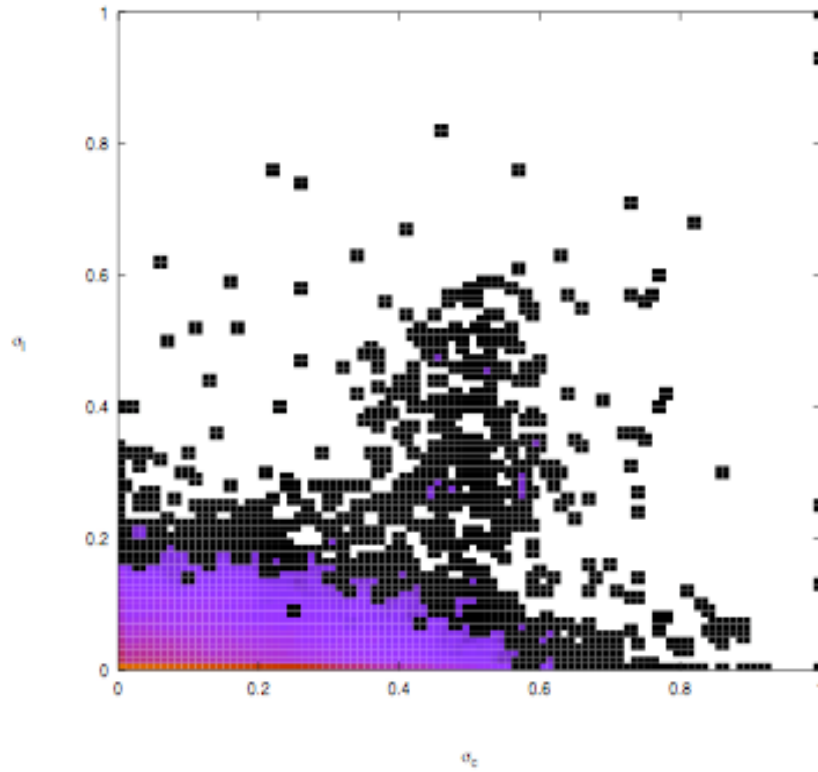
Summing
semantic
similarity

Business

σ_ℓ

log Recall

Precision



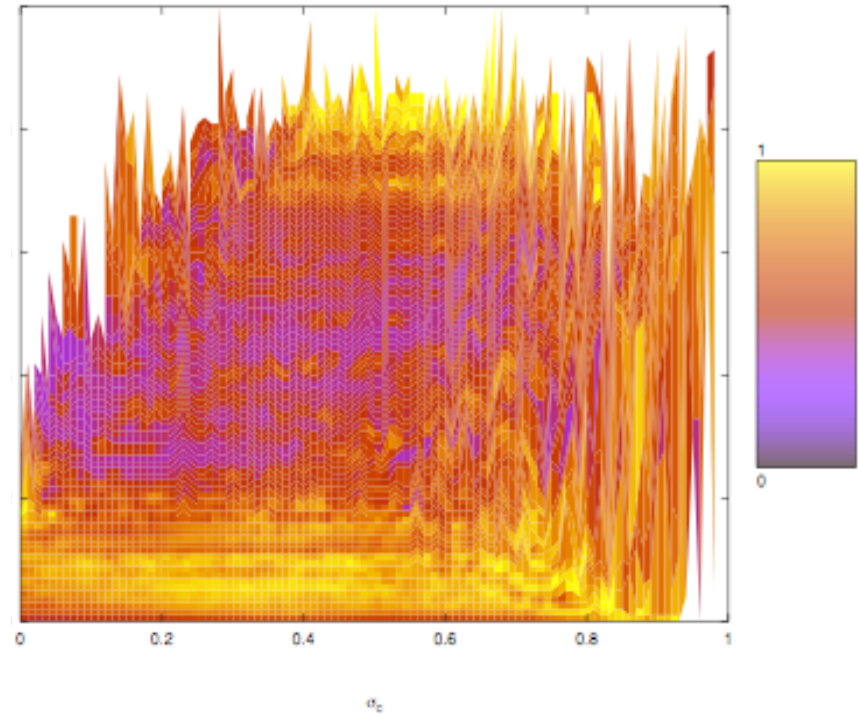
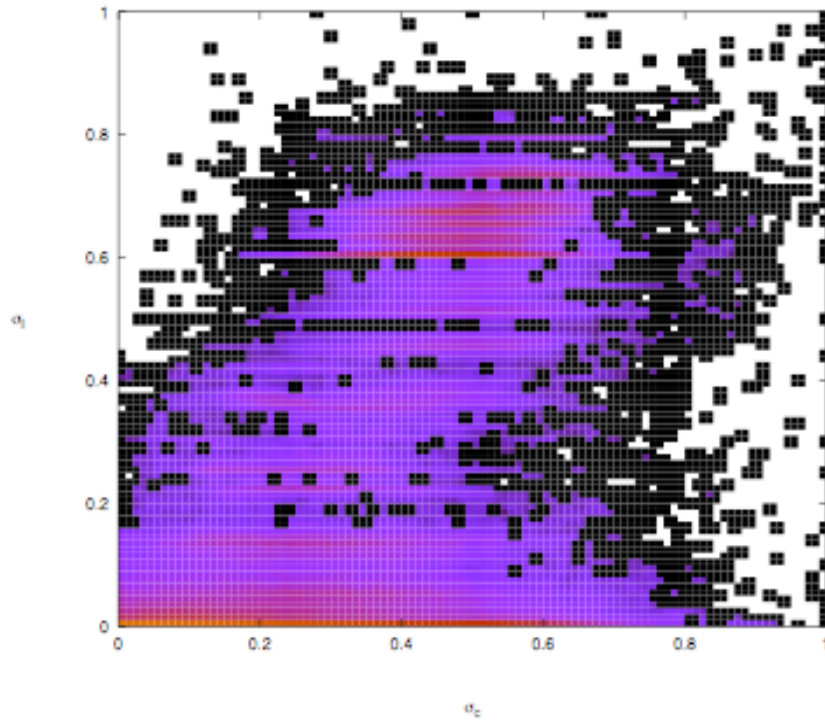
σ_c

Home



log Recall

Precision

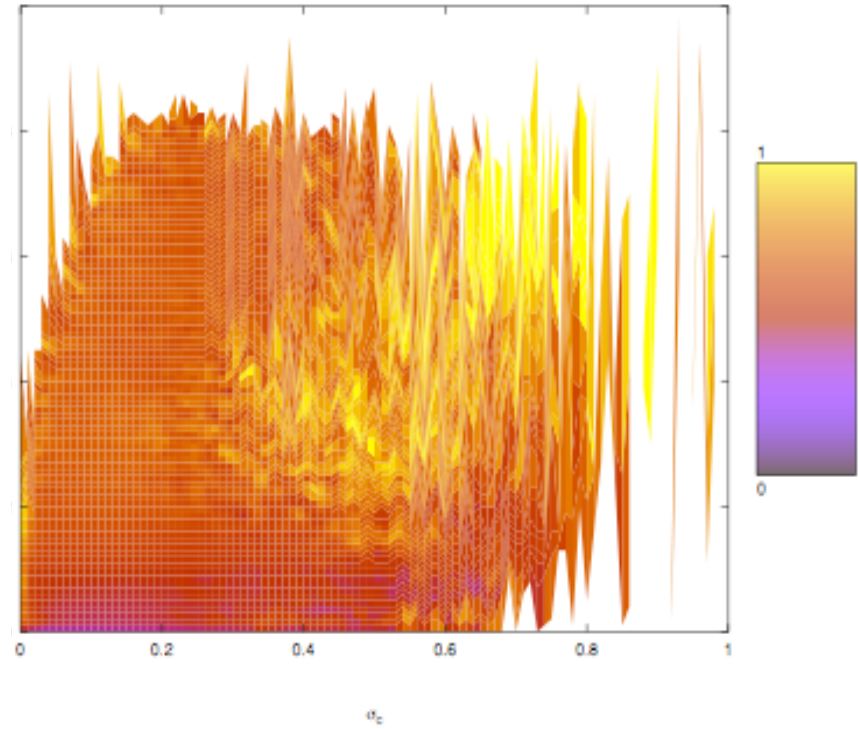
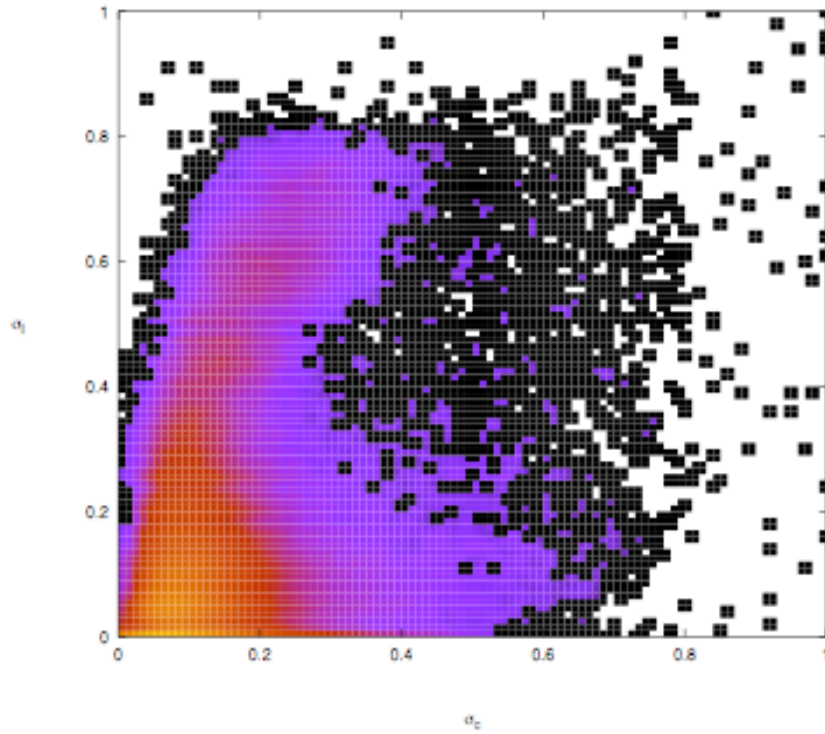


News

σ_ℓ

log Recall

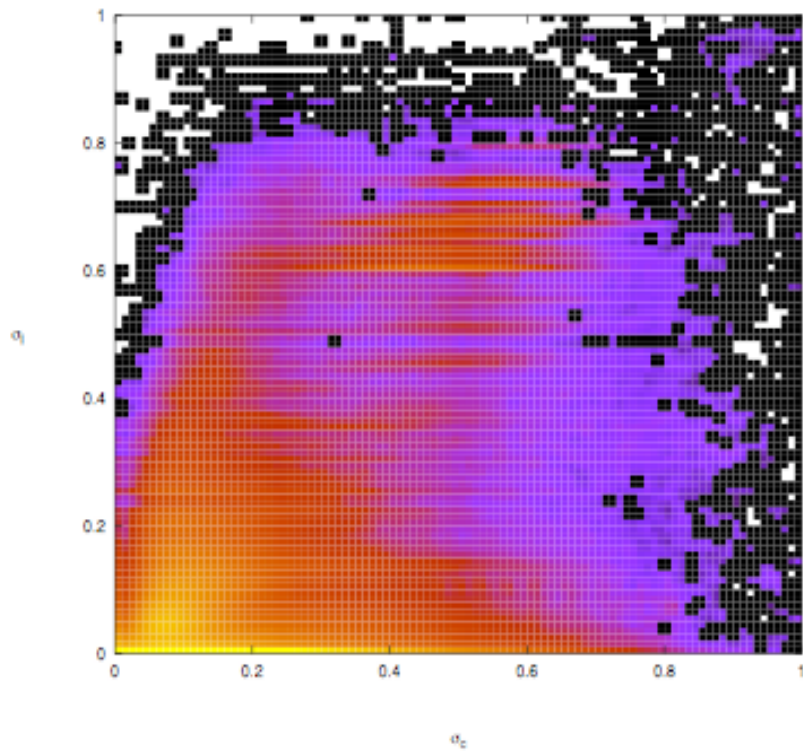
Precision



σ_c

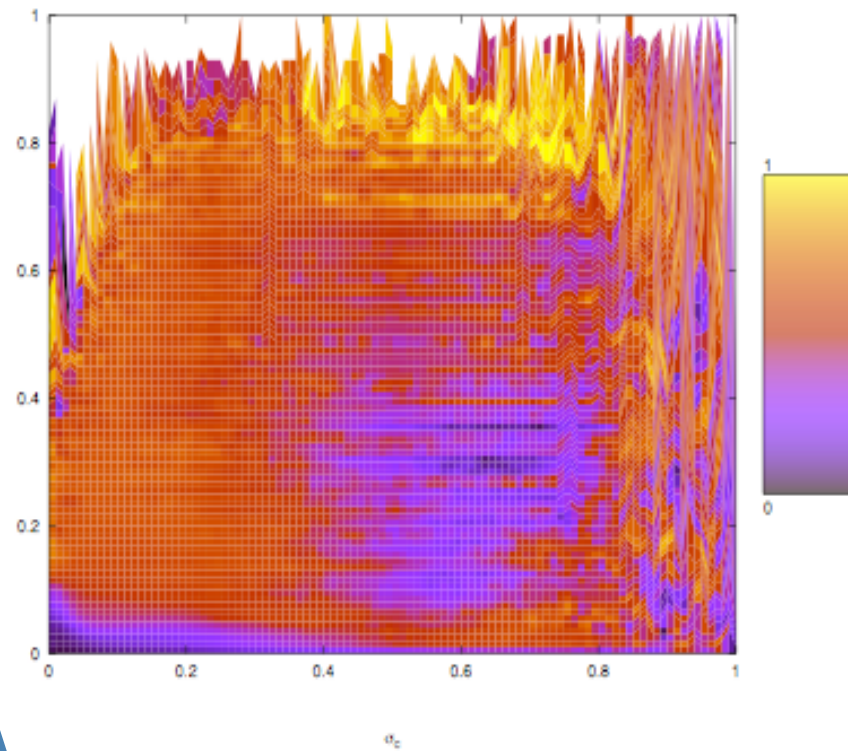
All pairs

log Recall

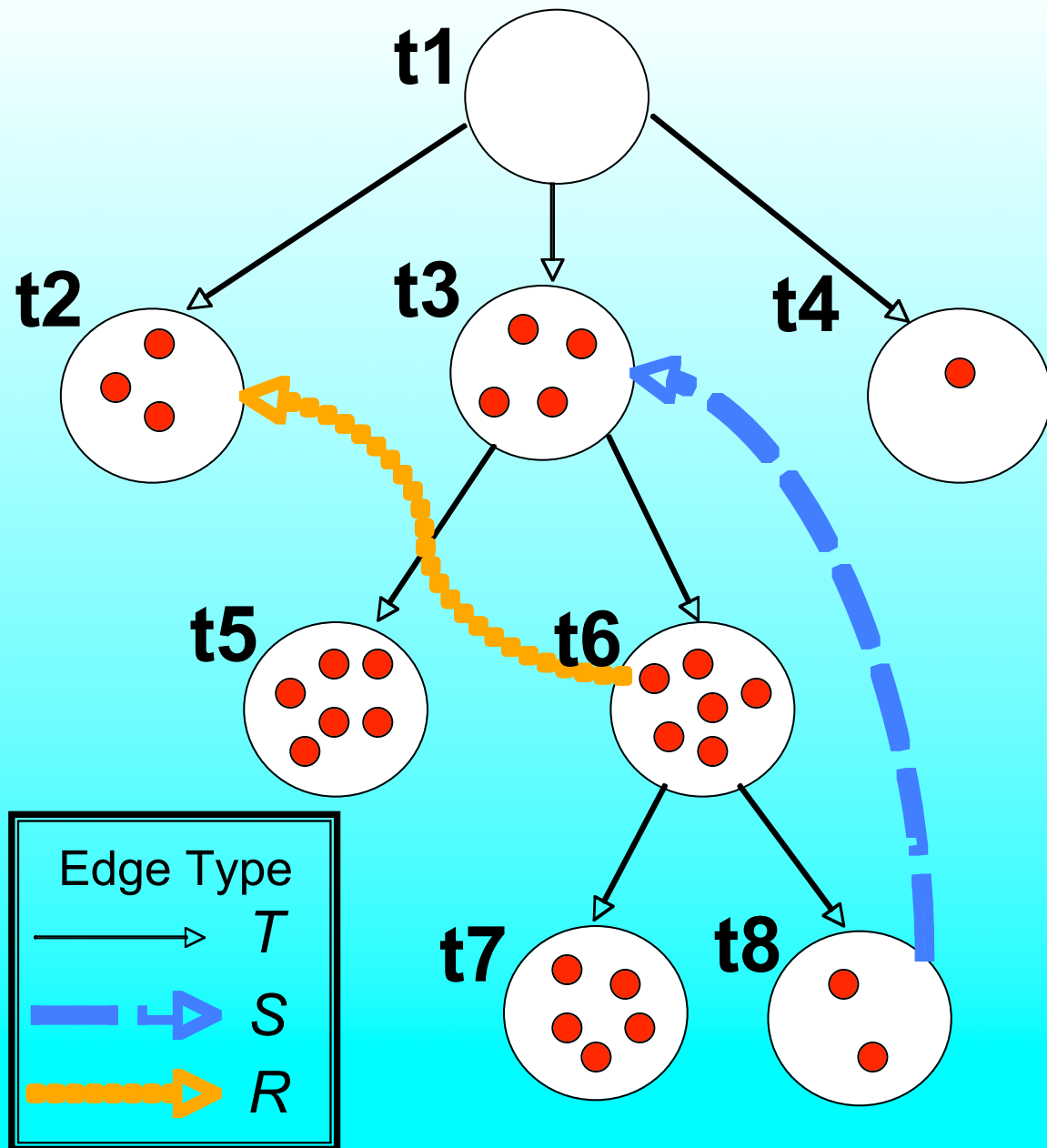


σ_l

Precision



σ_c



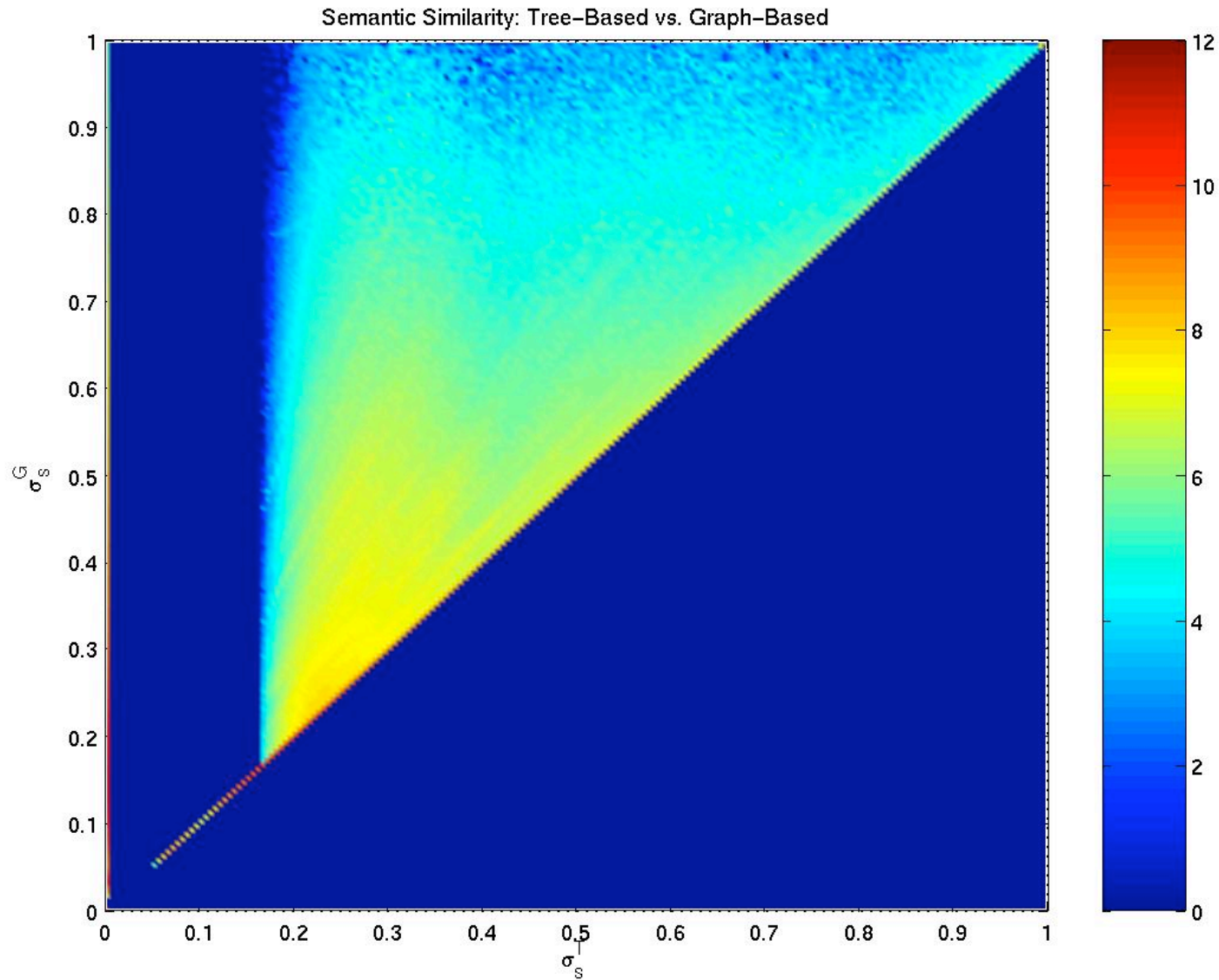
Better semantic similarity measure

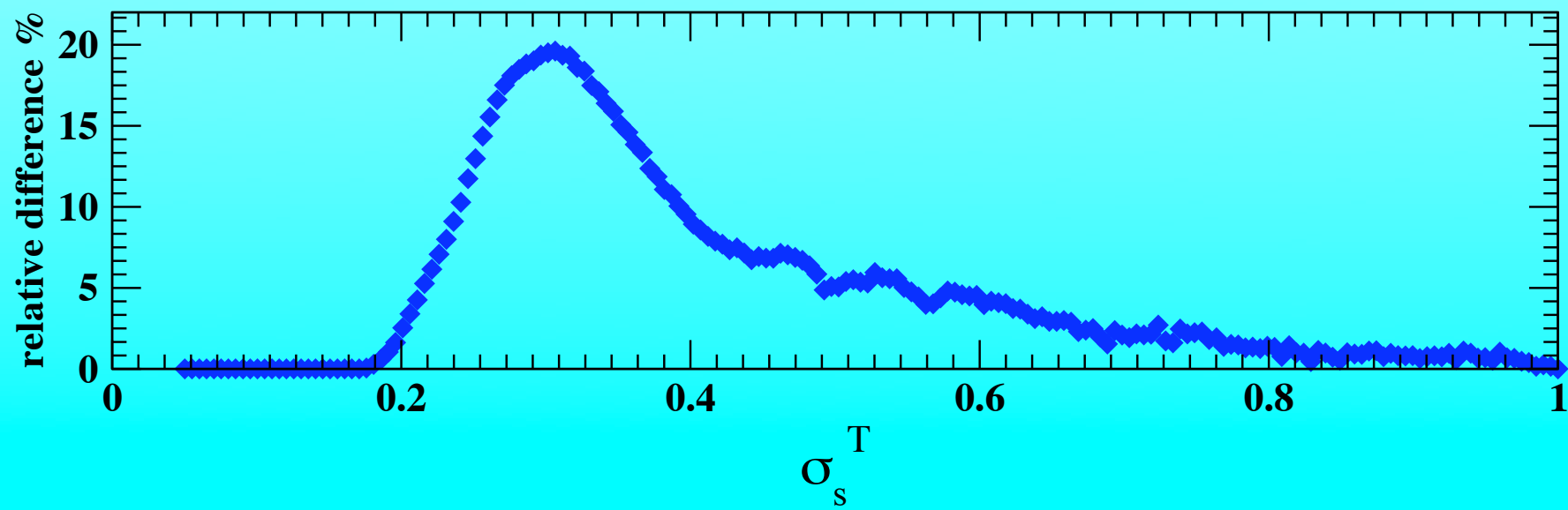
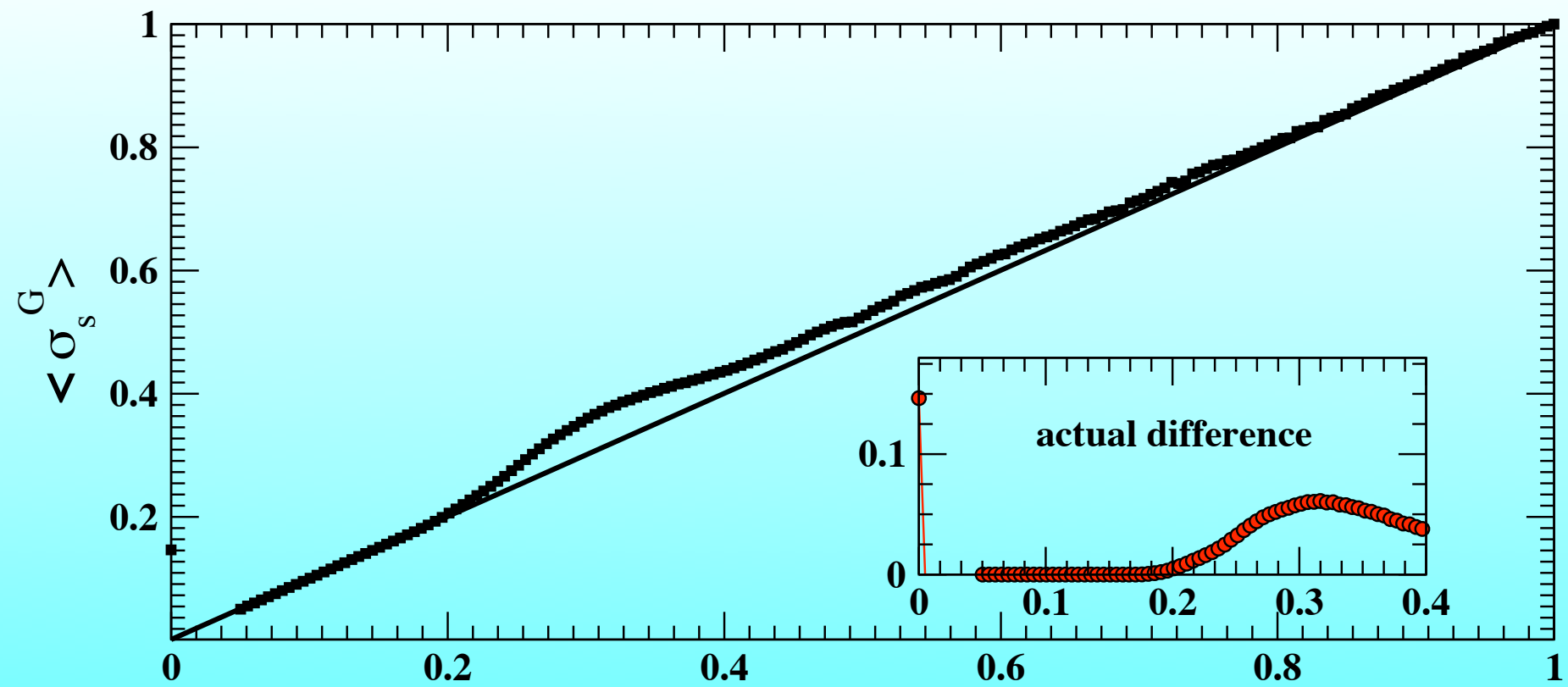
- Work w/ Ana Maguitman, Heather Roinestad, Alex Vespignani
- Include cross-links (symbolic) and see-also links (related)
- Transitive closure of topic graph
- Compute entropy based on fuzzy membership matrix

$$W = T^+ \odot G \odot T^+$$


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

Differences





Web Page Relatedness




MuppetsOnline.com


Welcome

Home
Shop
Music
Pictures
Sounds

The Eagle Speaks

Hello and welcome to MuppetsOnline.com. This is Sam the Eagle speaking. You can travel around this site merely by clicking the buttons to your left. Home will take you back here. Each section is looked after by a Muppet, so look around as there is lots to do. Visit the Muppet Shop and buy some of our work for yourself or a loved one. It's patriotic and American.




amazon.co.uk
Amazon Recommends:

[The Muppet Christmas Carol](#)
Michael Caine
£12.99

CLICK HERE TO BUY THIS MOVIE

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

"YOUR ONE SOURCE FOR QUALITY FAMILY ENTERTAINMENT"



THE LIVE CAST OF SESAME STREET

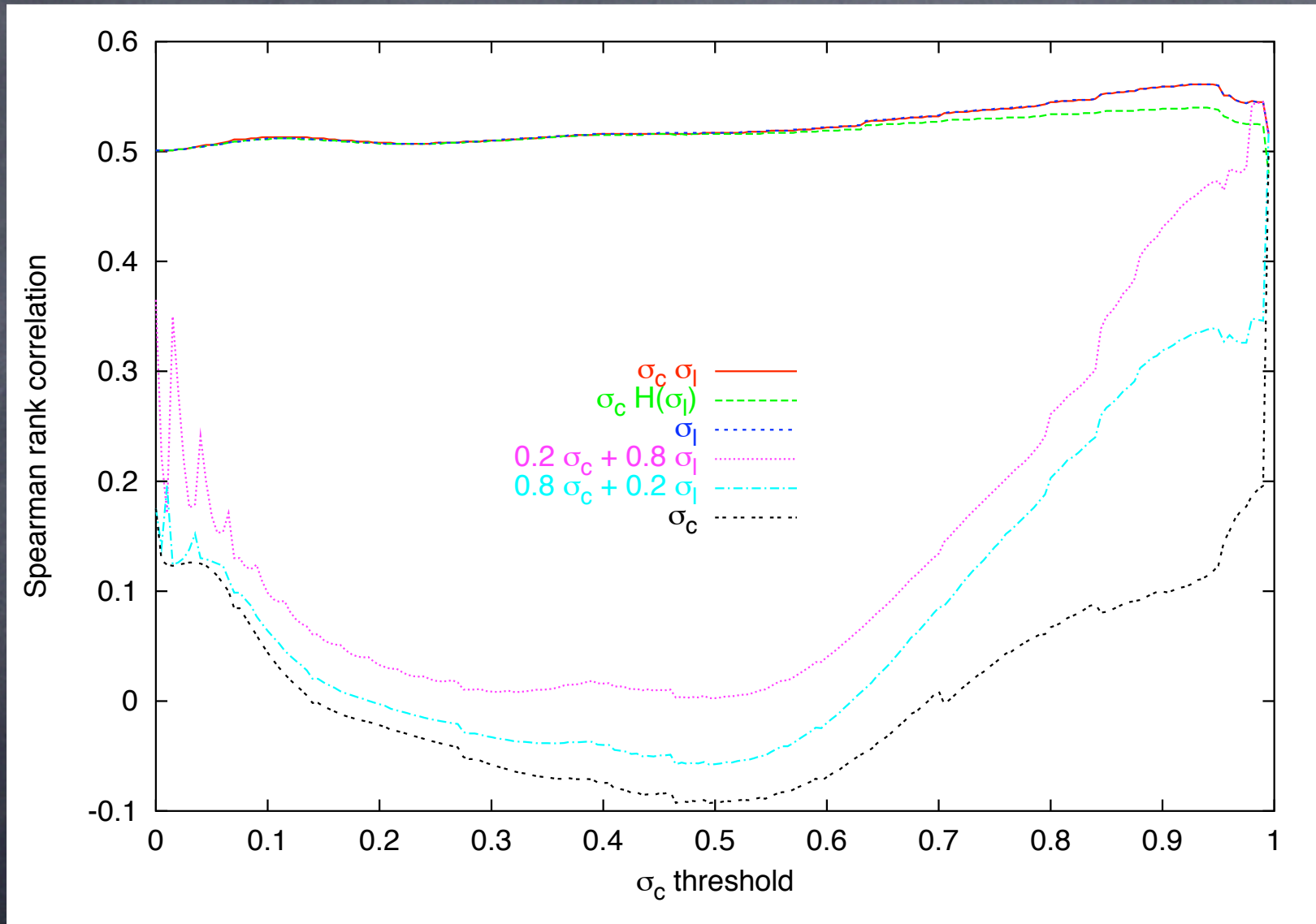
The accolades have all been written many, many times.

Quite simply, there has never been or exists now a television show like **SESAME STREET**.

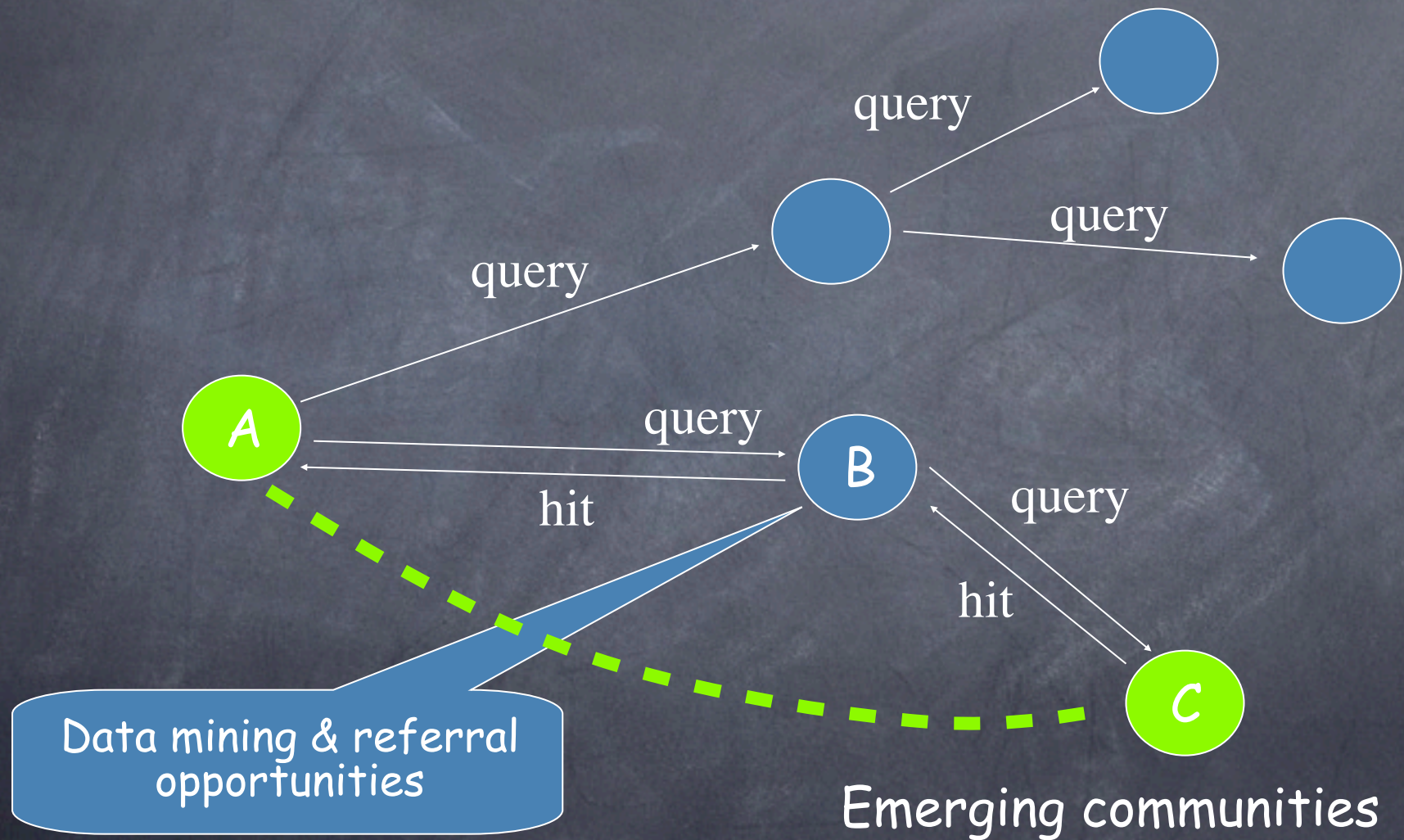
SESAME STREET has educated and entertained millions of children all over the world for thirty years! Its name has become synonymous with high quality. Much of the show's success is due to an extremely talented cast of actors.

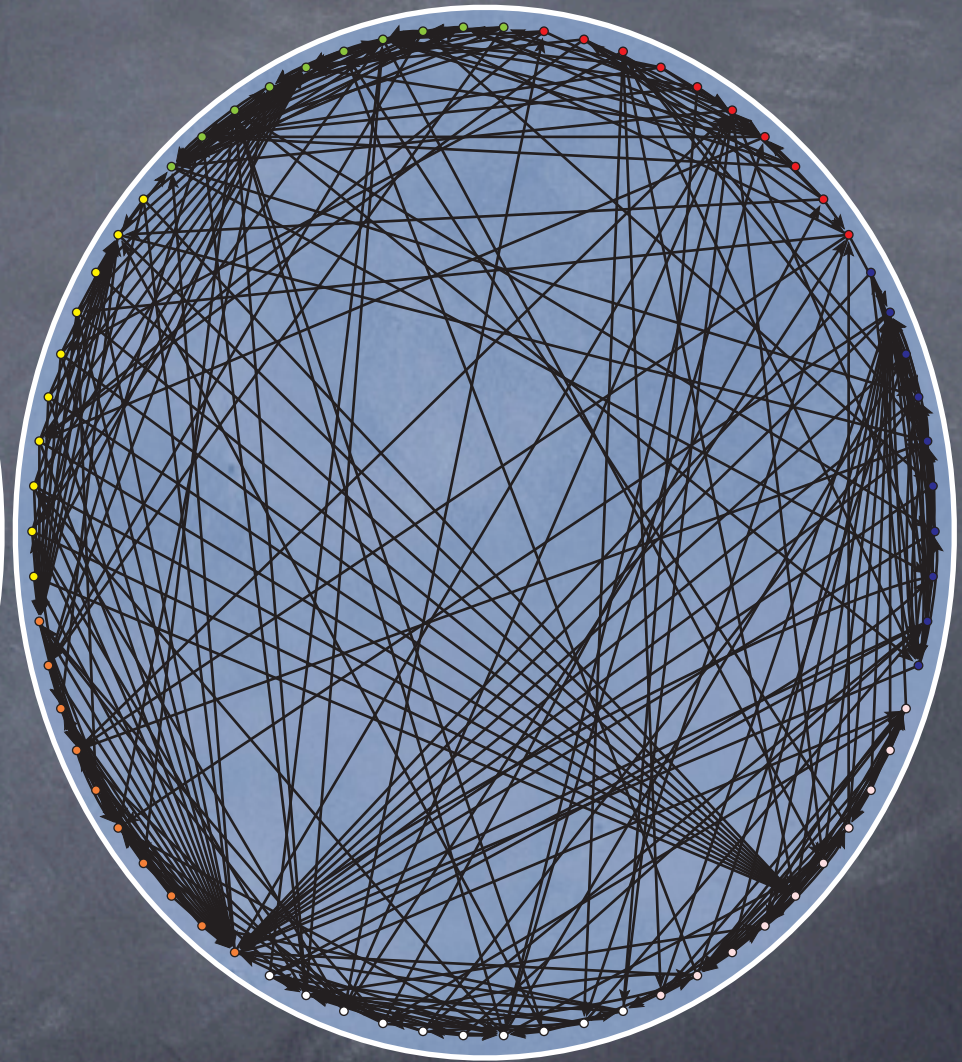
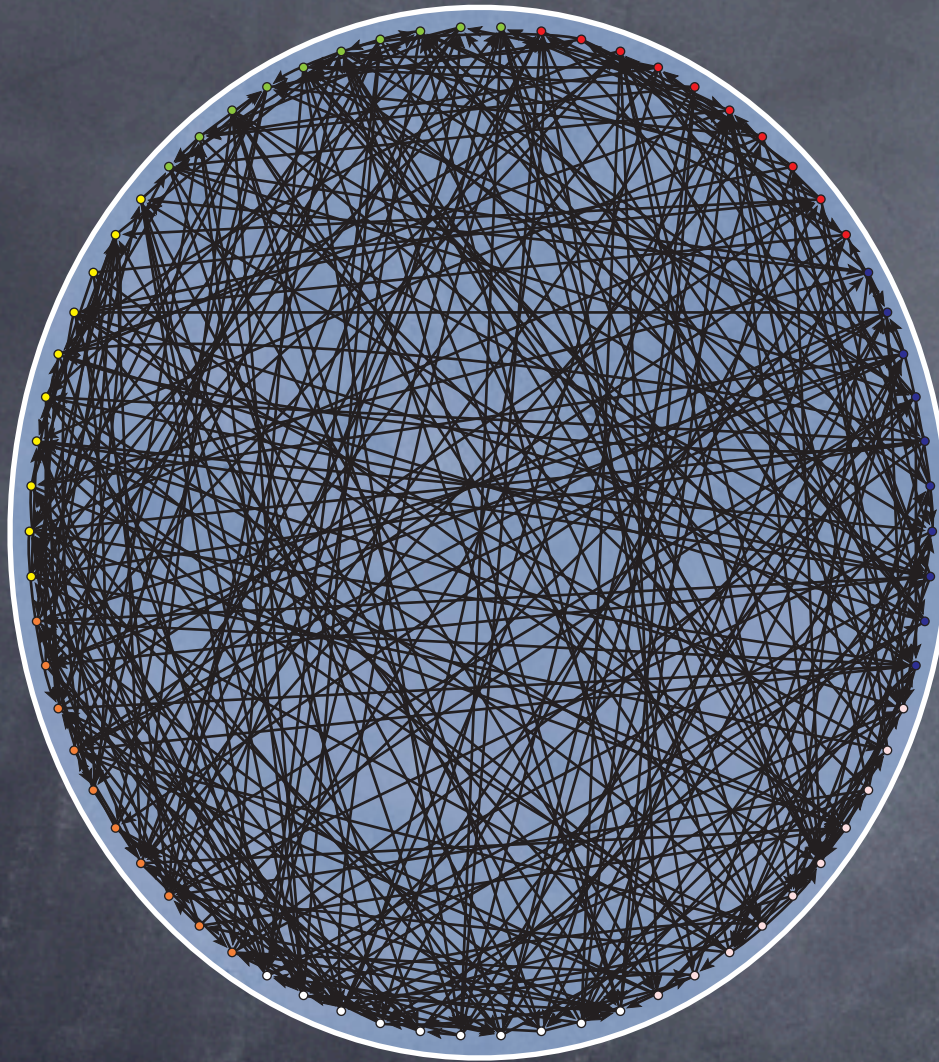
[ALL ABOUT](#) - [PHOTOS](#) - [UPCOMING EVENTS](#) - [LINKS](#) - [CONTACT US](#)

Combining content & links

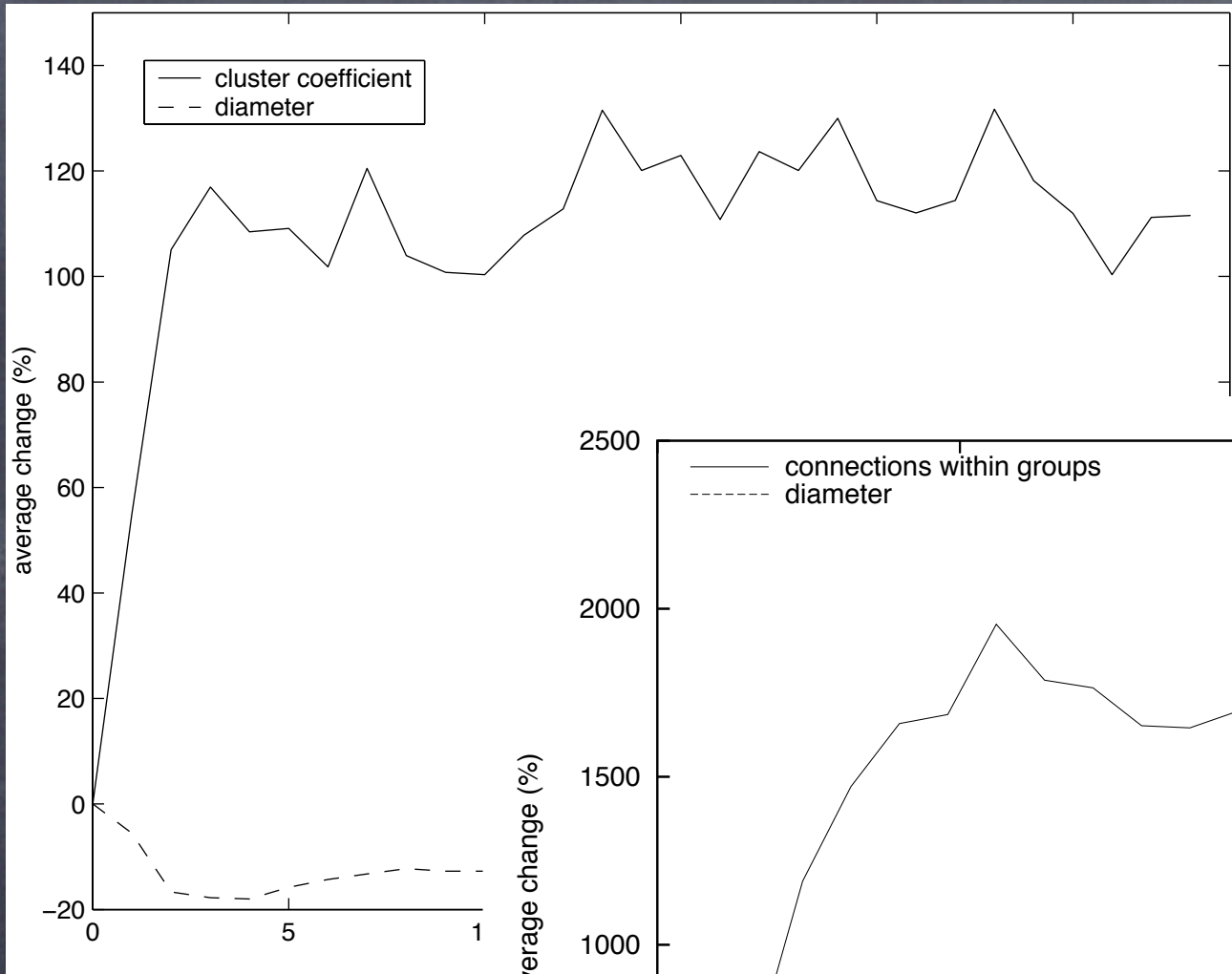


6S: peer distributed crawling and collaborative searching



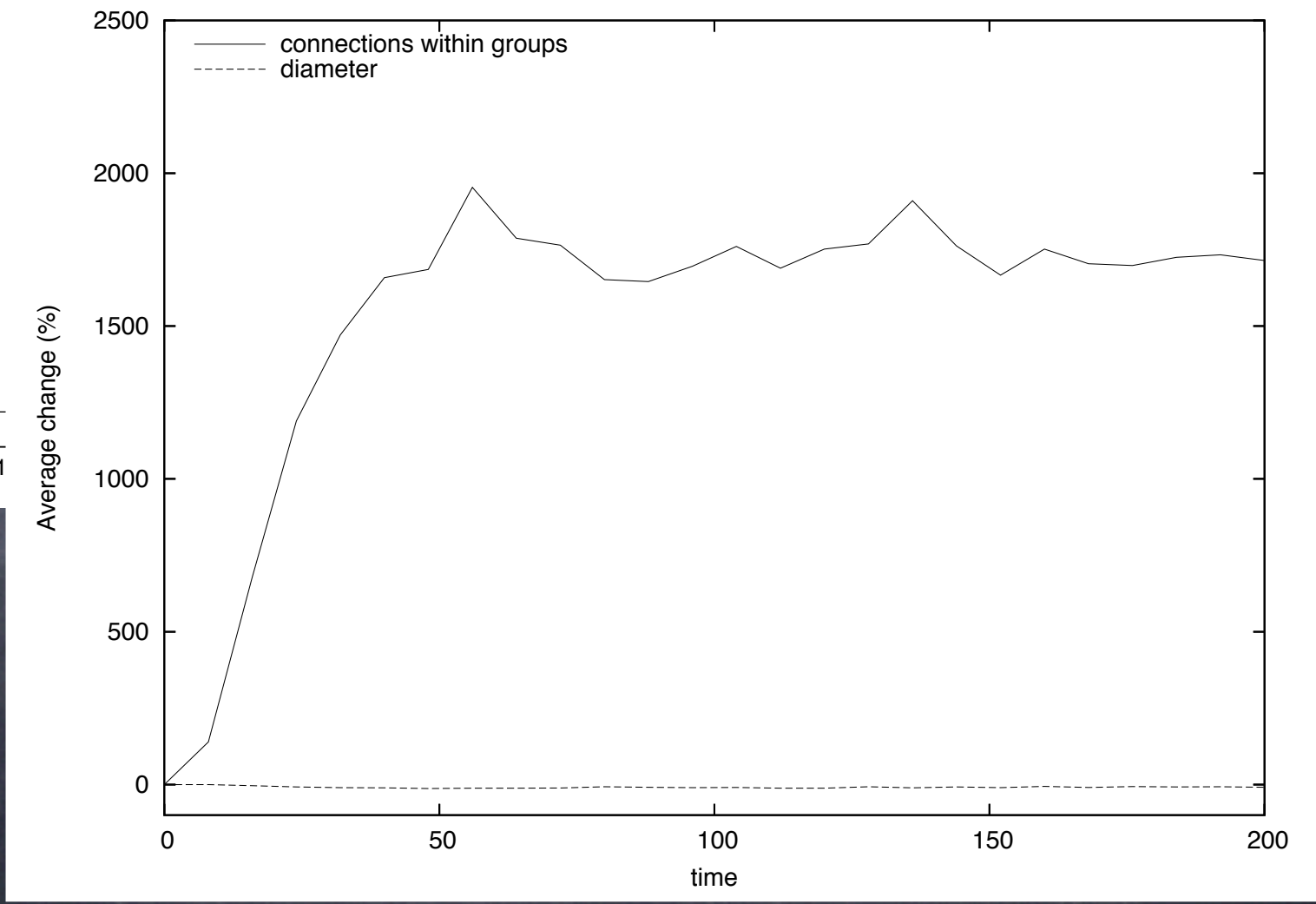


Work with Le-Shin Wu & Ruj Akavipat

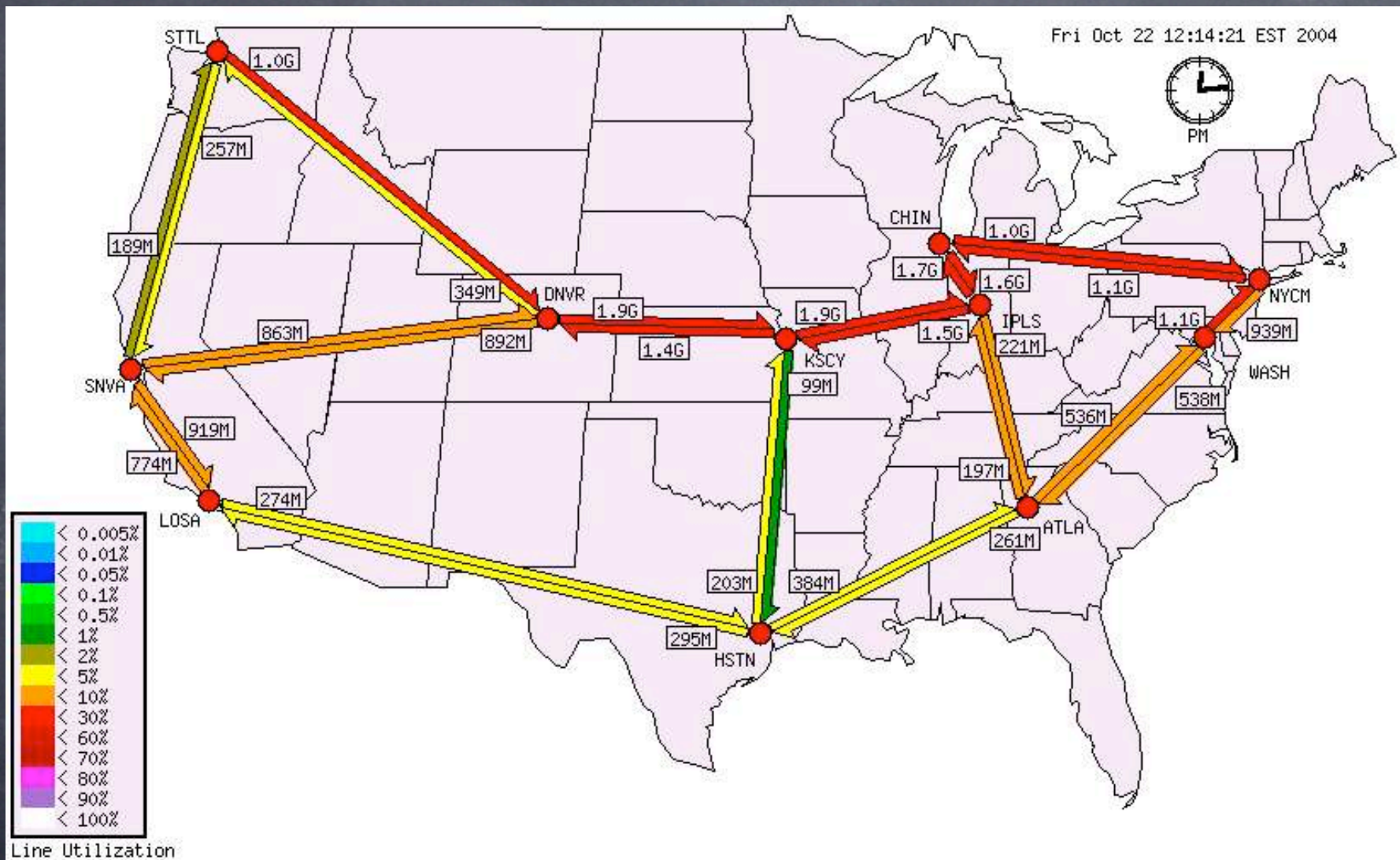


70 peers

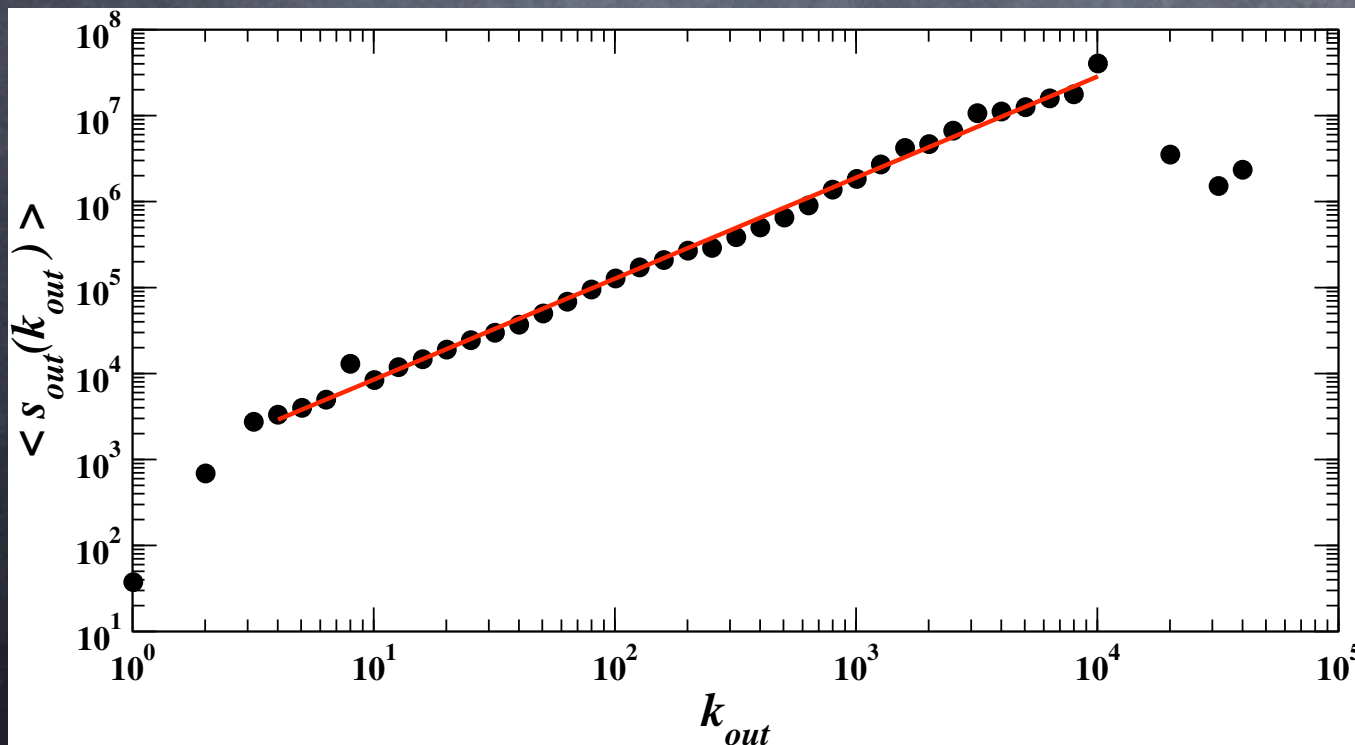
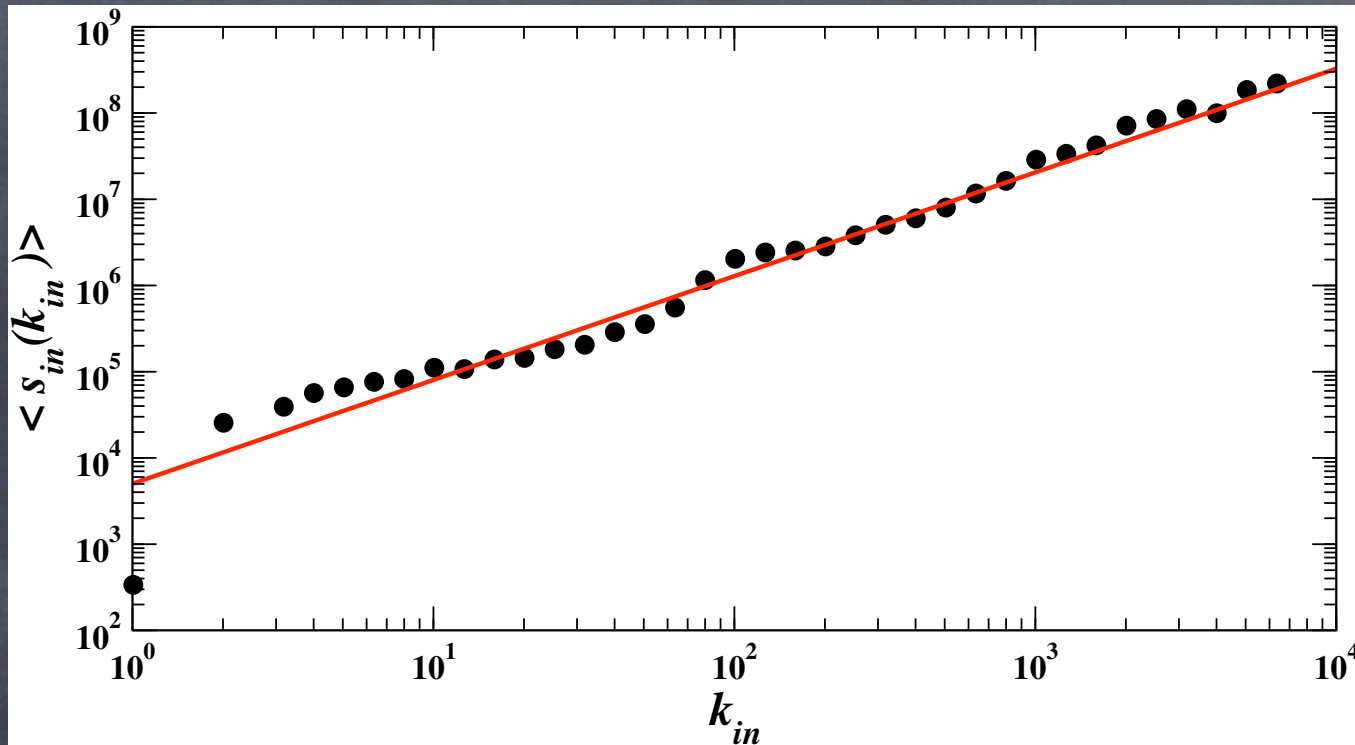
500 peers



WWW traffic network on Internet2



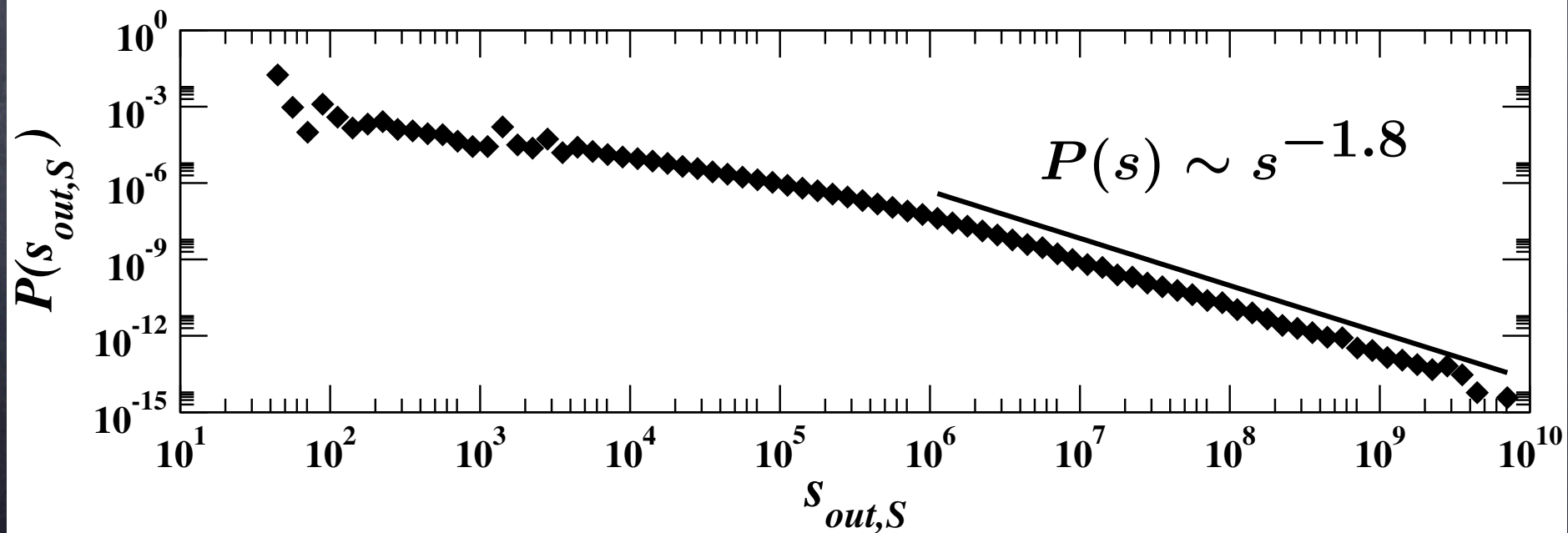
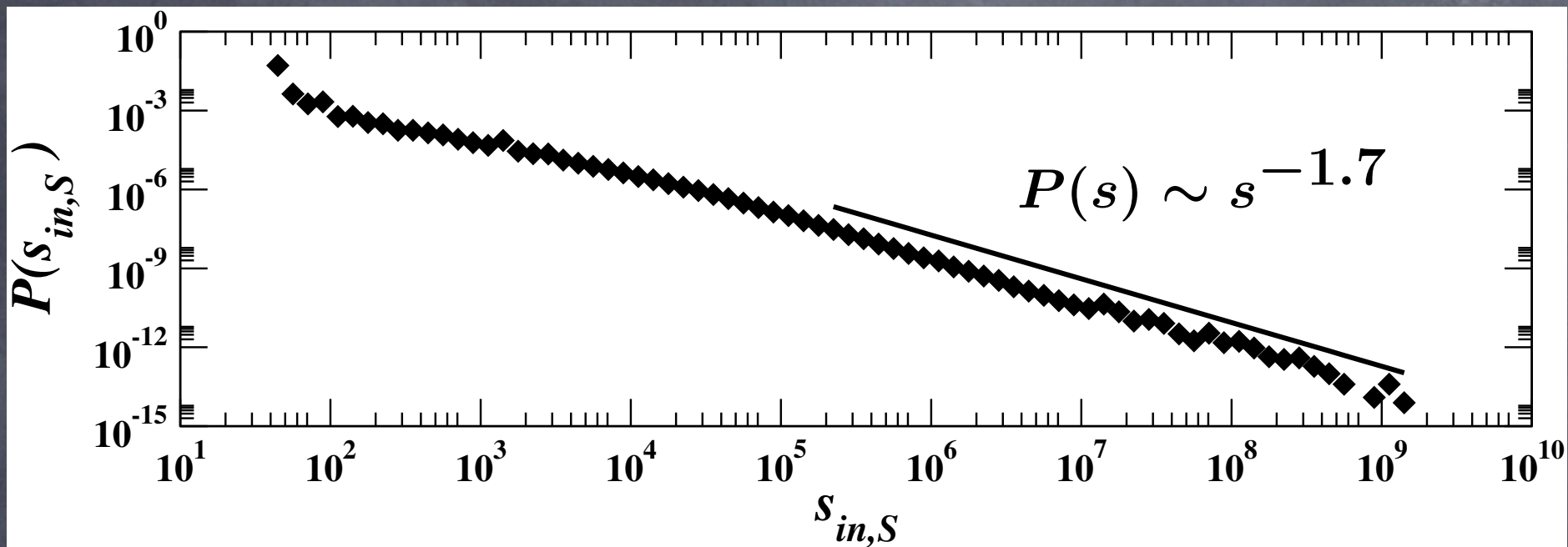
Work with Mark Meiss & Alex Vespignani



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