How does the Web really grow?

Filippo Menczer

School of Informatics

Department of Computer Science

Indiana University

http://informatics.indiana.edu/fil/

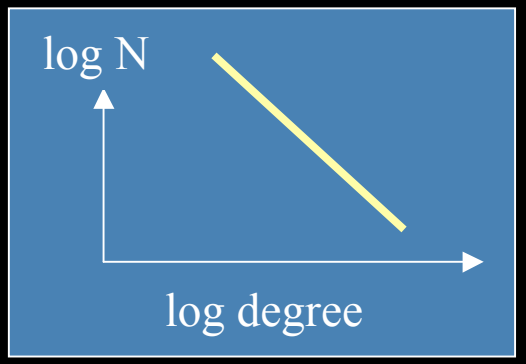
Research supported by NSF
CAREER Award IIS-0133124

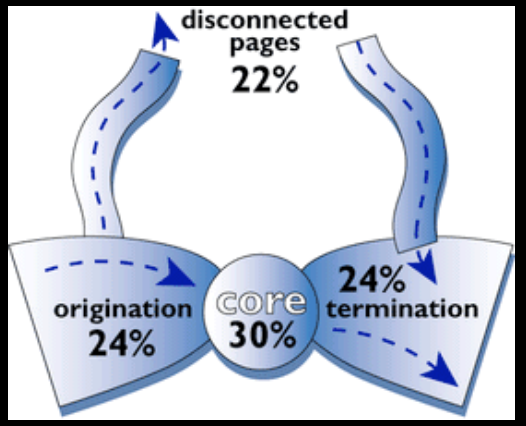
Conclusion

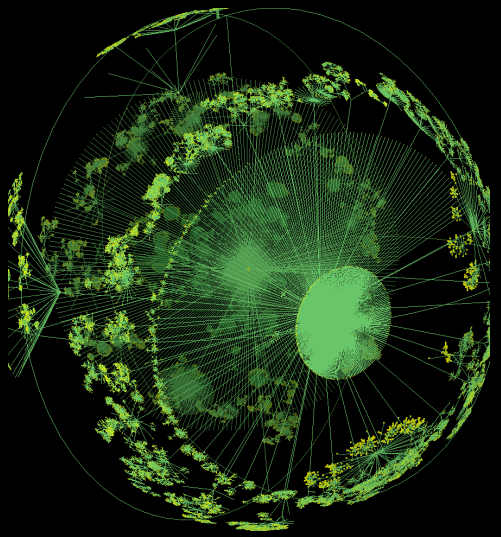
To explain the emergent topology of the Web, we model the cognitive processes behind Web dynamics – content matters!

Characterizing the Web's link structure

- The Web is a small-world: small diameter
  (Barabasi, Albert & al @ ND, Huberman,Adamic & al @ PARC/HP)
- The Web is a bow tie
  (Broder & al @ AltaVista/Compaq/IBM)
- The Web is scale-free: power laws
  (all of the above)
- Subsets of the Web are not scale free
  (Pennock, Lawrence, Giles & al @ NECI)







Web growth ~~models~~ classes

- 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)

$$\Pr(i) \quad k(i)$$

- Modified BA / fitness / hidden variable

  (Bianconi & Barabasi 2001, Adamic & Huberman 2000…)

$$\Pr(i) \quad \square(i)k(i)$$

- Mixture

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

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

- Web copying

  (Kleinberg, Kumar & al 1999, 2000)

$$\Pr(i) \quad \square \cdot \Pr(j \square \ i) + (1 \square \square) \cdot c$$

- Mixture with metric distance

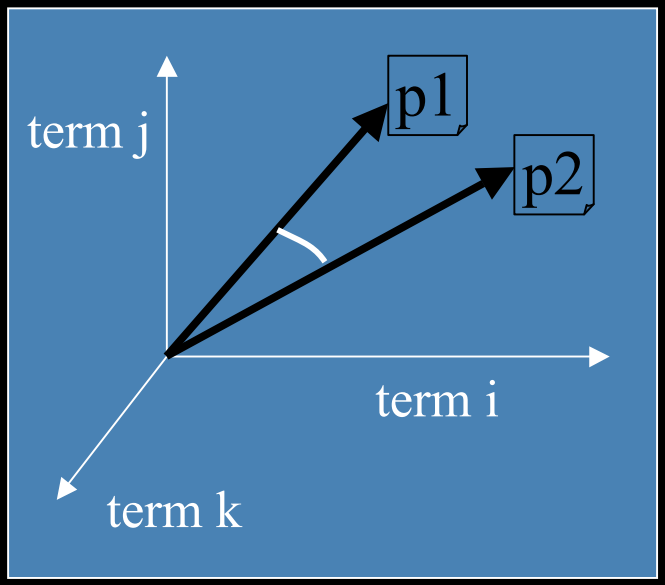
  (Fabrikant, Koutsoupias & Papadimitriou 2002,
  Aldous 2003)

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

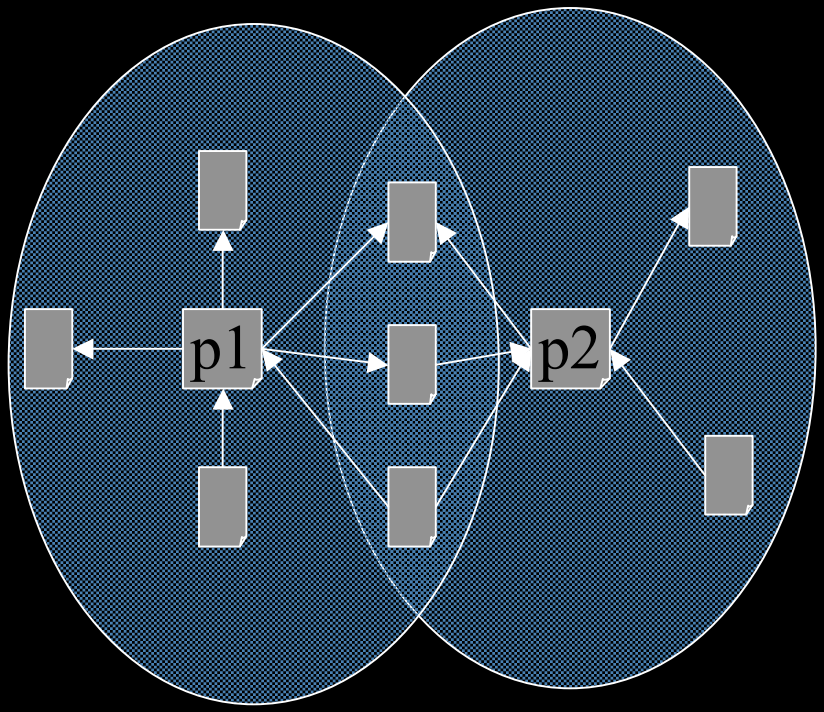
Mapping the relationship between topologies

- Given any pair of pages, need 'similarity' or 'proximity' metric for each topology:
  - Content: textual/lexical (cosine) similarity
  - Link: co-citation/bibliographic coupling
- 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

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



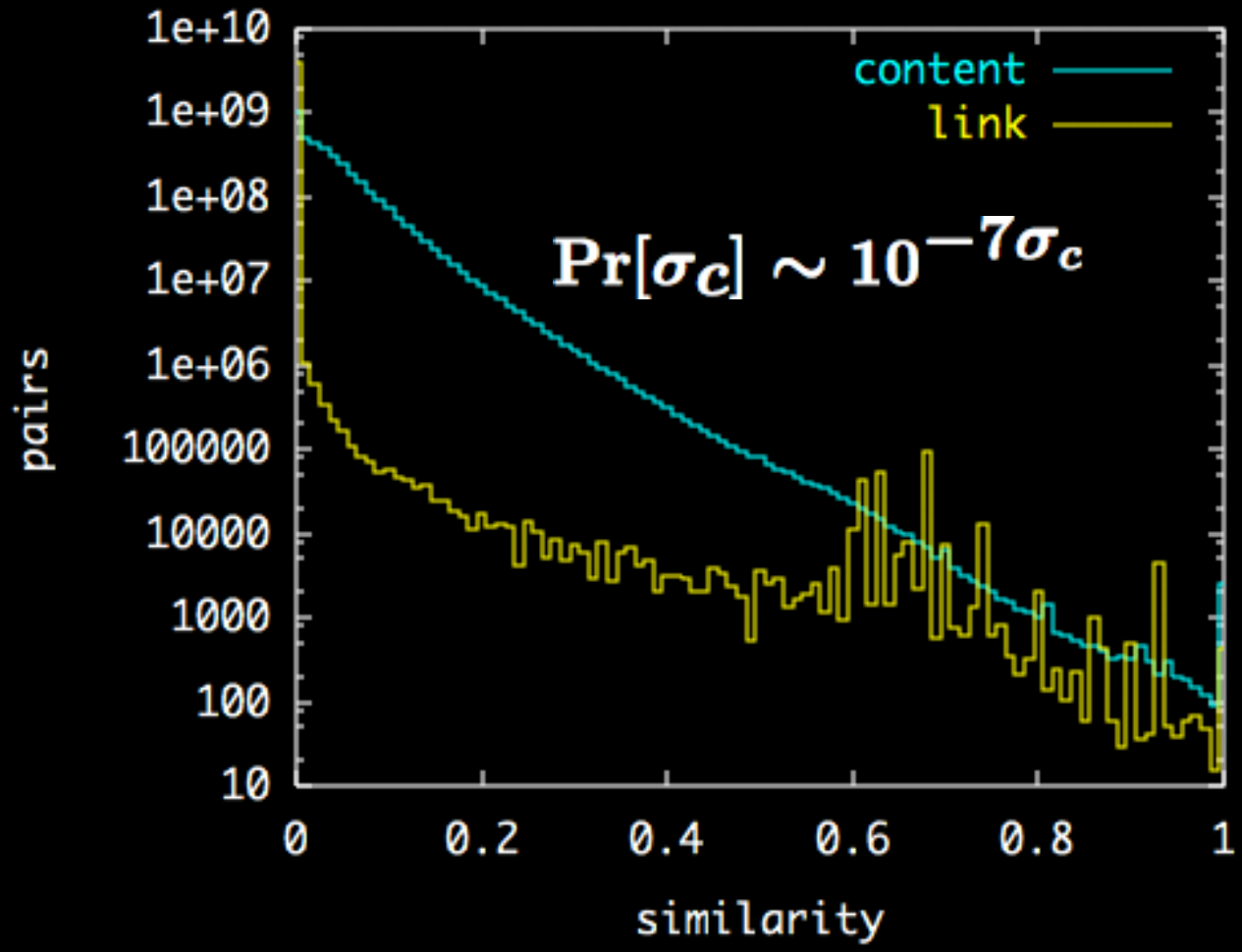
Content similarity

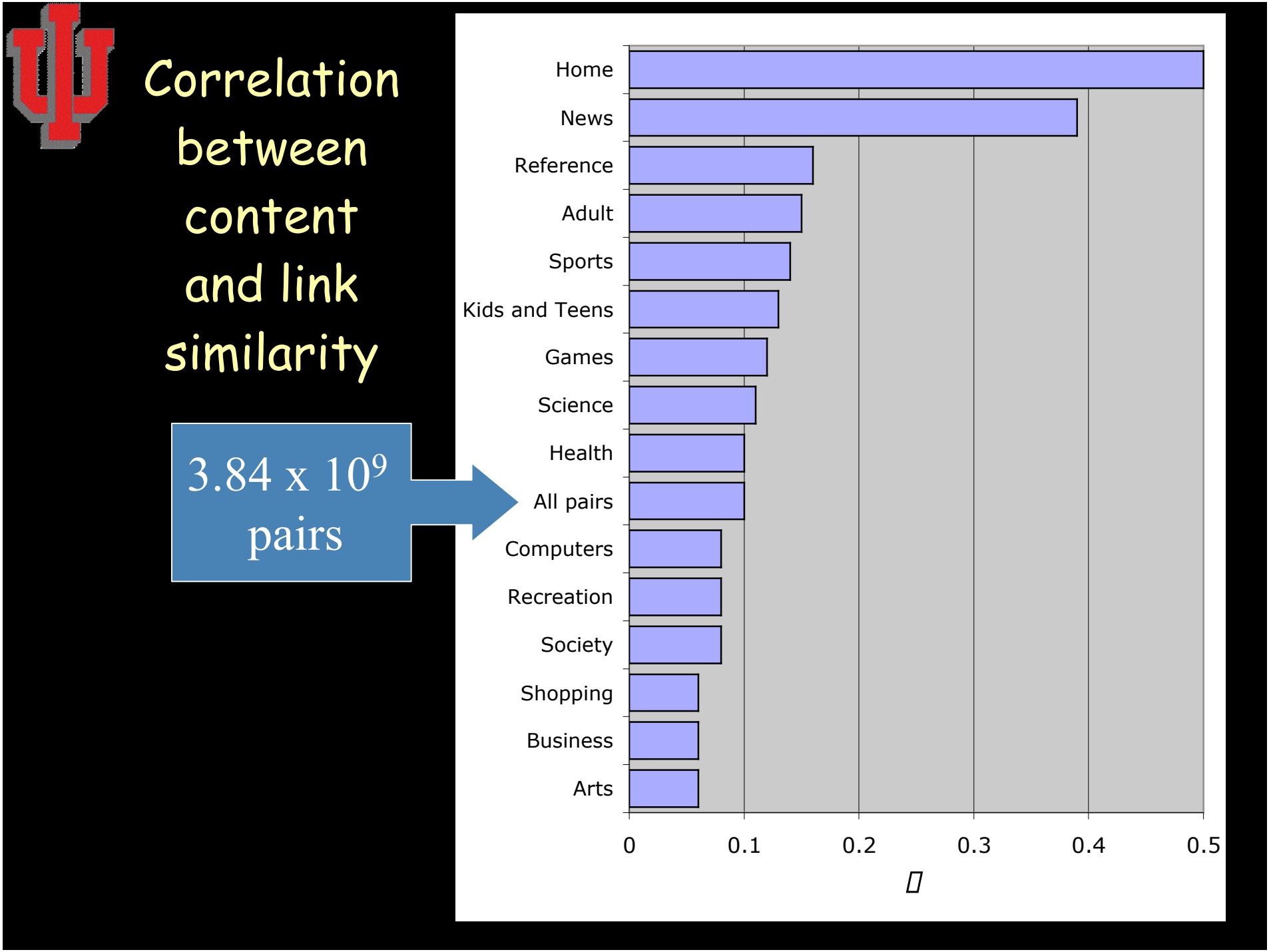


Link similarity

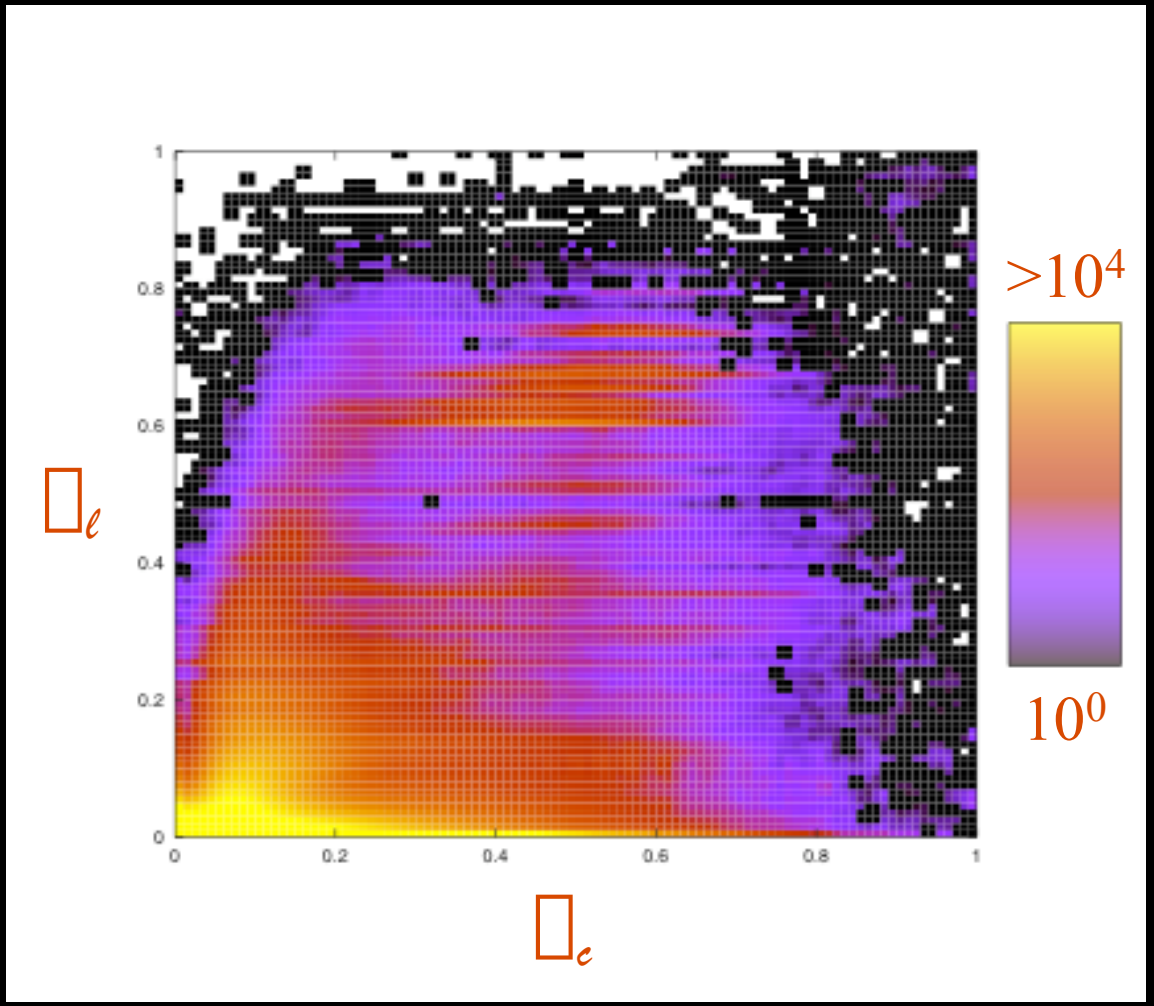
$$\sigma_l(p_1, p_2) = \frac{\left|U_{p_1} \cap U_{p_2}\right|}{\left|U_{p_1} \cup U_{p_2}\right|}$$

Individual metric distributions




Correlation between content and link similarity
3.84 x 10^9 pairs
Home
News
Reference
Adult
Sports
Kids and Teens
Games
Science
Health
All pairs
Computers
Recreation
Society
Shopping
Business
Arts
0
0.1
0.2
0.3
0.4
0.5

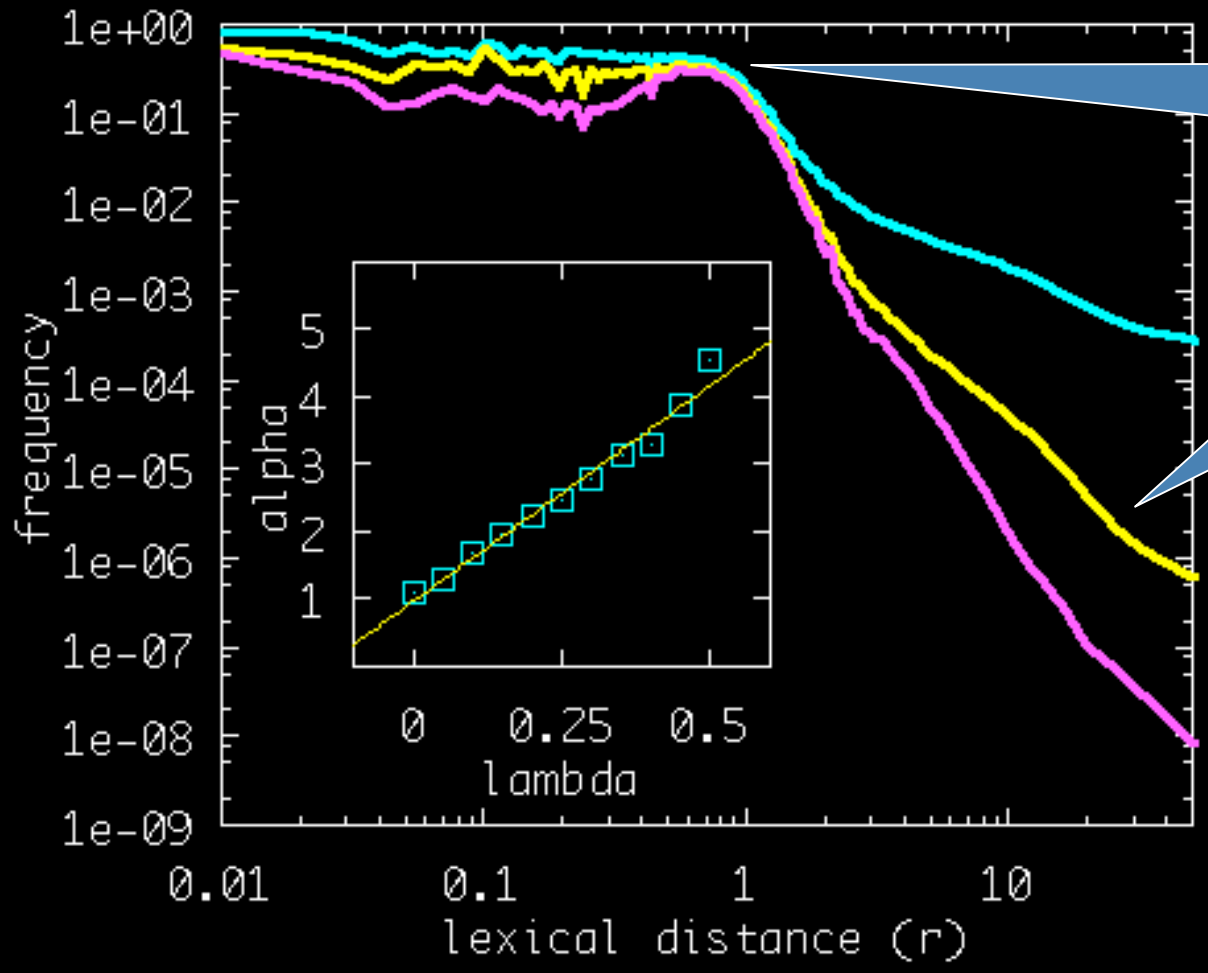
Joint distribution map



Link probability vs lexical distance

$$r = 1/\lambda_c \quad \lambda \geq 1$$

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



Phase transition λ^*

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

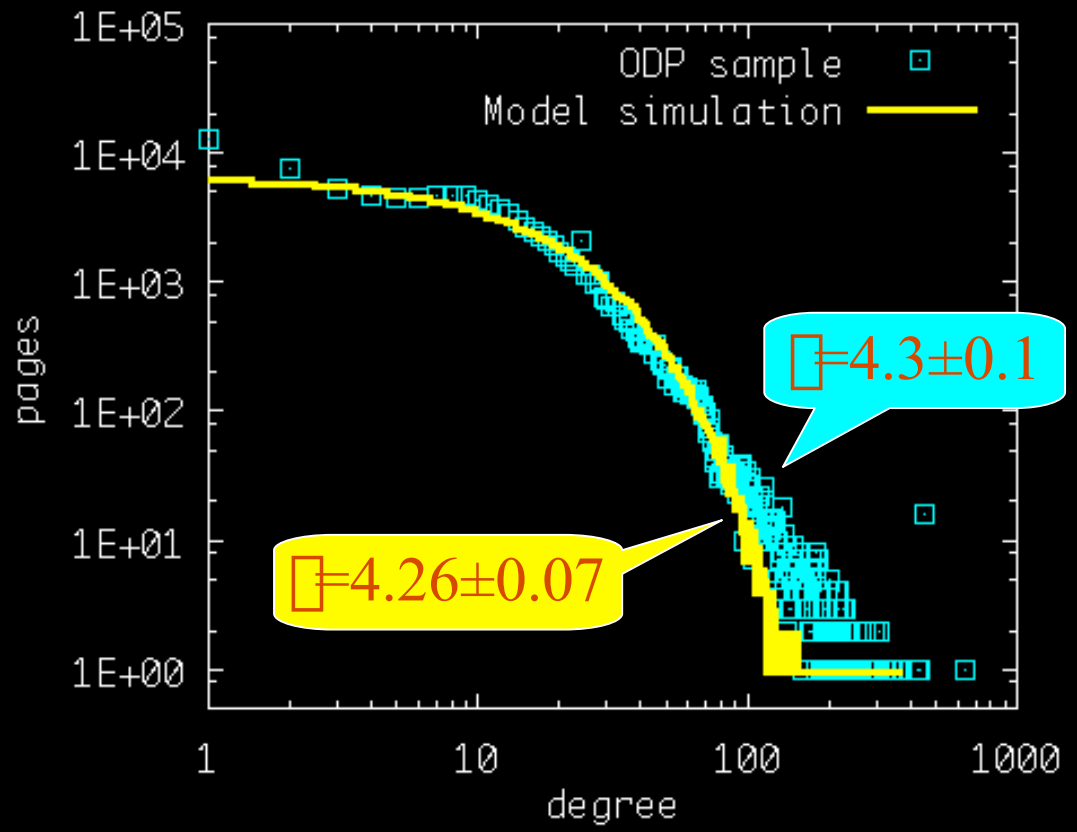
Proc. Natl. Acad. Sci. USA 99(22): 14014-14019, 2002

Local content-based growth model

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

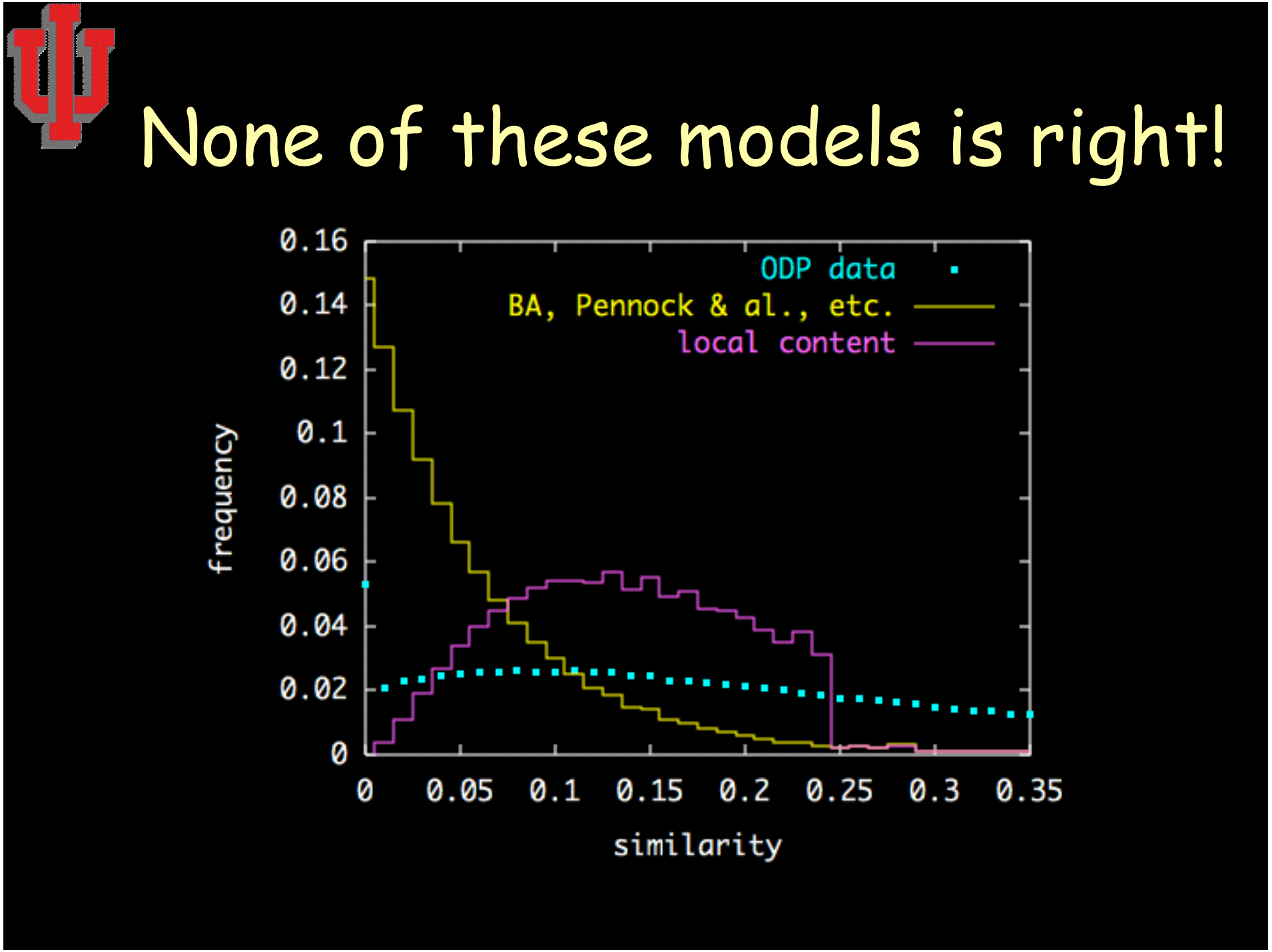
- Similar to preferential attachment (BA)
- At each step t add page p_t
- Create m new links from p_t to existing pages
- Use degree (k) info only for nearby pages

  (popularity/importance of similar/related pages)



So, many models get the degree distribution right

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



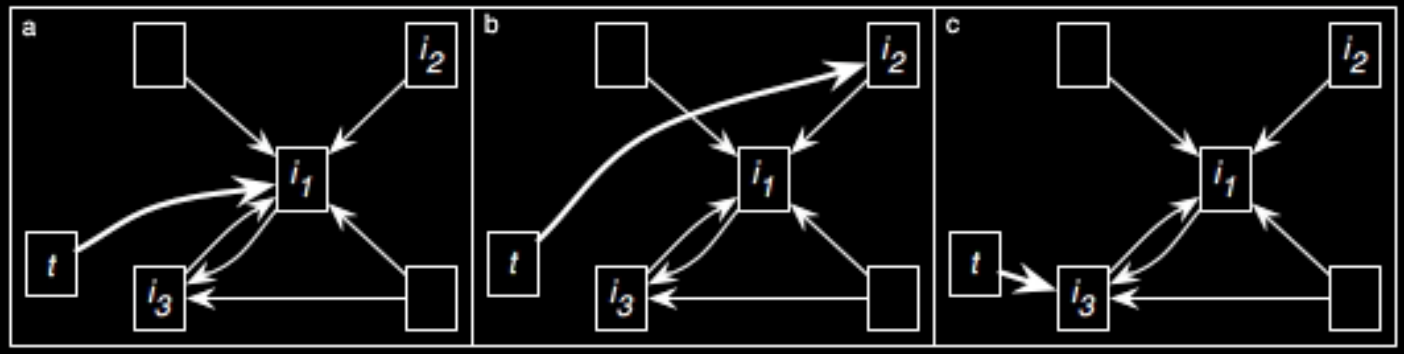
None of these models is right!

Back to the mixture model

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

- Pennock & al. 2002
- Cooper & Frieze 2001
- Dorogovtsev & al 2000

- Extra free parameter
  - Can tune popularity with "random"



  - Bias choice by content similarity instead of uniform distribution

Mixture model class

$$\Pr(i) = \psi \frac{k(i)}{2mt} + (1 - \psi)\hat{\Pr}(i)$$

- Degree-uniform mixture:

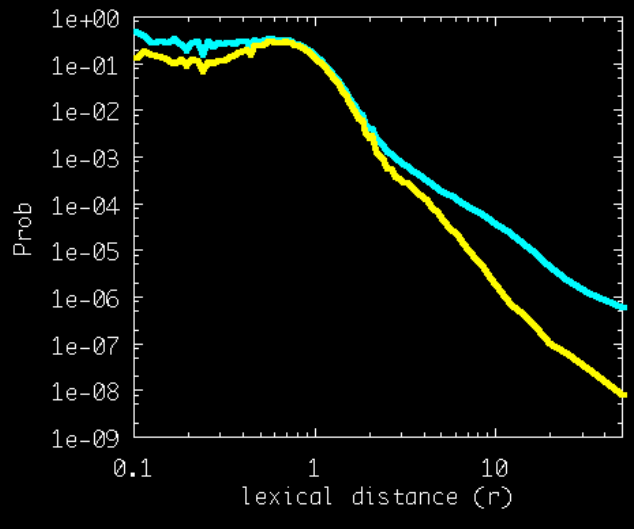
$$\hat{\Pr}(i) = \frac{1}{t}$$

$$\psi = 0.3$$

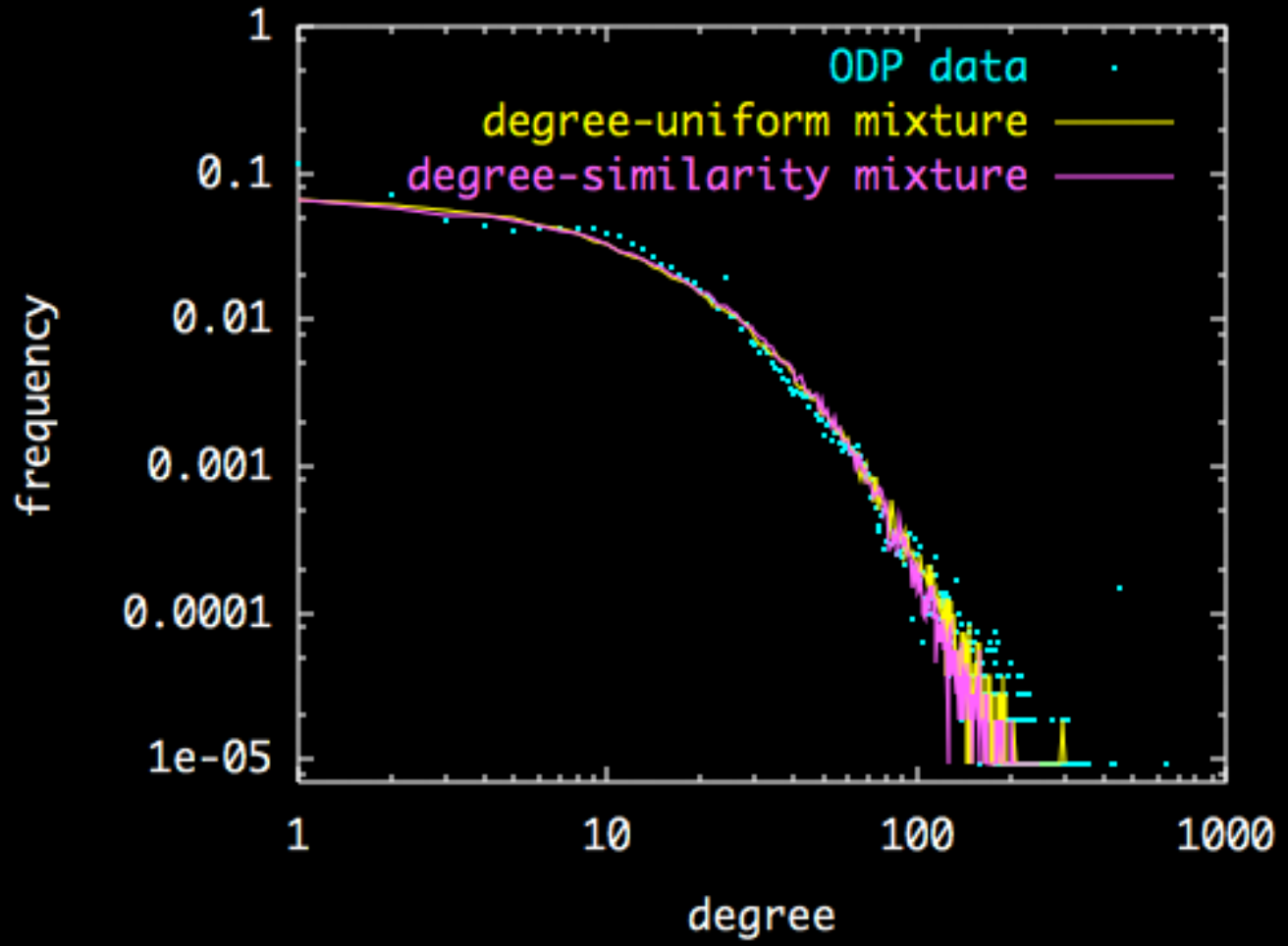
- Degree-similarity mixture:

$$\hat{\Pr}(i) \propto \left(\frac{1}{\sigma_c(i,t)} - 1 \right)^{-\alpha}$$

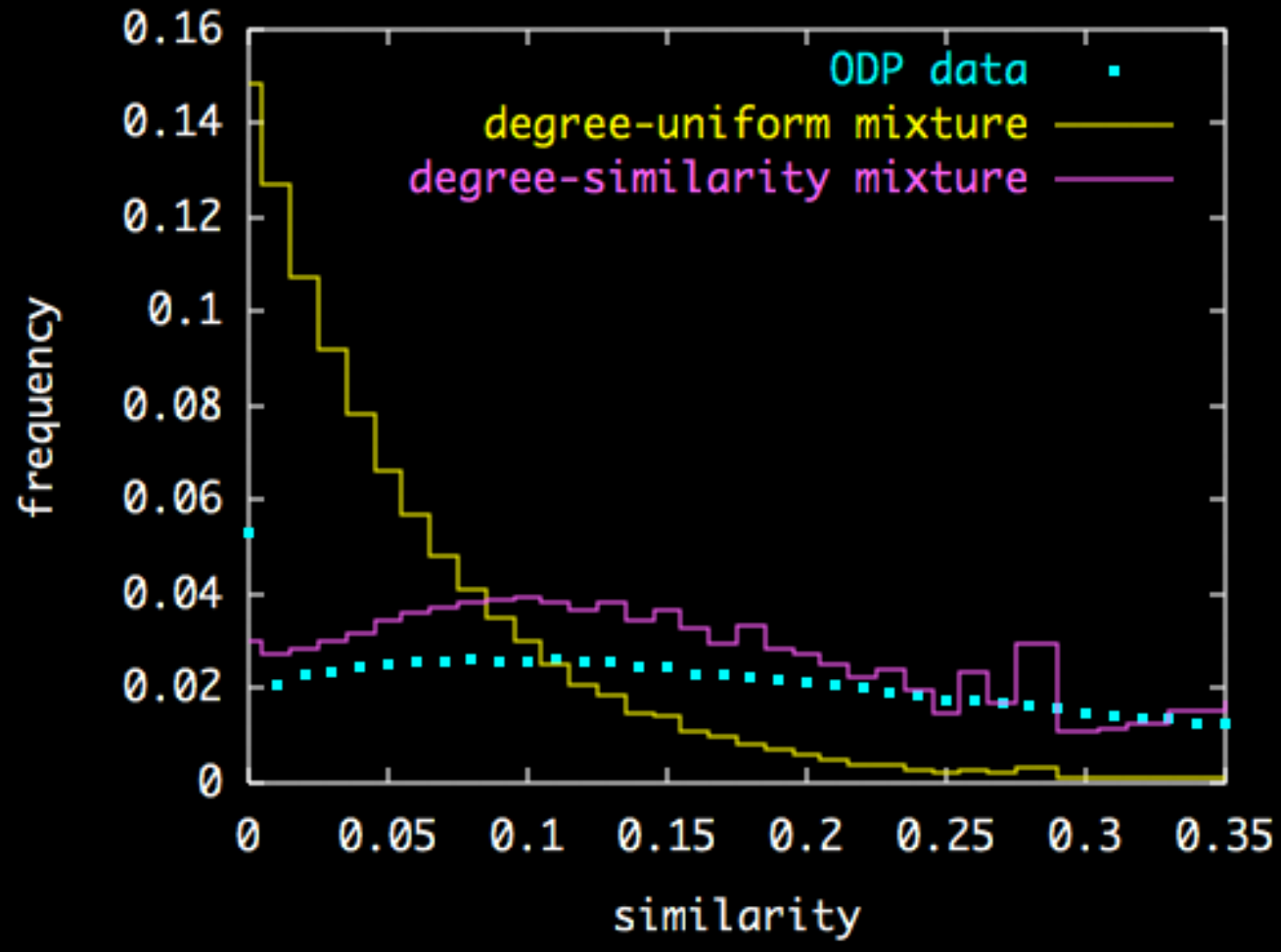
$$\psi = 0.2,\ \alpha = 1.7$$



Both mixture models get the degree distribution right…



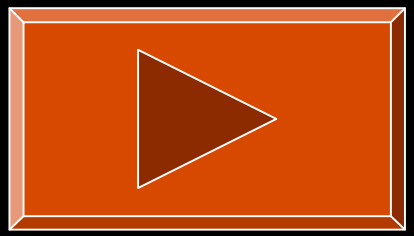
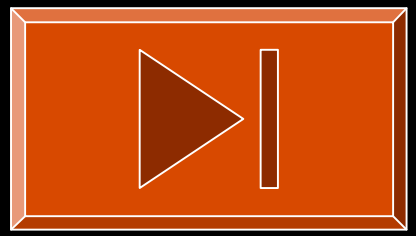
…but the degree-similarity mixture model predicts the similarity distribution much better



Proc. Natl. Acad. Sci. USA, forthcoming

Conclusion

- To explain the emergent topology of the Web, we model the cognitive processes behind Web dynamics – content matters!
- What about paper citation networks?

Questions?

http://informatics.indiana.edu/fil/

Research supported by NSF
CAREER Award IIS-0133124