

# INFO 368

# Introduction to network science

## Filippo Menczer

---



SCHOOL OF Informatics and Computing

- **Review**

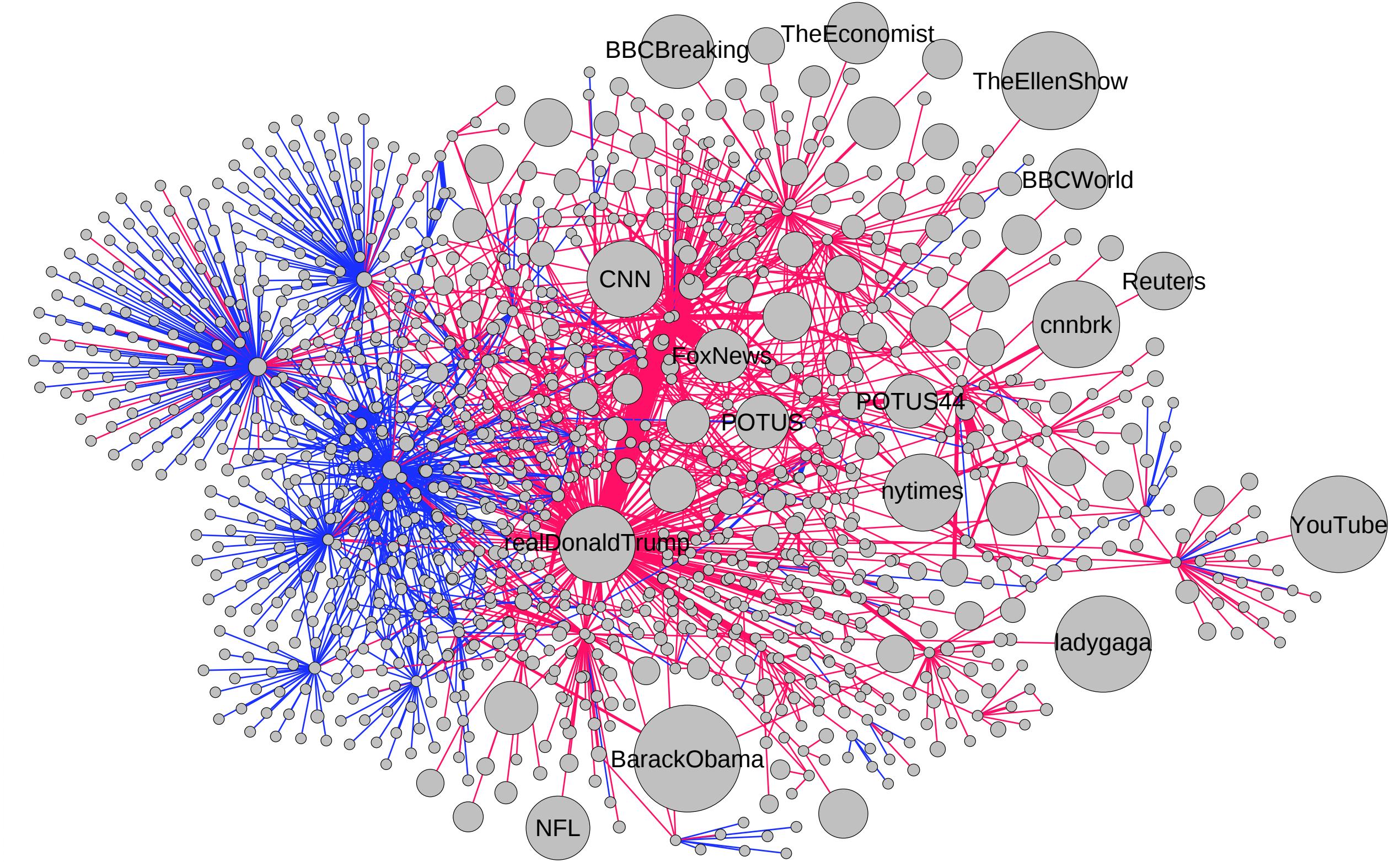
- SIR vs threshold models of diffusion
- Network role in spreading processes
- Halting an epidemic
- Selective immunization

# Plan for next 4 weeks

- \* Information networks
- \* Web structure
- \* Homophily and topical locality
- \* Web search engines
- \* PageRank
- \* Web traffic networks
- \* Search in networks

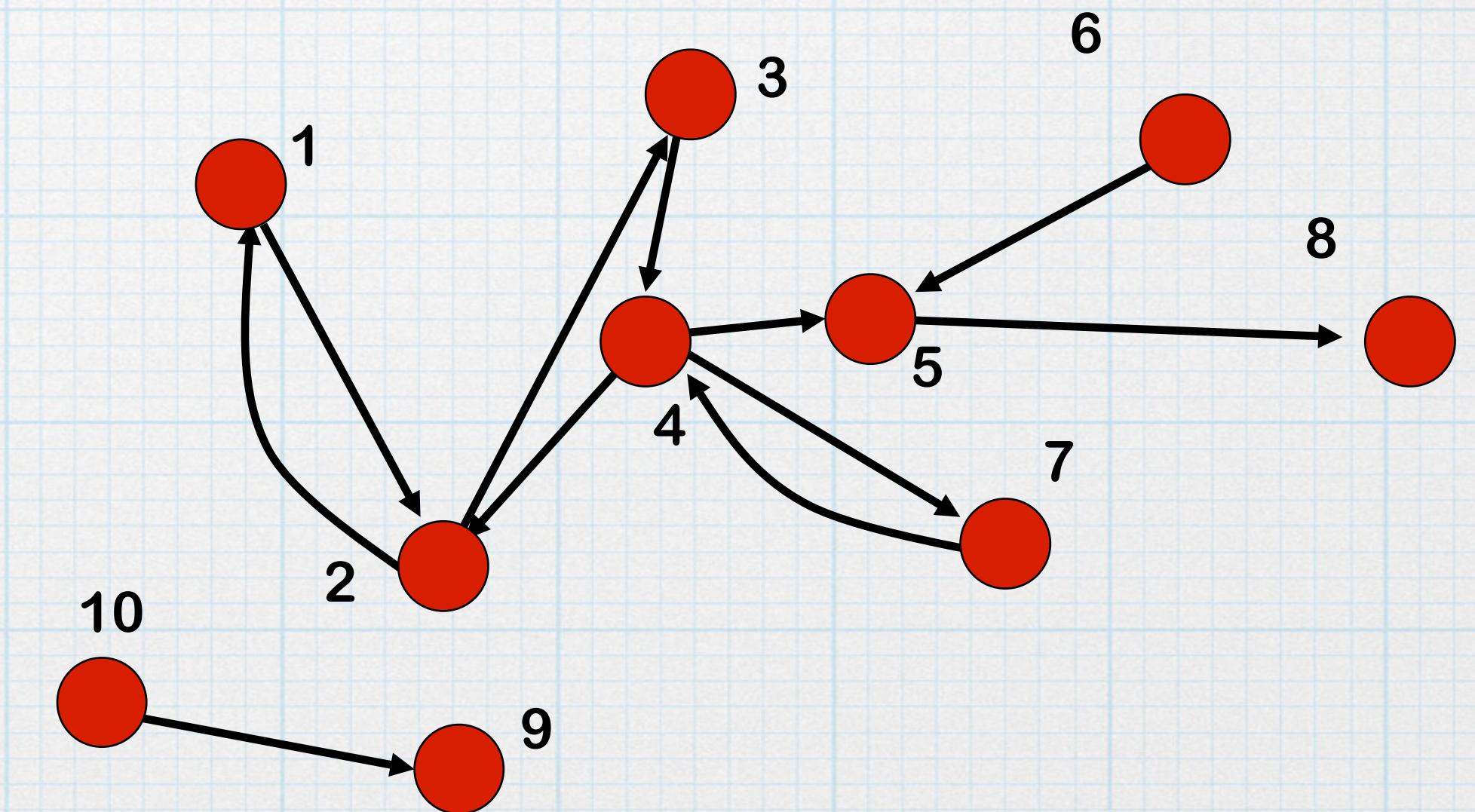
# Review: weighted networks

- \* Each link has a weight
- \* Examples?
- \* How do we extend the notion of degree to the weighted case?
- \* How can we statistically characterize a weighted network?
- \* Name a couple centrality measures

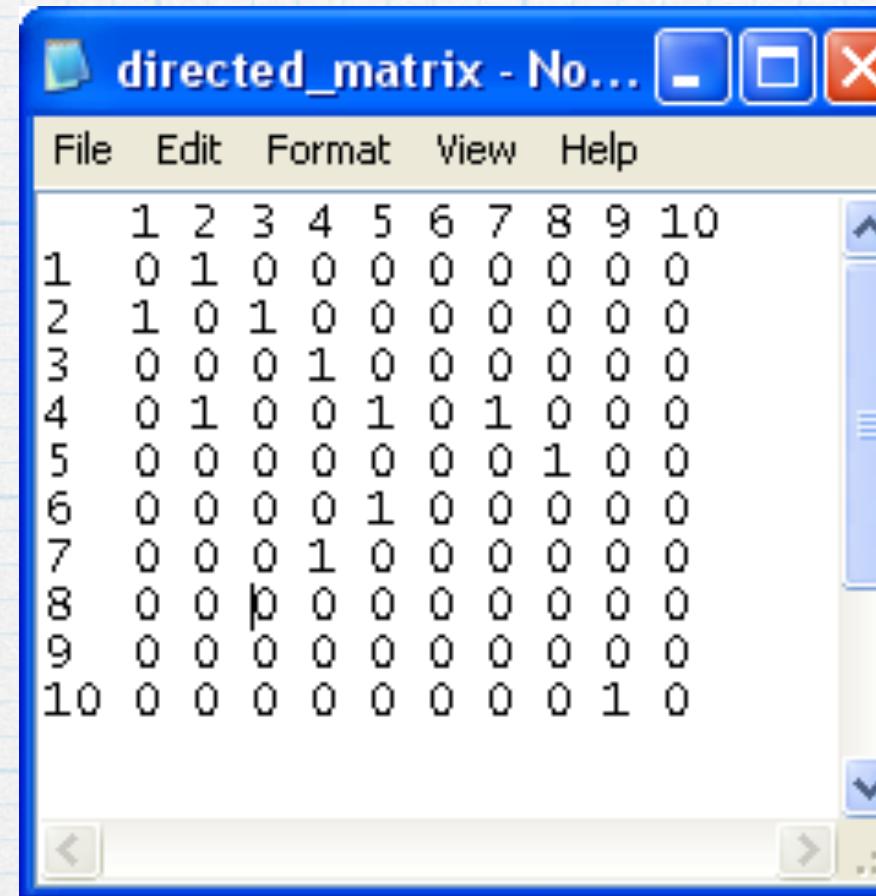


# Review: directed networks

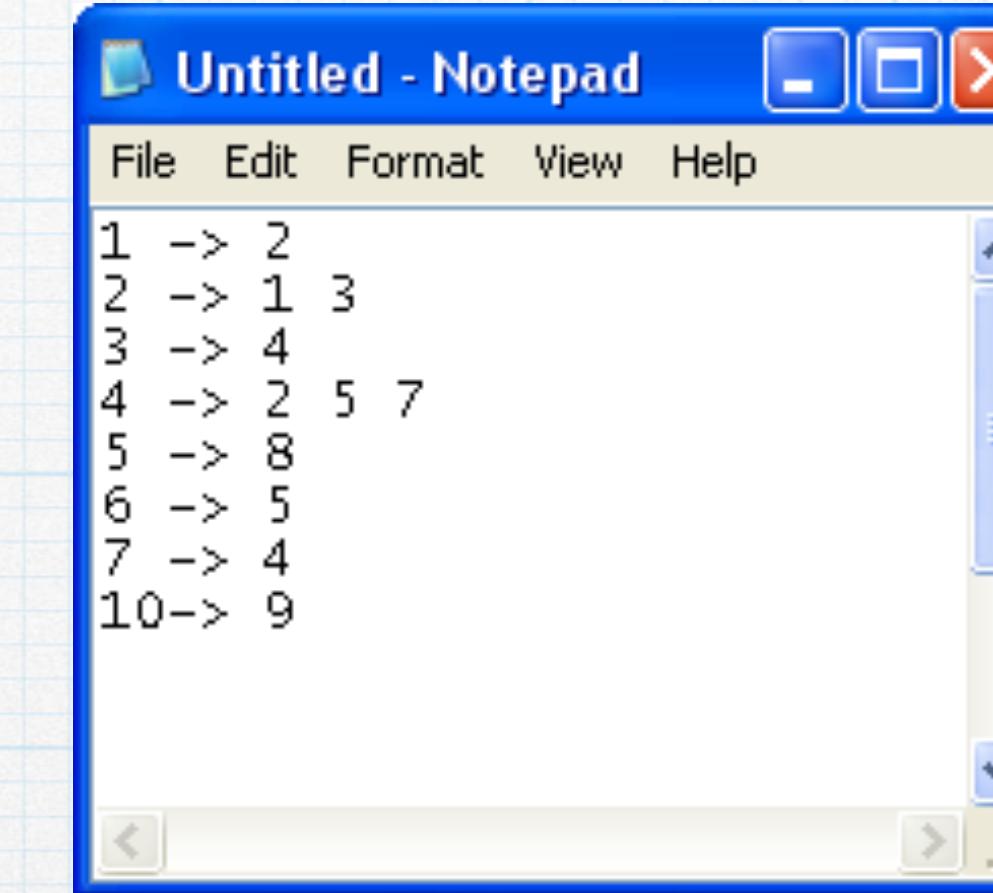
- \* Each link has a source and a destination
- \* Examples?
- \* How do we extend the notion of degree to the directed case?
- \* Name a couple centrality measures



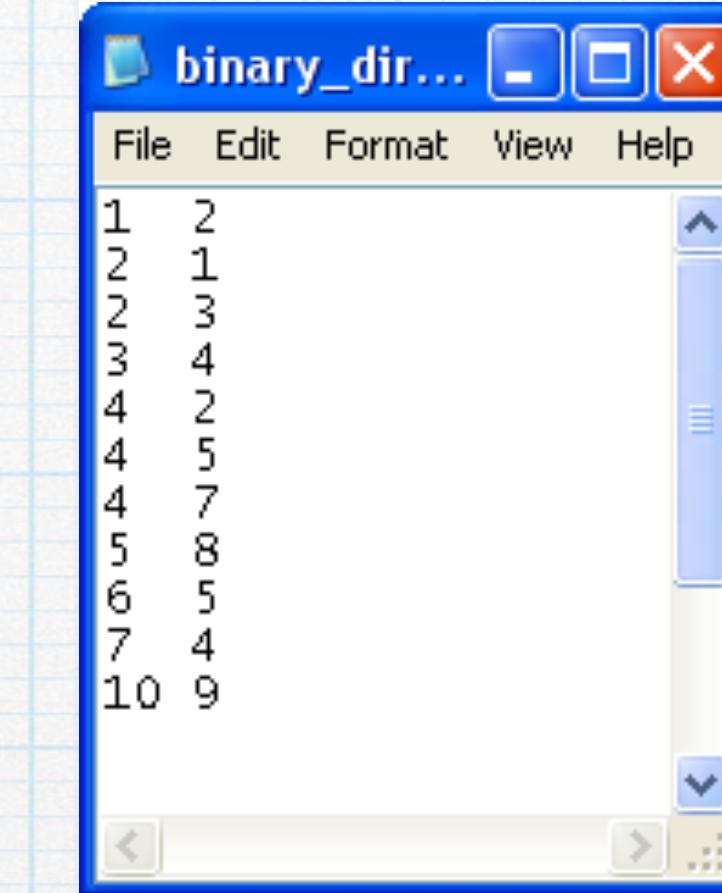
# Directed networks



1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	0	0	0	0	0
2	1	0	1	0	0	0	0	0	0
3	0	0	0	1	0	0	0	0	0
4	0	1	0	0	1	0	1	0	0
5	0	0	0	0	0	0	1	0	0
6	0	0	0	0	1	0	0	0	0
7	0	0	0	1	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	1

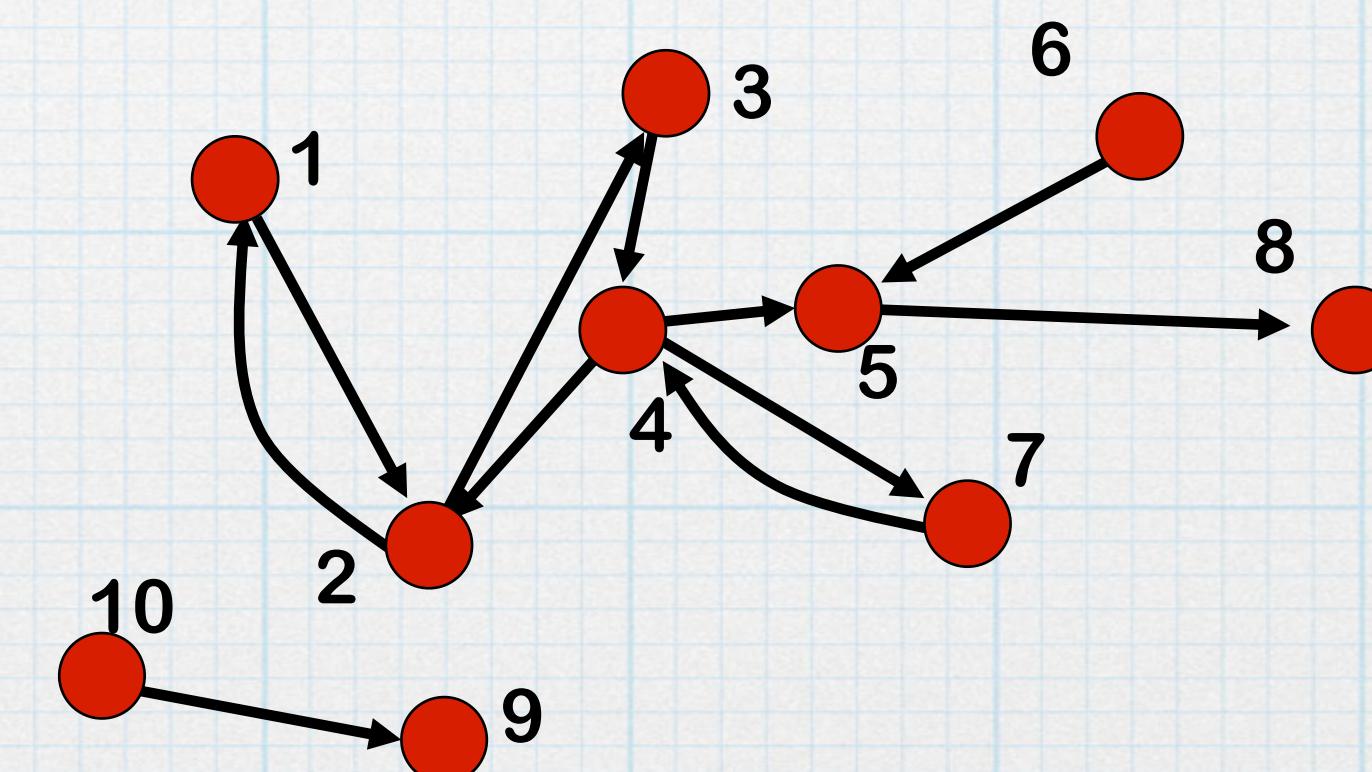


```
1 -> 2
2 -> 1
3 -> 4
4 -> 2
4 -> 5
4 -> 7
5 -> 8
6 -> 5
7 -> 4
10-> 9
```



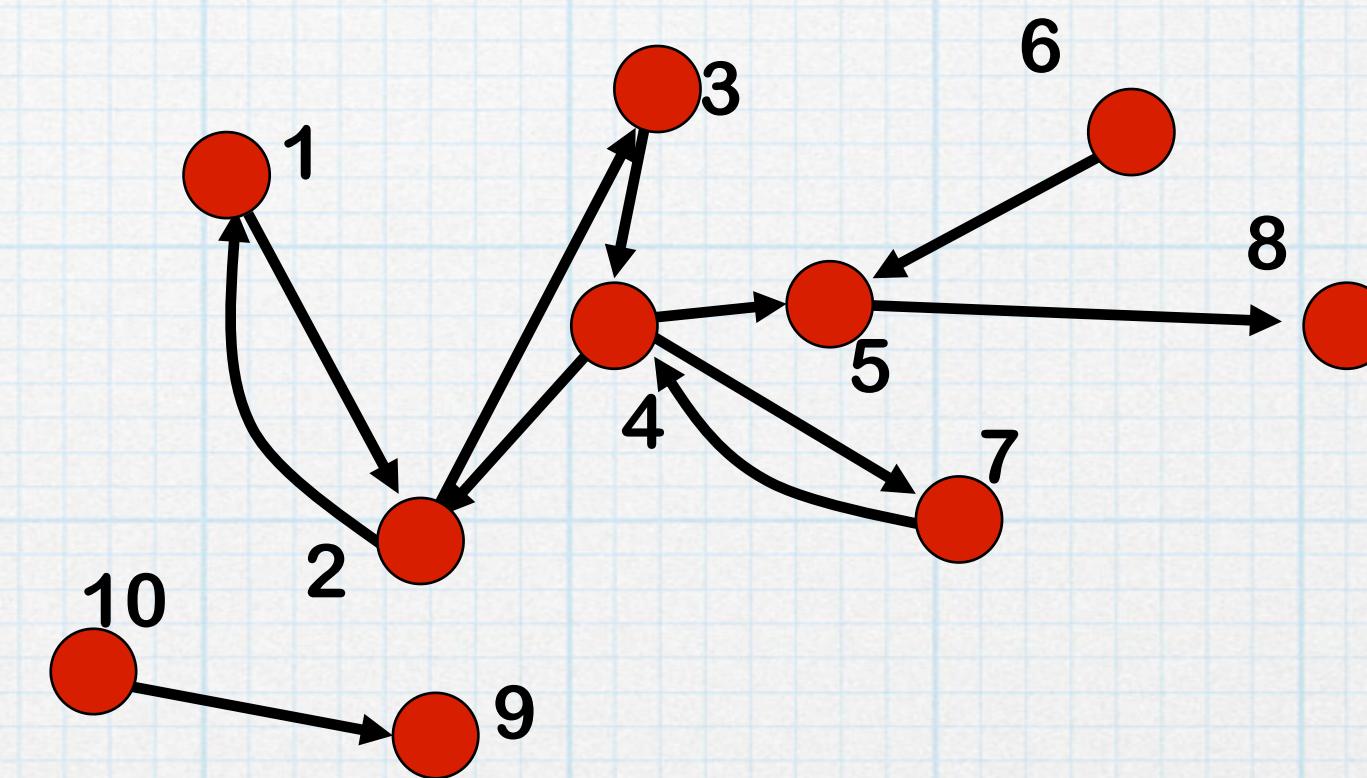
```
1 2
2 1
2 3
3 4
4 2
4 5
4 7
5 8
6 5
7 4
10 9
```

- \* Representations:
  - \* Adjacency matrix
  - \* Adjacency list
  - \* Edge list



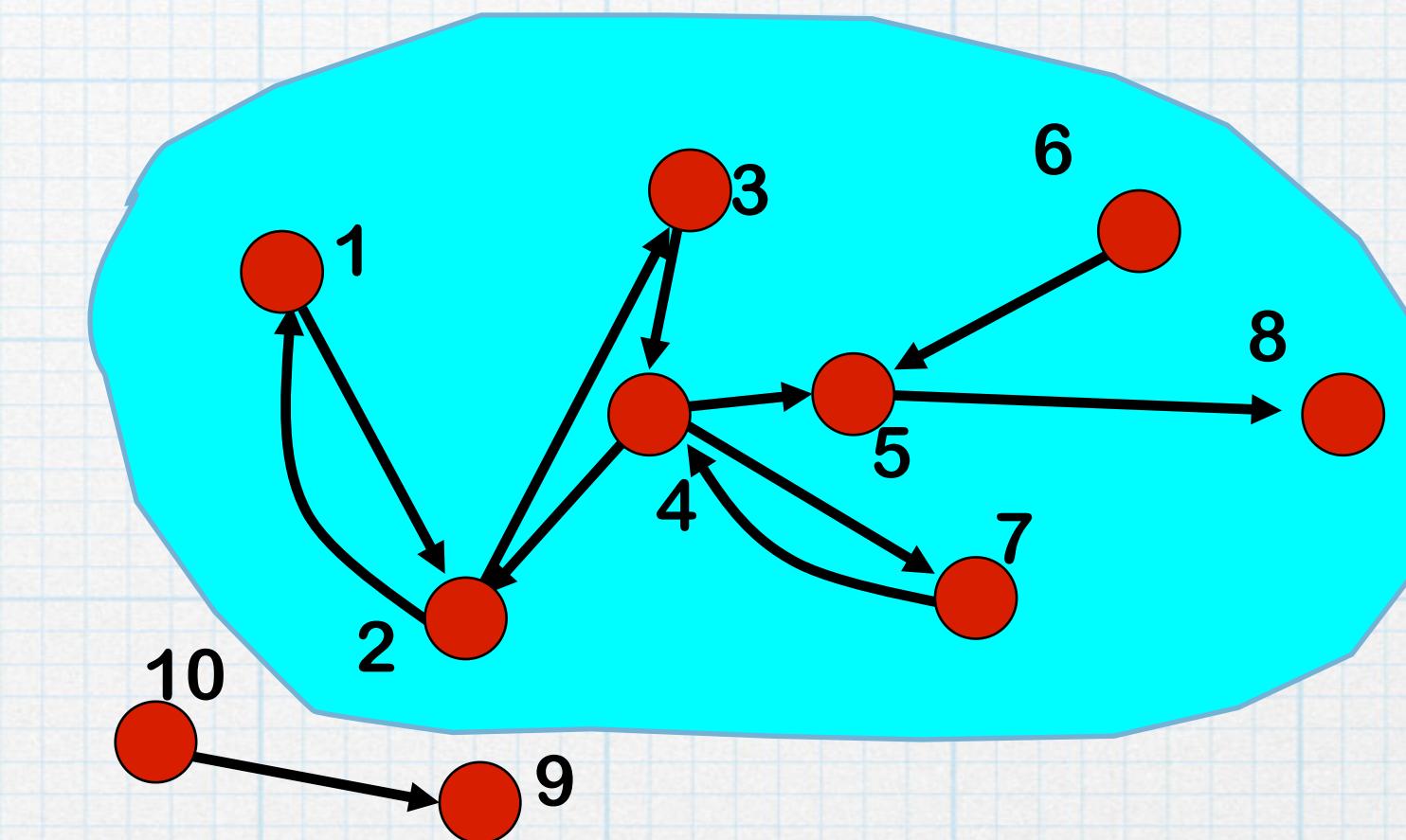
# Directed networks

- \* Paths are directed
- \* There may be a path from A to B but not from B to A
- \* Strongly and weakly connected components



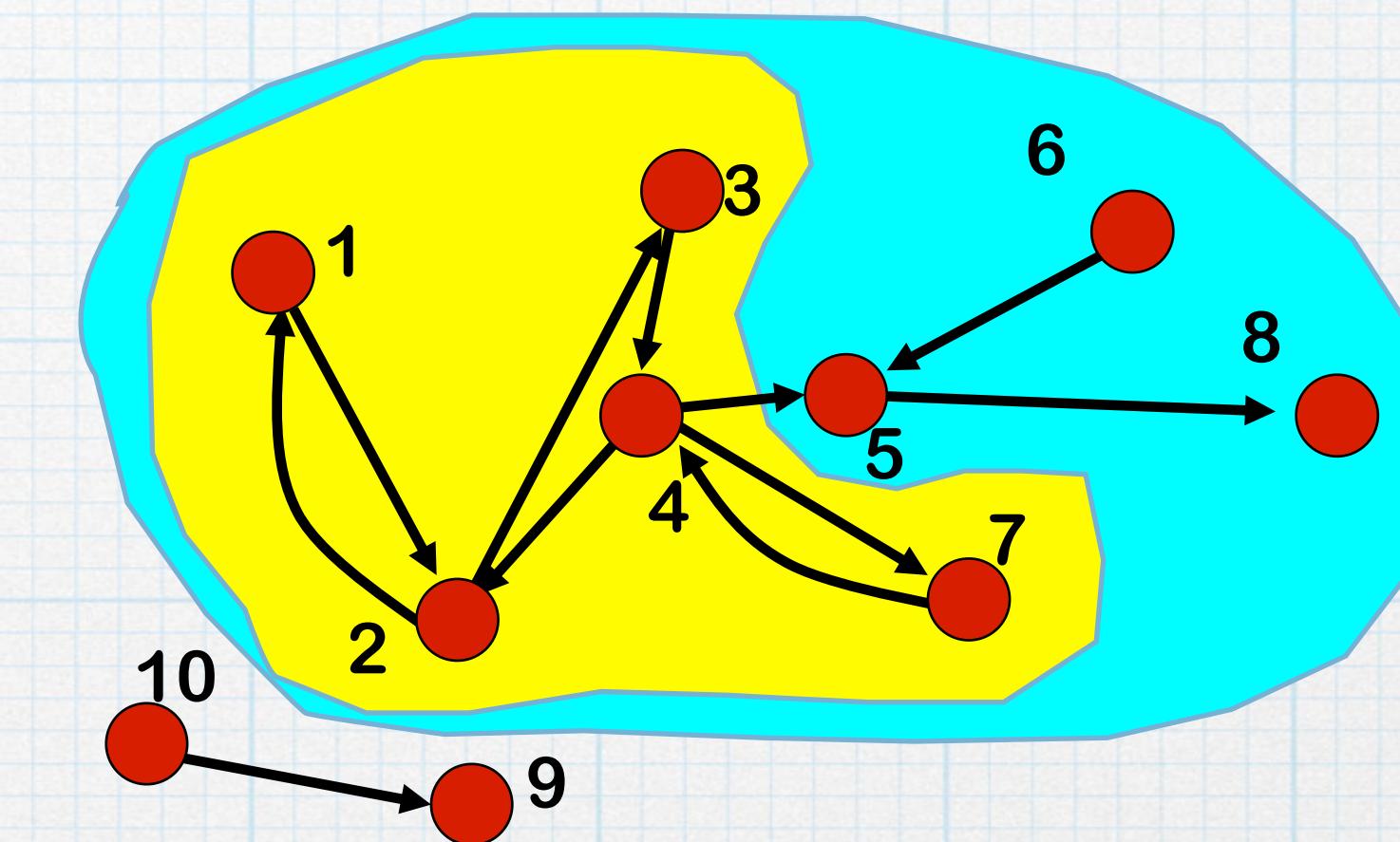
# Directed networks

- \* Paths are directed
- \* There may be a path from A to B but not from B to A
- \* Strongly and weakly connected components



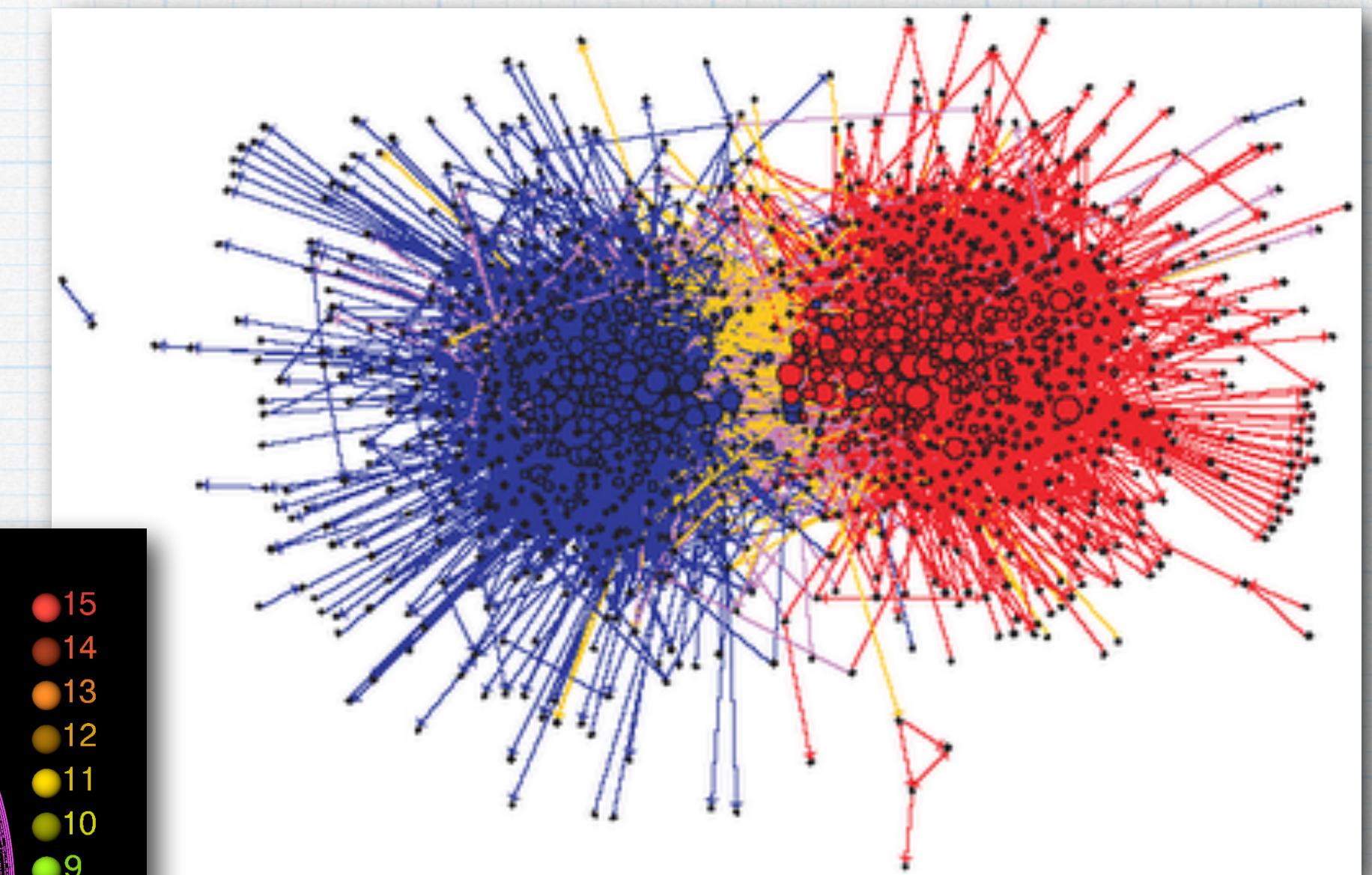
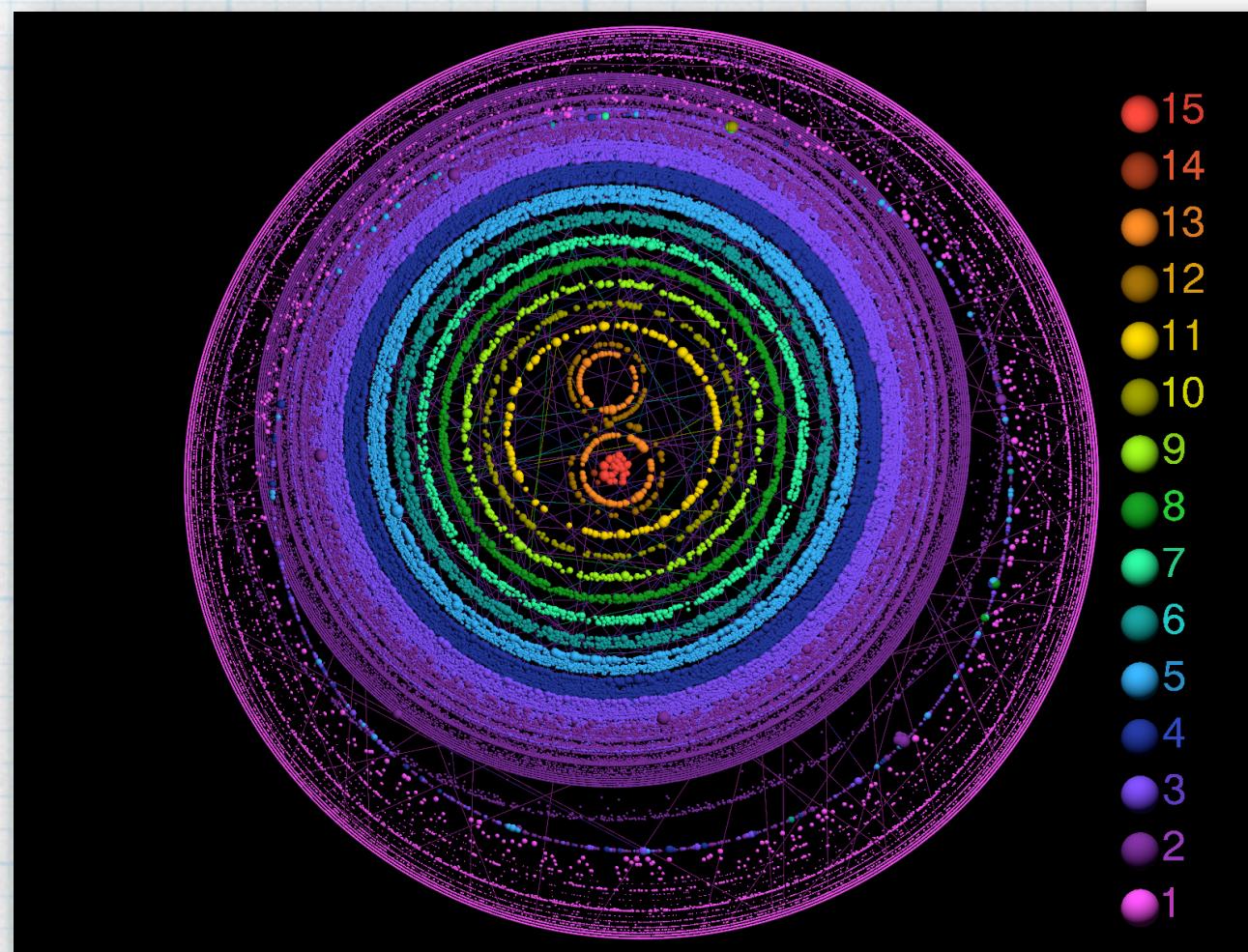
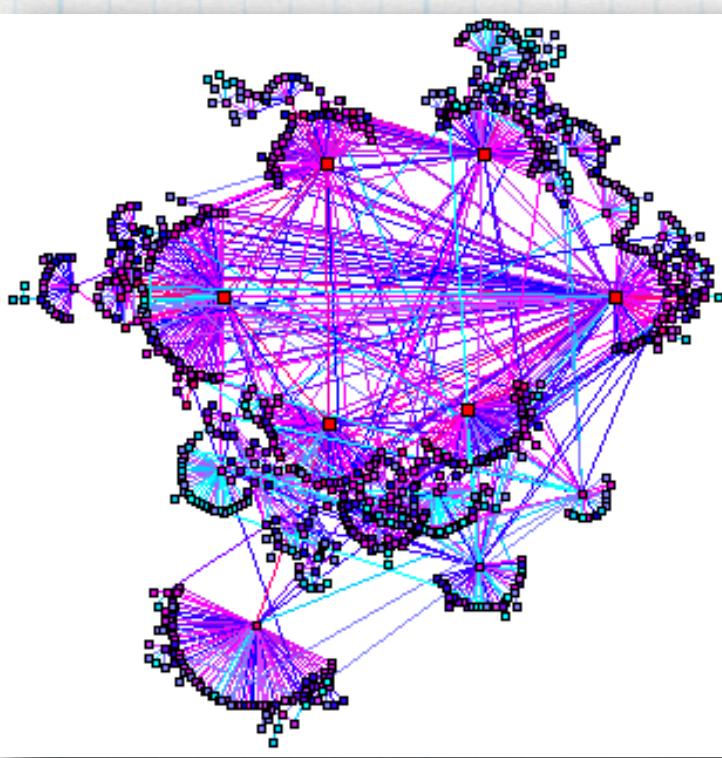
# Directed networks

- \* Paths are directed
- \* There may be a path from A to B but not from B to A
- \* Strongly and weakly connected components

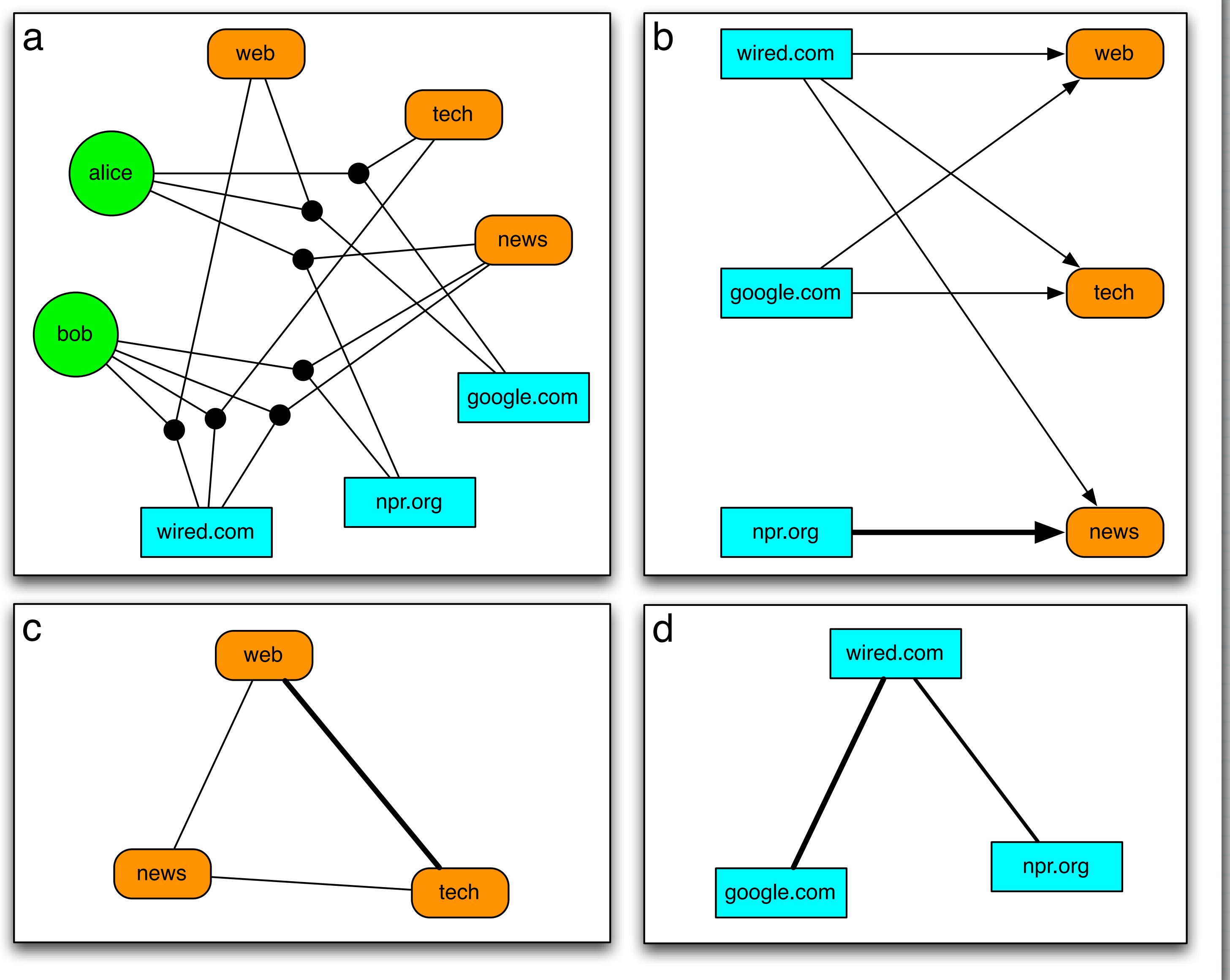


# Information networks: Web, Wikipedia, blogs...

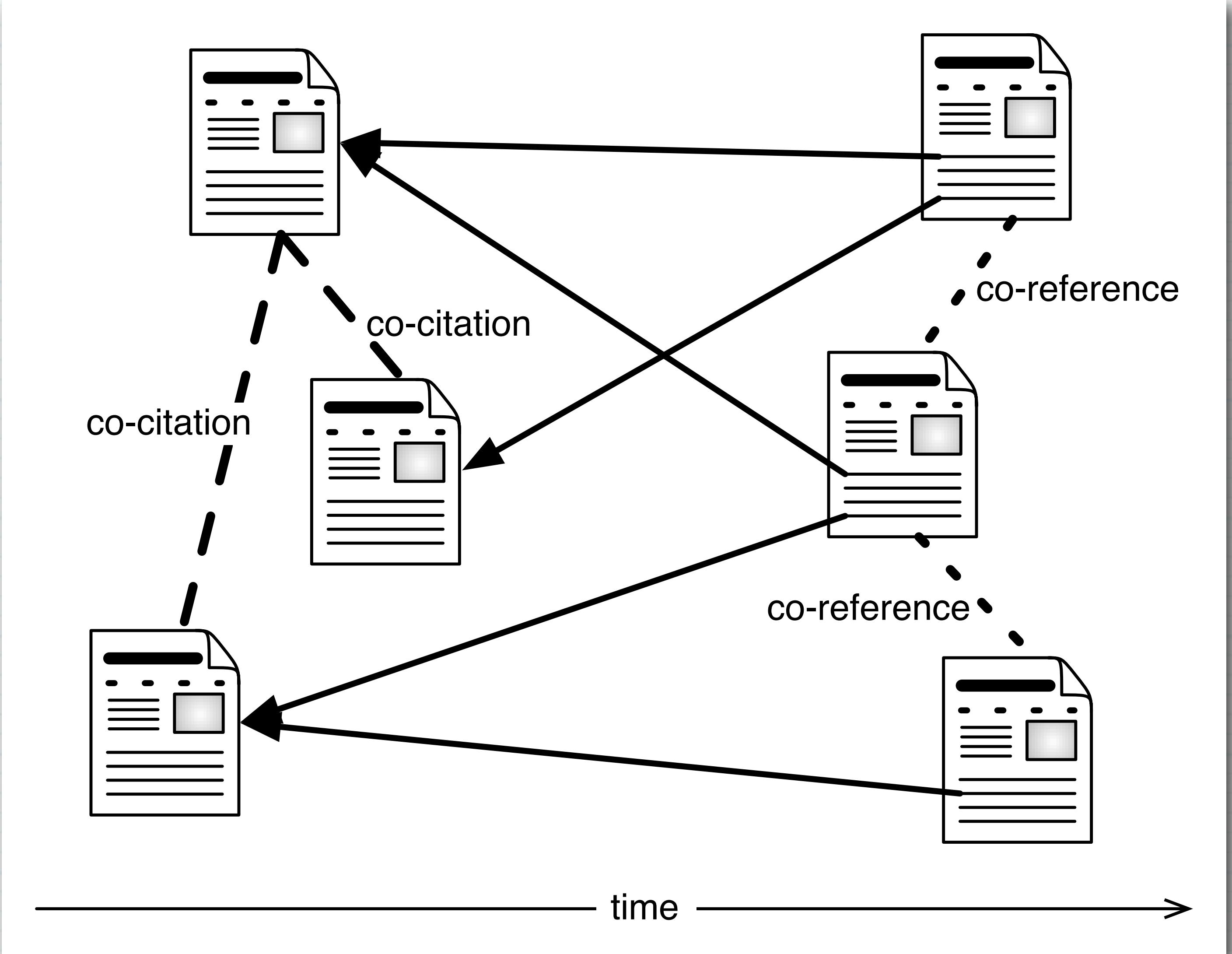
- \* What are the nodes?
- \* What are the edges?
- \* How do they form?



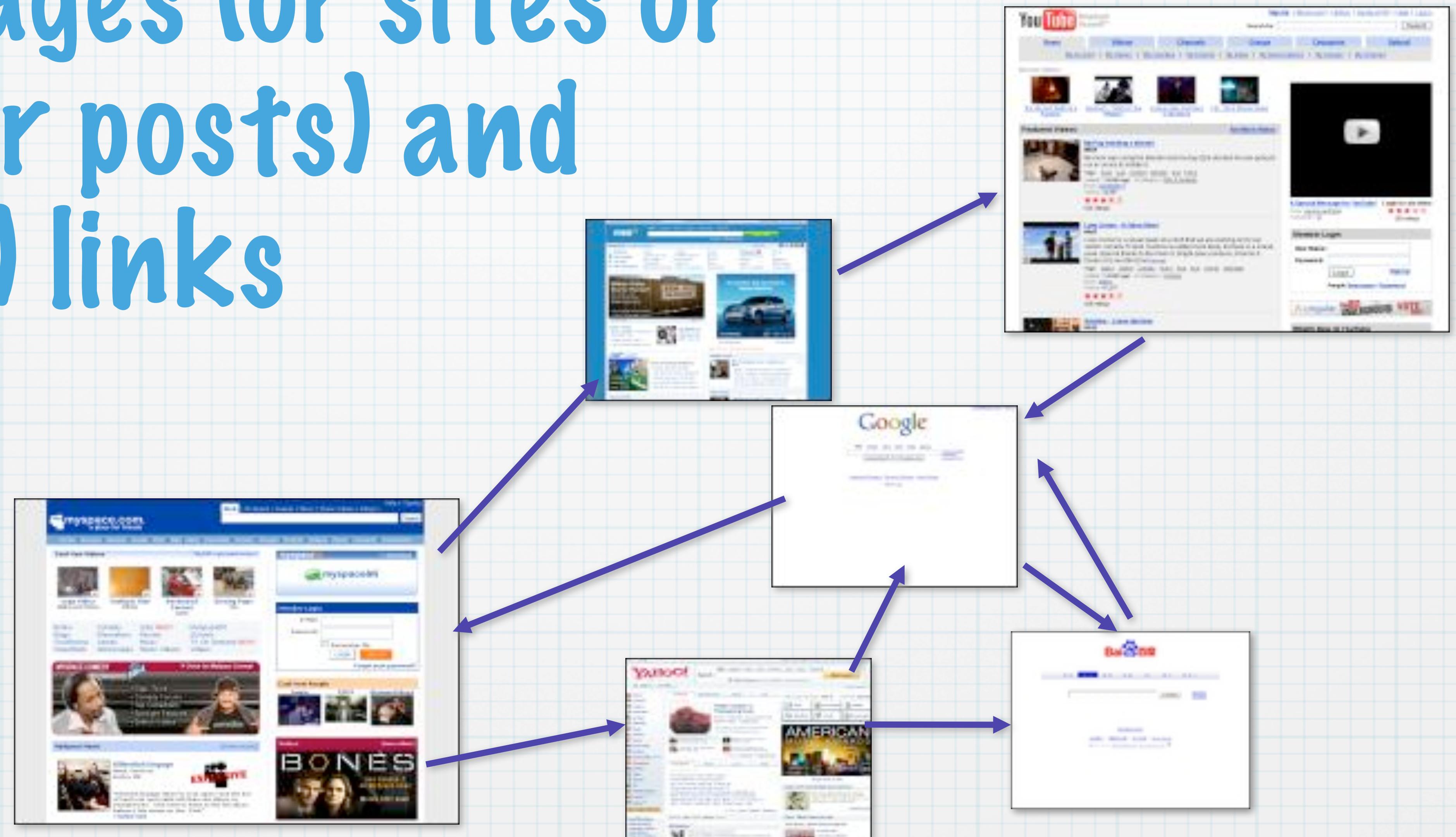
# Folksonomies & tag networks



# Citations



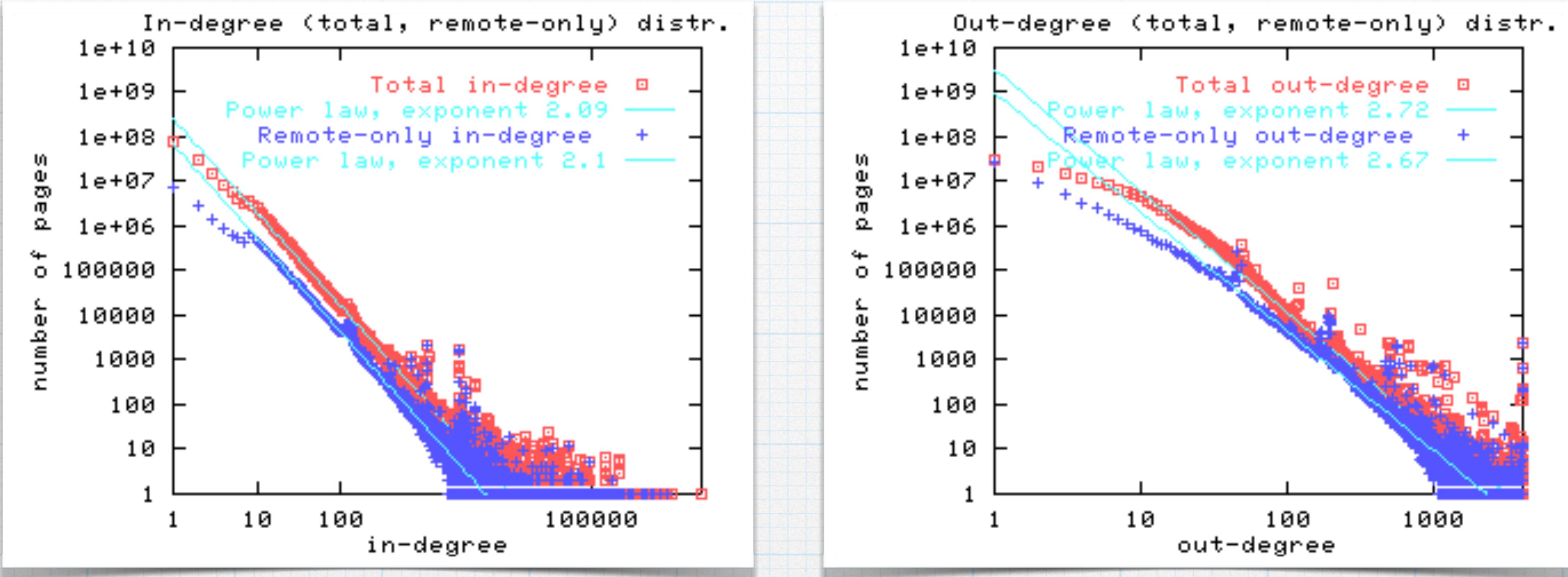
# Web pages (or sites or blogs or posts) and (hyper) links



# How big is the Web?

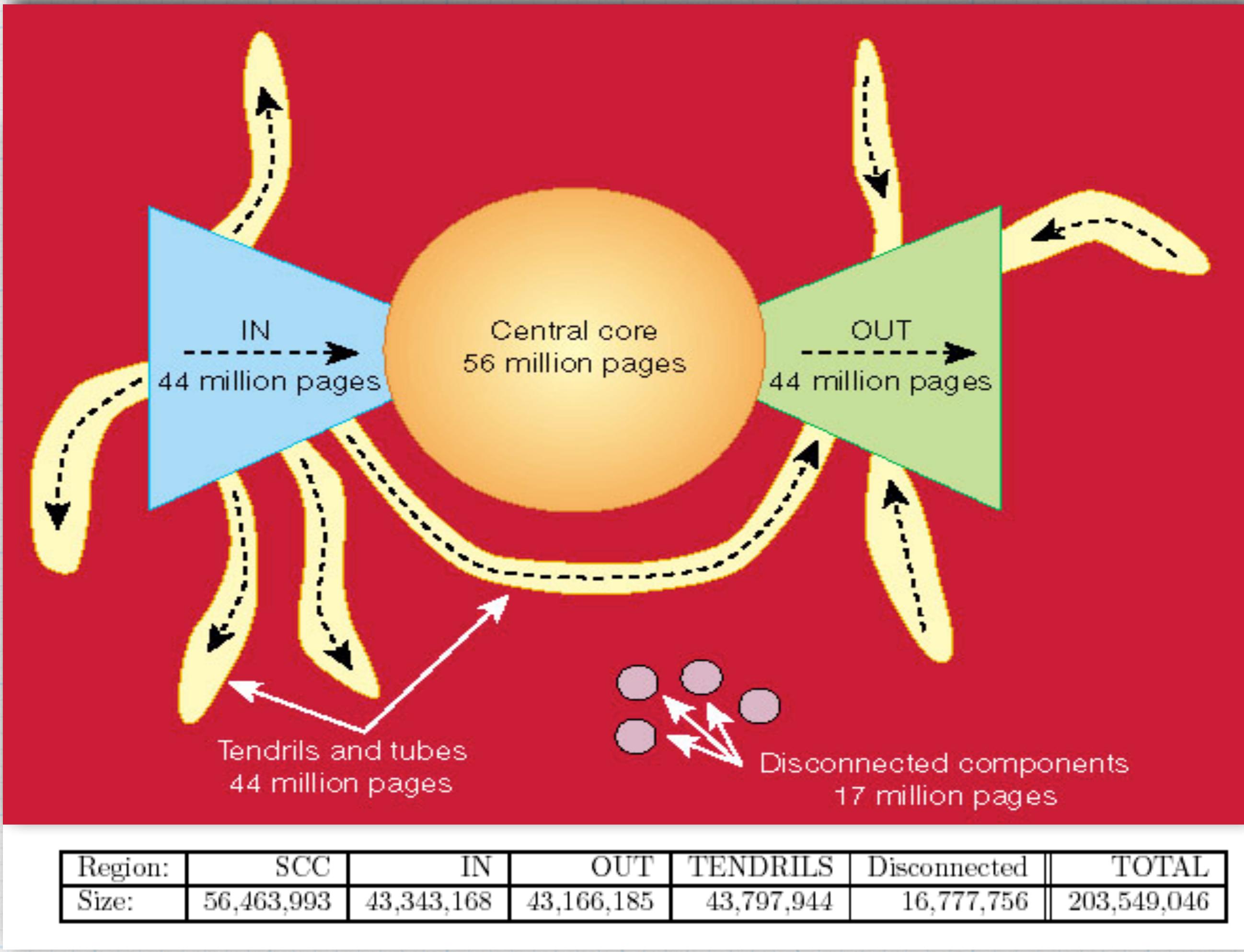
- \* Lawrence & Giles 1998: 320 millions (search engine sampling)
- \* Lawrence & Giles 1999: 800 millions (IP sampling)
- \* Caveat: “Static” Web, “indexable” Web...
- \* Later: 5 billion? 9 billion? 11 billion? 20 billion?
- \* Question no longer makes sense
  - \* Search engines stopped reporting # pages indexed
  - \* Dynamic pages
  - \* Hidden Web

# Scale free Web structure



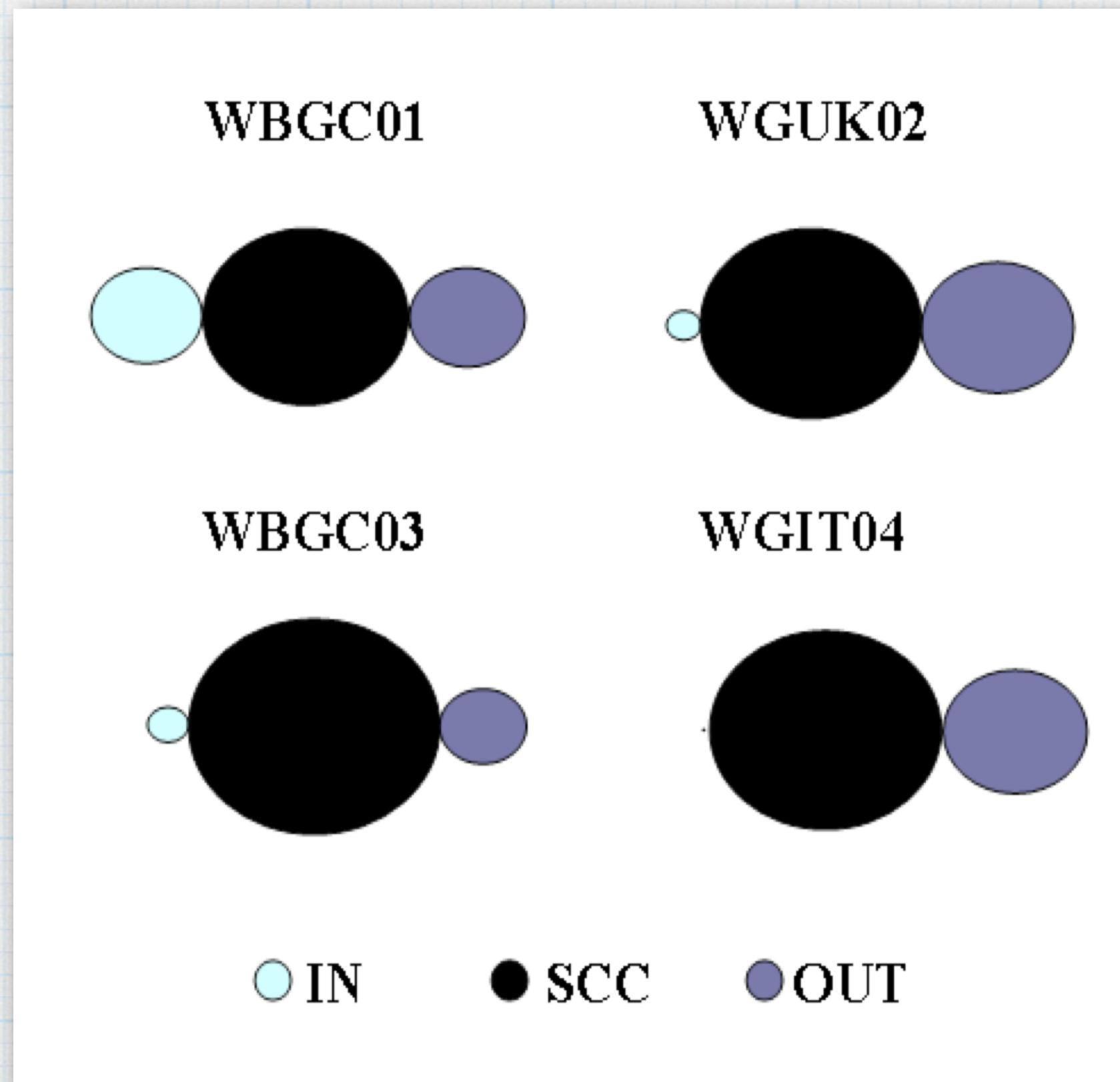
- \* Degree is broadly distributed (Broder & al. 2000):
  - \* In-degree is a power law  $P(k_{in}) \sim k_{in}^{-2.1}$
  - \* Out-degree not so scale-free: why?

# The Web looks like a "bow-tie" (why?)



The component size is also broadly (power-law) distributed

# Web structure from different crawls



$$K = \langle k^2 \rangle / \langle k \rangle$$

Data set	WBGC01	WGUK02	WBGC03	WGIT04
# nodes	80571247	18520486	49296313	41291594
# links	752527660	292243663	1185396953	1135718909

Data set	WBGC01	WGUK02	WBGC03	WGIT04
IN	17.24	1.69	2.28	0.03
SCC	56.46	65.28	85.87	72.30
OUT	17.94	31.88	11.26	27.64
MAIN	91.64	98.85	99.41	99.98

Data set	WBGC01	WGUK02	WBGC03	WGIT04
$\langle k_{in} \rangle$	9.3	15.8	24.1	27.5
$k_{in}^{max}$	788632	194942	378875	1326744
$\sigma_{in}$	200.2	143.3	421.6	881.4
$\kappa_{in}$	4298.6	1317.5	7414.9	28269.9
$\gamma_{in}$	1.9	1.7	2.2	1.6

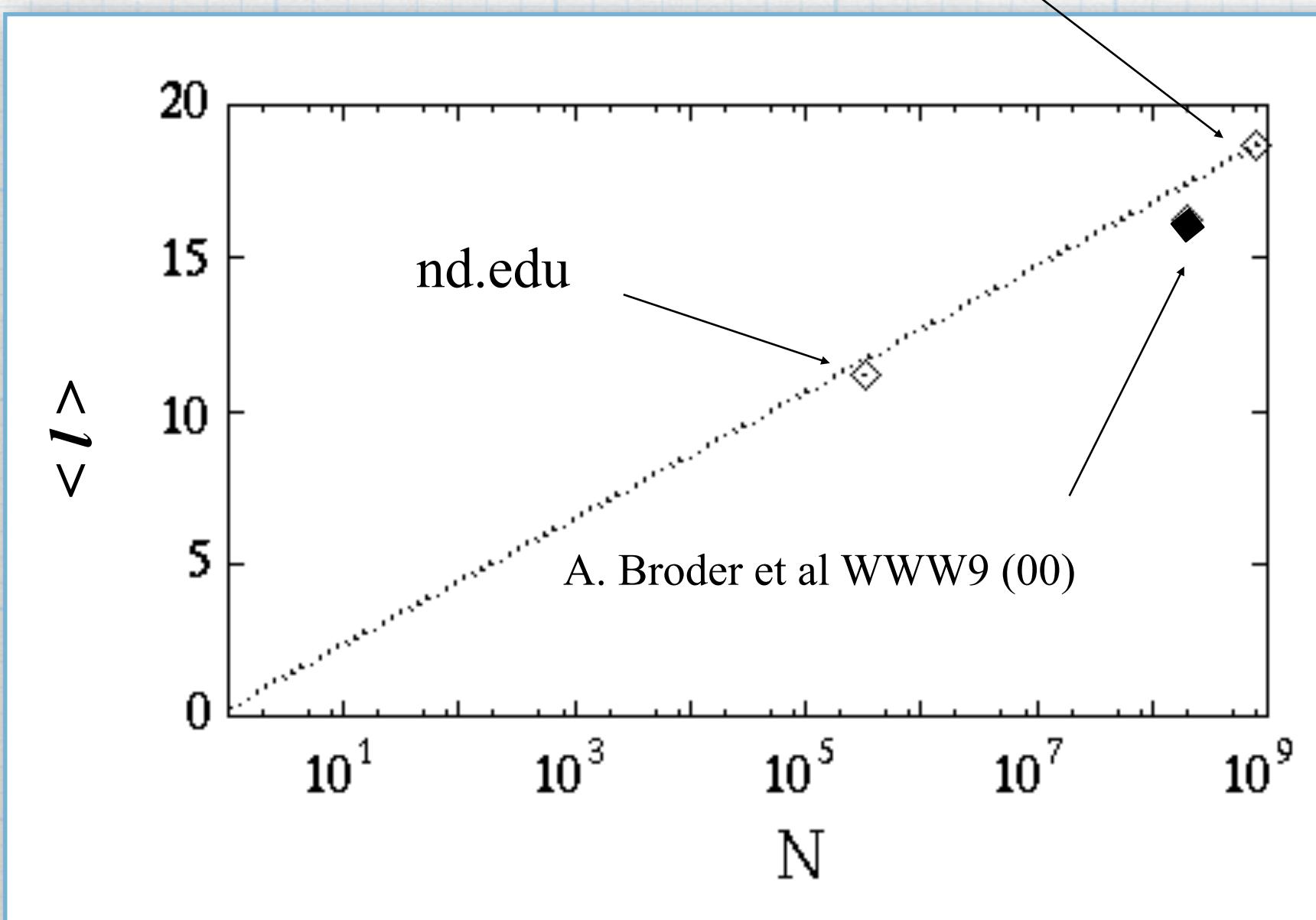
	WBGC01	WGUK02	WBGC03	WGIT04
$\langle k_{out} \rangle$	9.3	15.8	24.1	27.5
$k_{out}^{max}$	552	2449	629	9964
$\sigma_{out}$	13.1	27.4	29.5	67.1
$\kappa_{out}$	27.7	63.4	60.3	191.0
$\gamma_{out}$	$\infty$	$\infty$	$\infty$	$\infty$

Serrano & al. 2007

# Web average path length

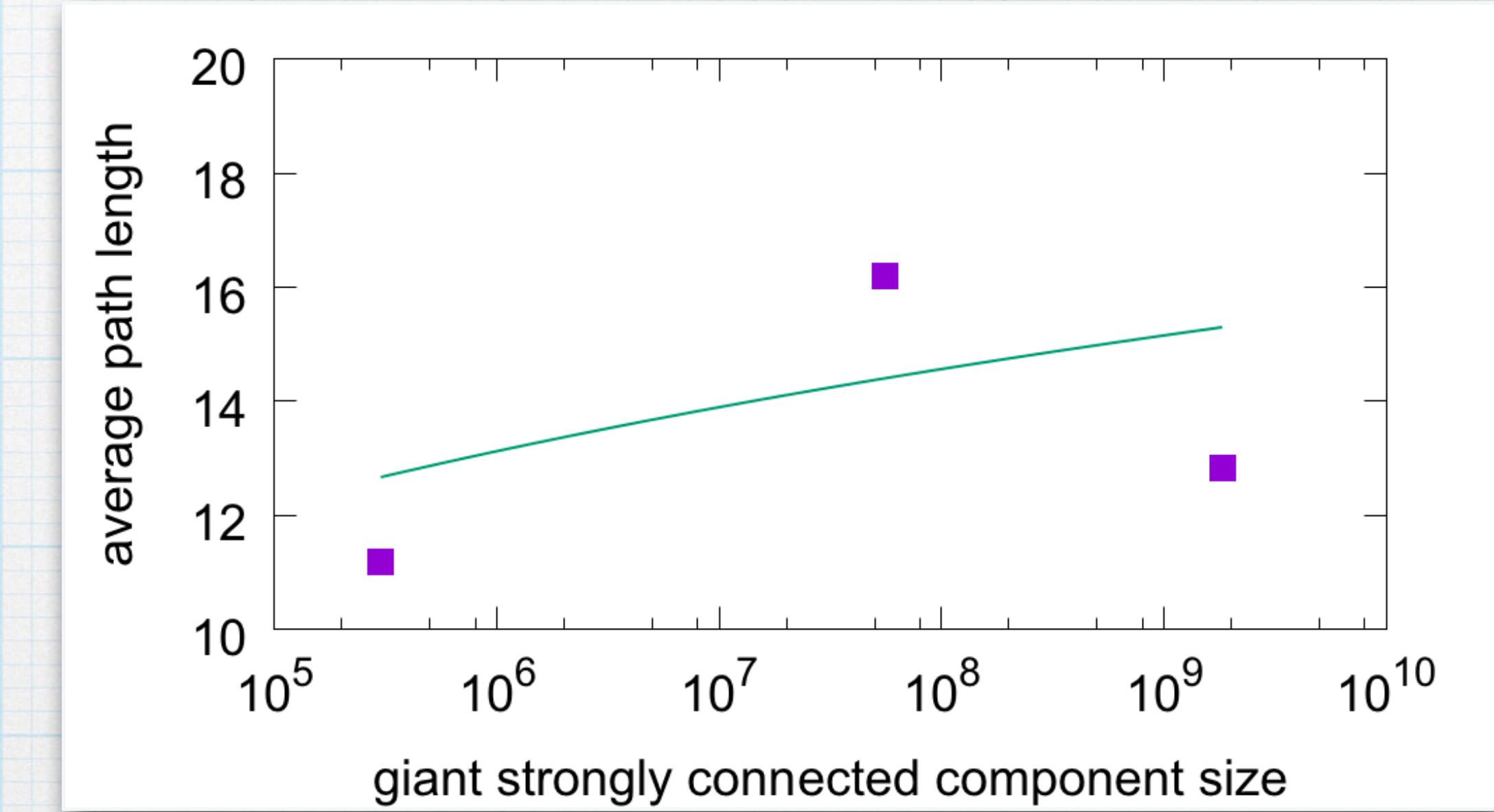
## For the core (SCC)

19 degrees of separation  
Albert et al 1999



$$\langle l \rangle \sim \log(N)$$

2012 commoncrawl.org  
Meusel et al 2015



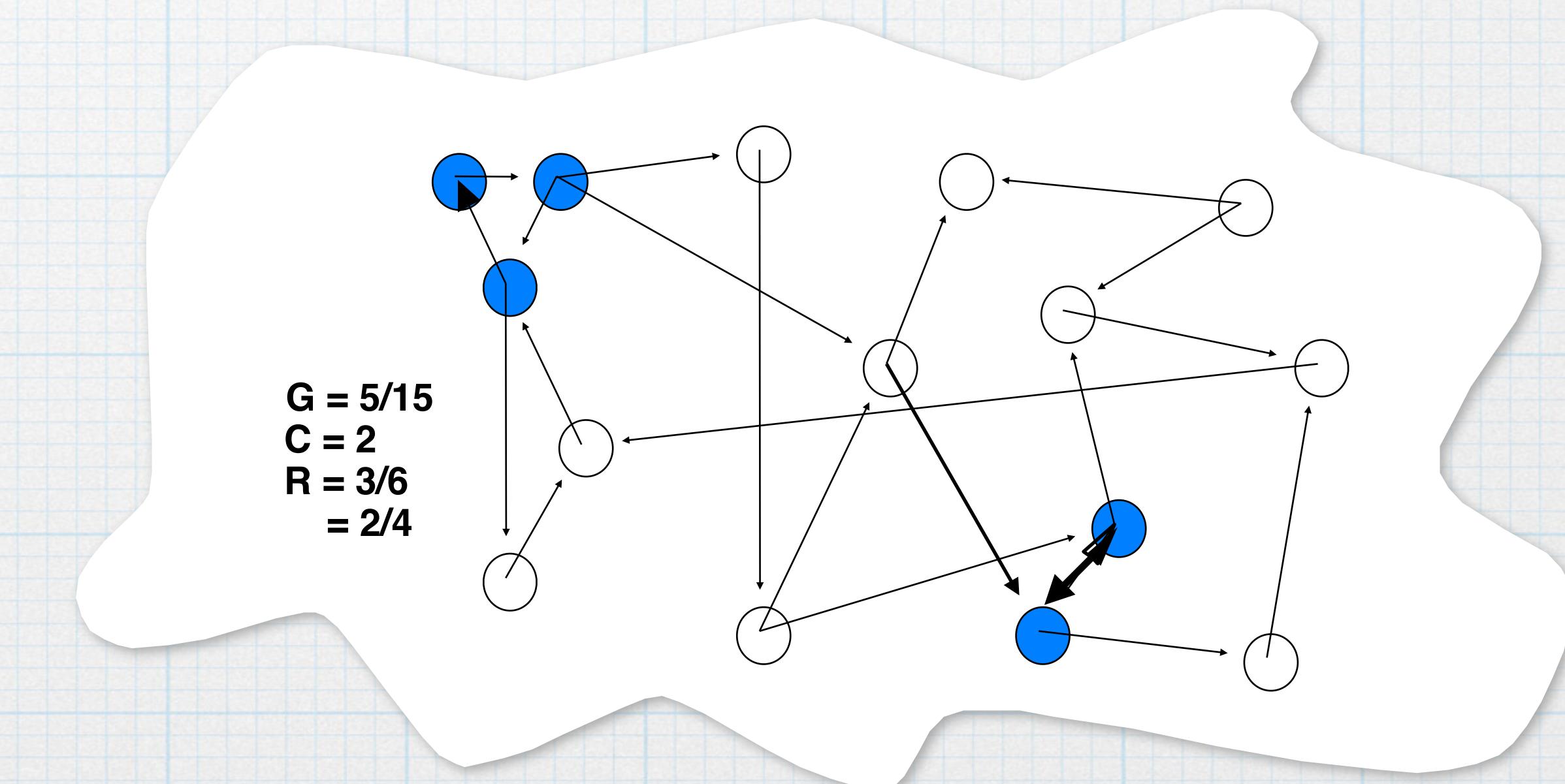
$$\langle l \rangle \sim \log(\log(N))$$

# Homophily

- \* We have seen in social networks that two nodes are more likely to be connected if they are similar, or share features. This is called "homophily" (like of same)
- \* "A man is known by the company he keeps"
- \* "Birds of a feather flock together"
- \* Social media and advertisers can determine your political leanings, sexual preference, product tastes, and much more, by looking at your friends
- \* Plays a role in "echo chambers" and "filter bubbles" in social media, as we've seen
- \* What is the equivalent of homophily on the Web?

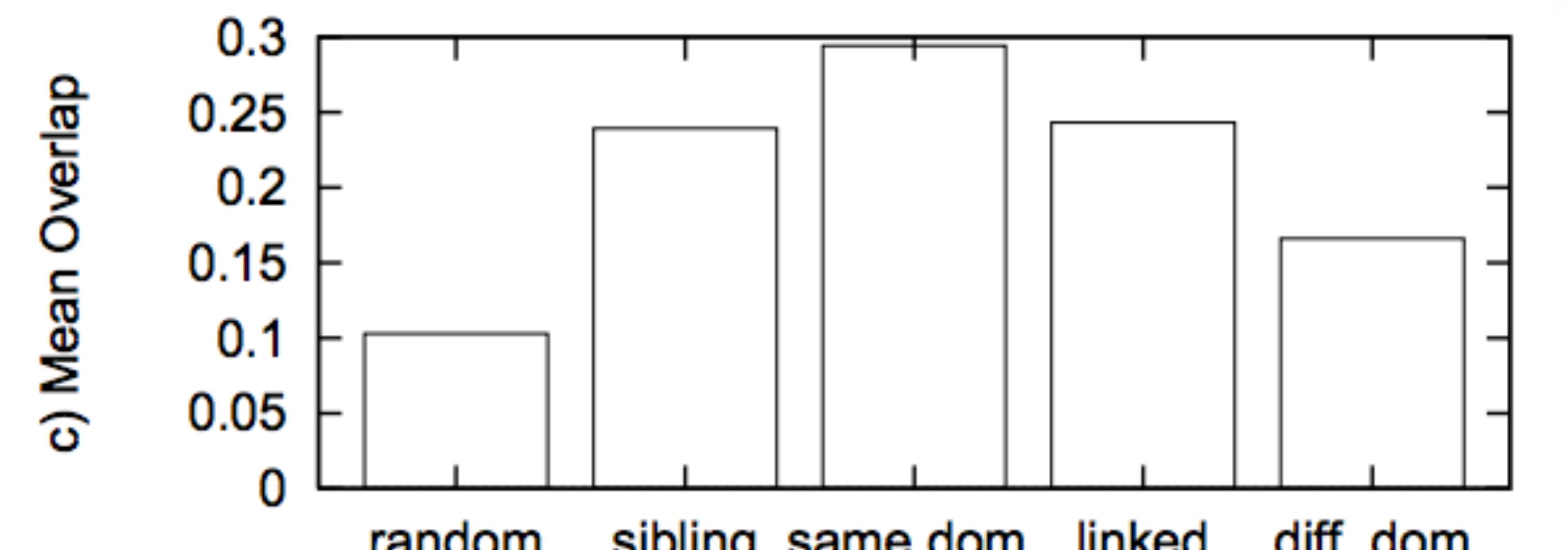
# The “link-cluster” conjecture

- Connection between **semantic** topology (relevance) and **link** topology (hypertext)
  - $G = \Pr[\text{rel}(p)]$  : fraction of relevant/topical pages (topic generality)
  - $R = \Pr[\text{rel}(p) \mid \text{rel}(q) \text{ AND } \text{link}(q,p)]$  : conditional probability given neighbor on topic
- Related nodes are **clustered** if  $R > G$ 
  - Necessary and sufficient condition for a random crawler to find pages related to start points



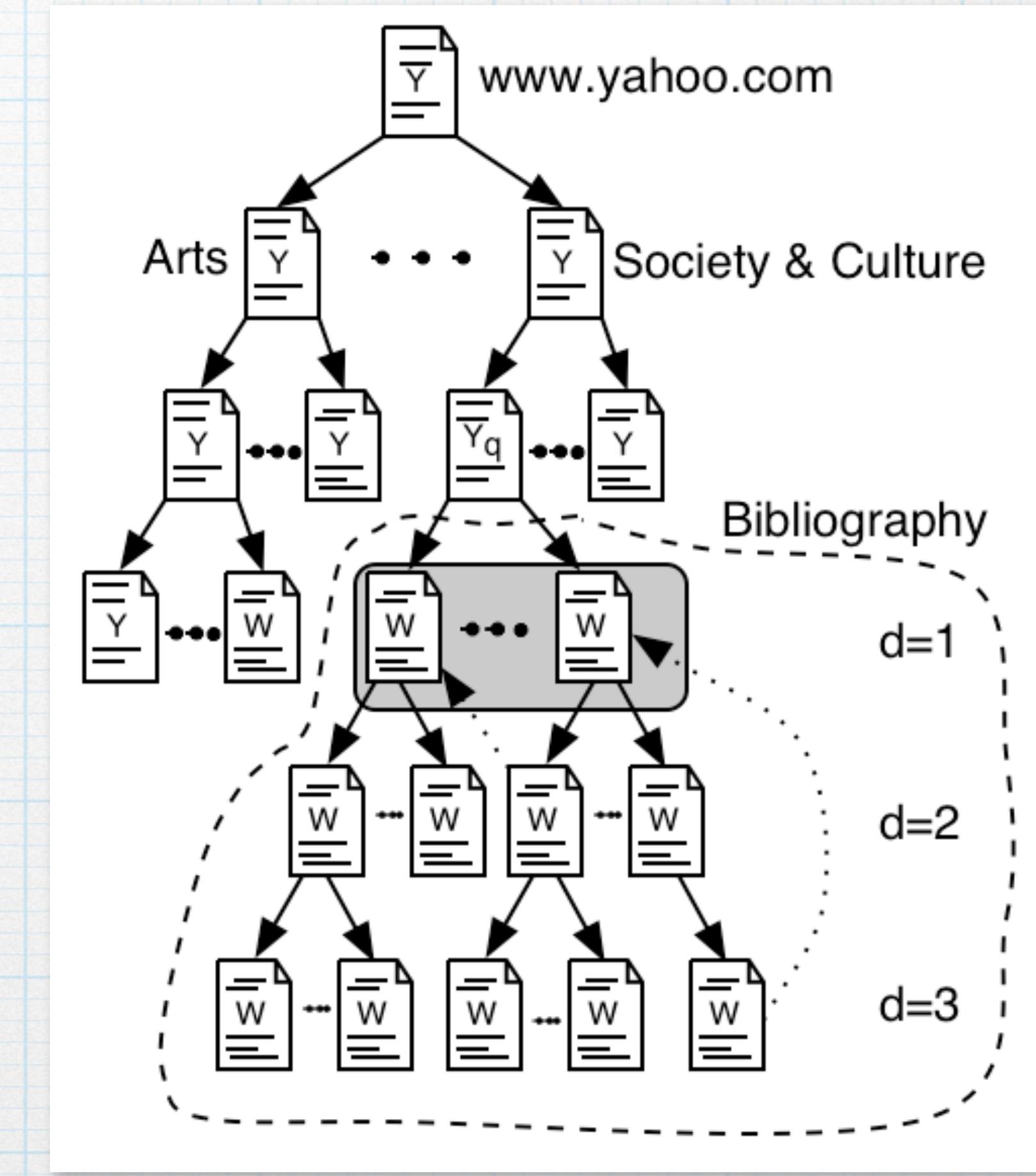
# Topical locality

- \* Topical locality means that links must encode semantic information, i.e. say something about neighbor pages; not random
- \* It is also a sufficient condition to find relevant pages if we start from “good” seed pages
- \* We know that Web topical locality is strong :
  - \* Indirectly: crawlers work and people surf the Web
  - \* Direct measurements (Davison 2000; Menczer 2004, 2005)



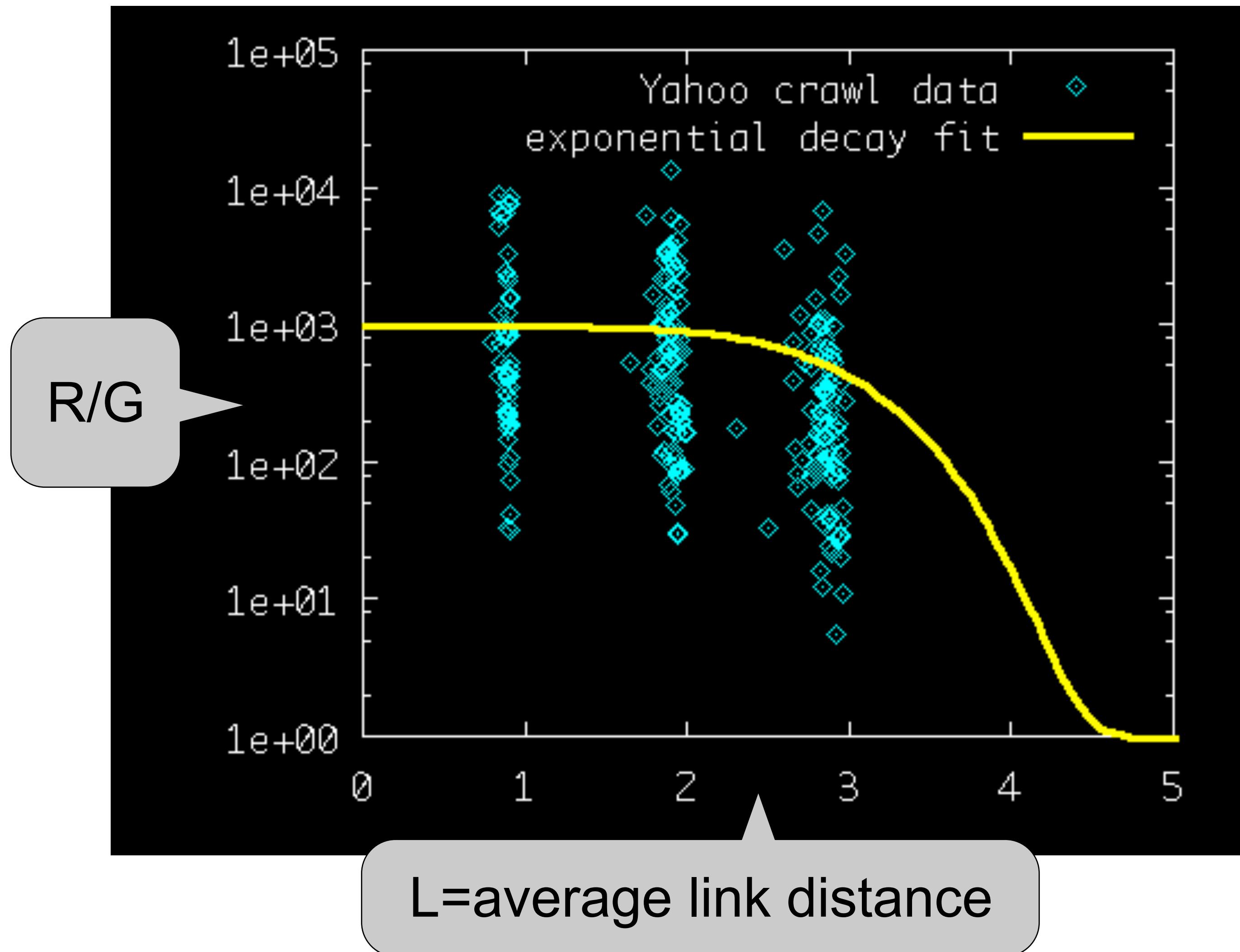
# Quantifying topical locality

- Different ways to pose the question:
  - **Semantic locality:** How quickly does the chance to visit a related page decay as we surf away from a starting page?
  - **Topic drift:** How quickly does content change as we surf away from a starting page?
- To answer these questions, let us consider **exhaustive breadth-first crawls** from 100 topic pages



# Semantic locality

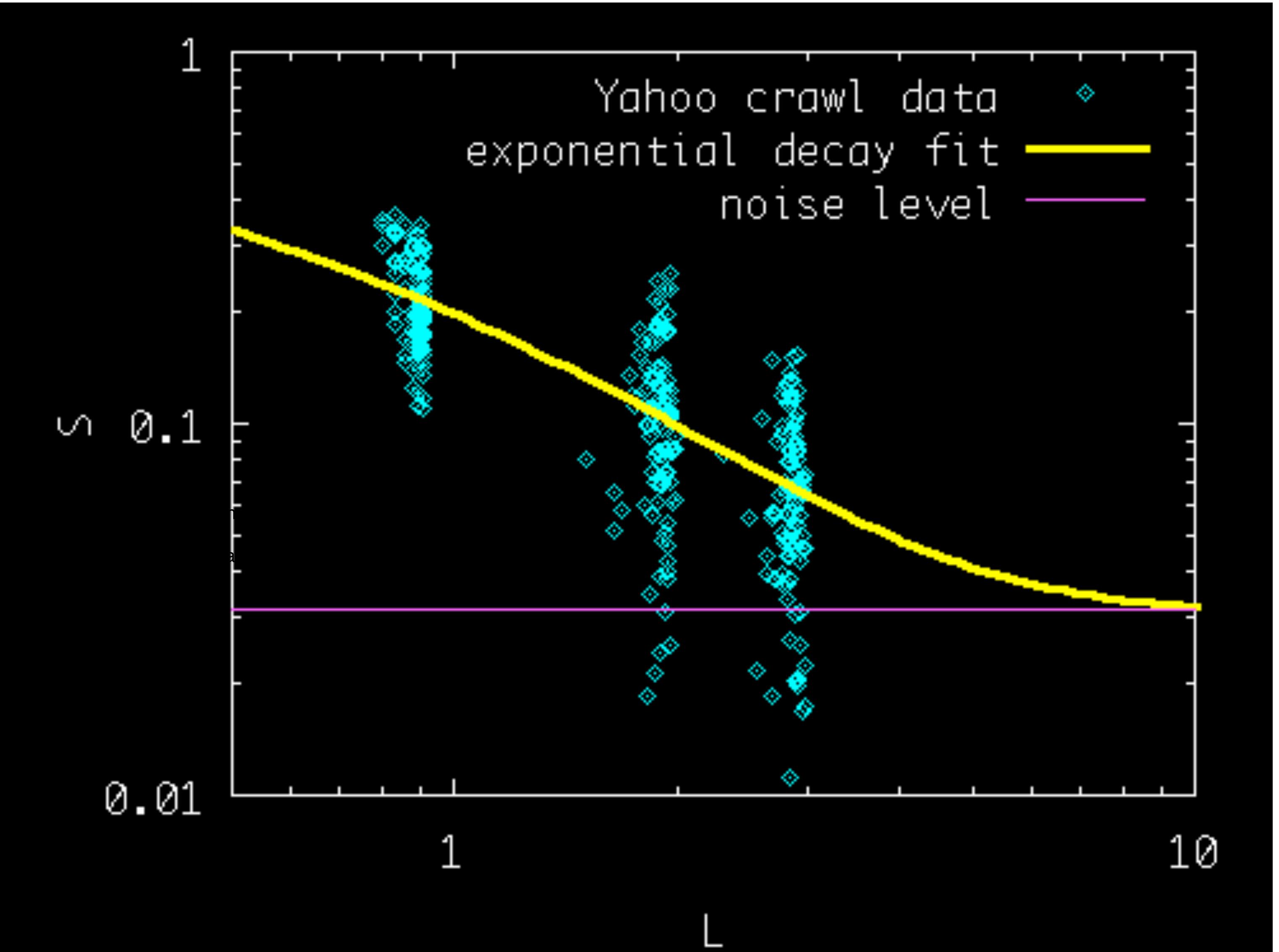
- Preservation of **semantics** (meaning) across links
- 1000 times more likely to be on topic if near an on-topic page!



Ψ

# Topic drift

- Correlation of **lexical (content)** and **link topology**
- **L**: average link distance
- **S**: average content similarity to start (topic) page

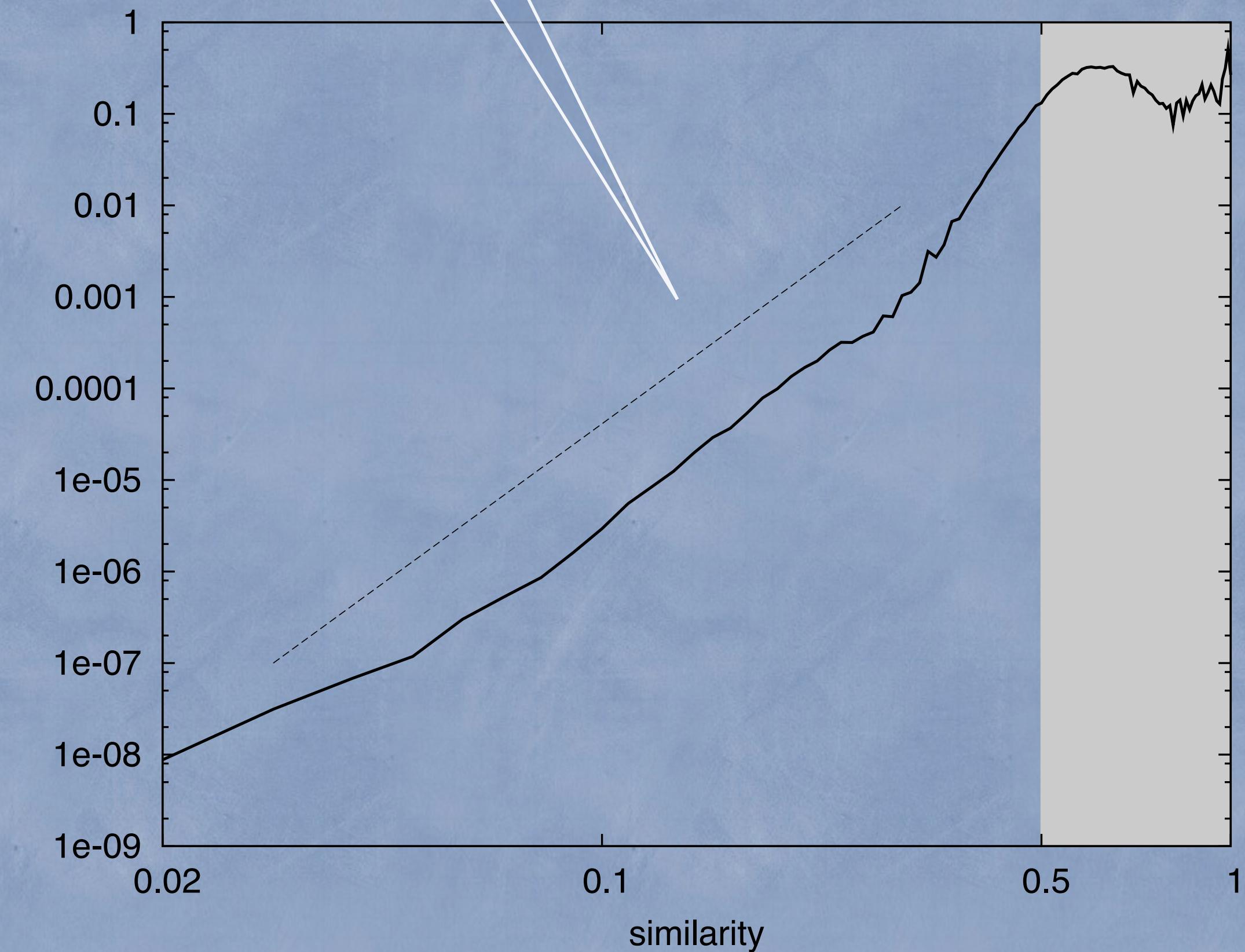


Ψ

# Topical locality

- \* What is the relationship between the similarity of two pages, and the probability they are linked?
- \* Are people more likely to create a link to a related page?

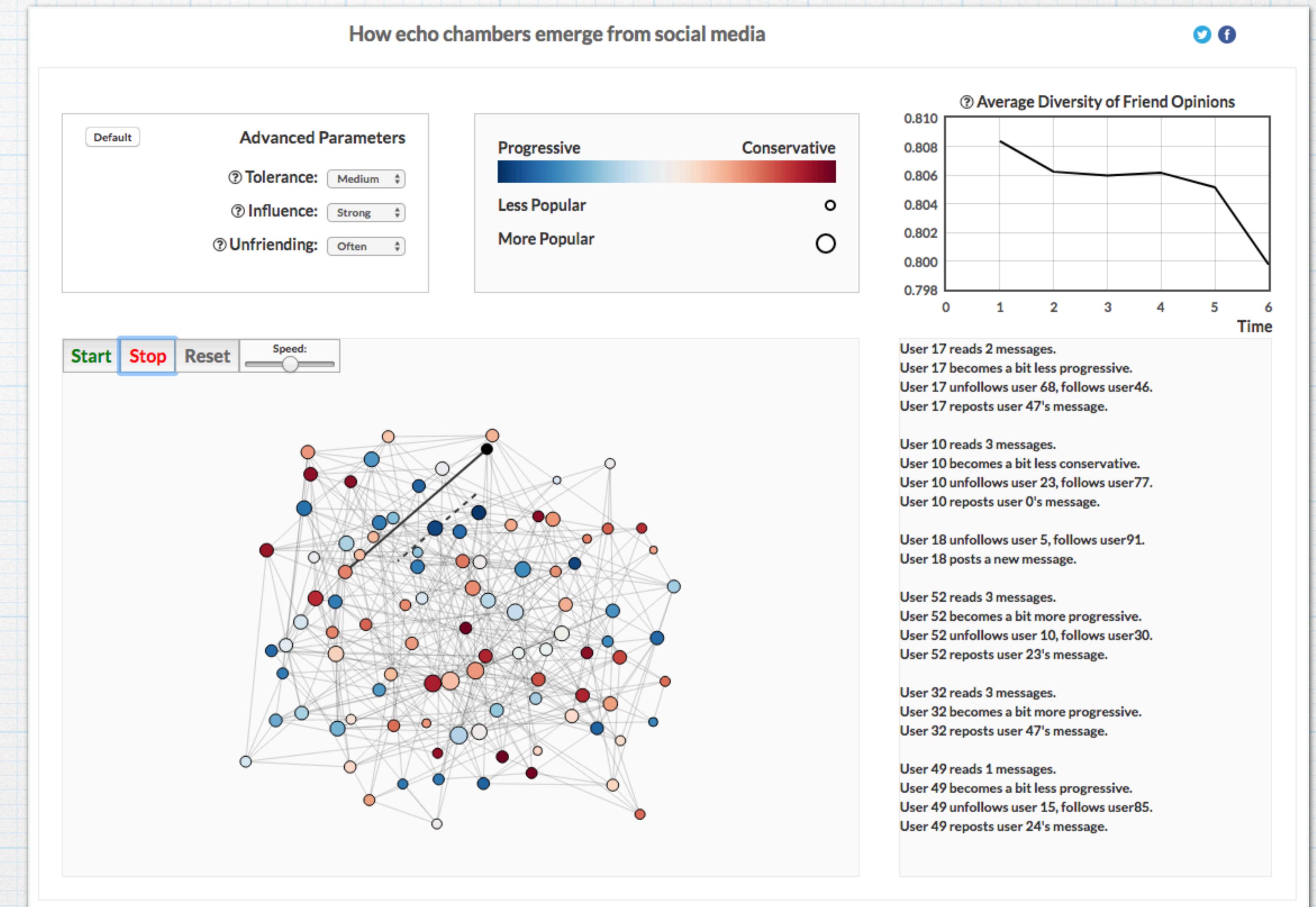
power law



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

# How might homophily arise in social media?

The extreme version: echo chambers



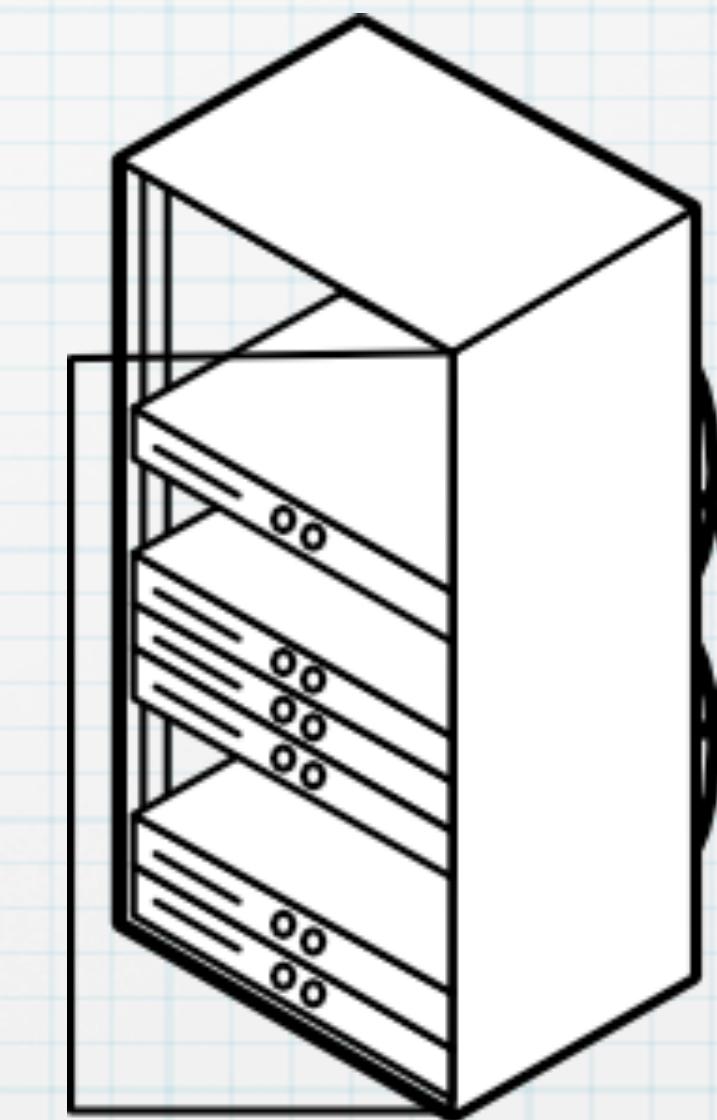
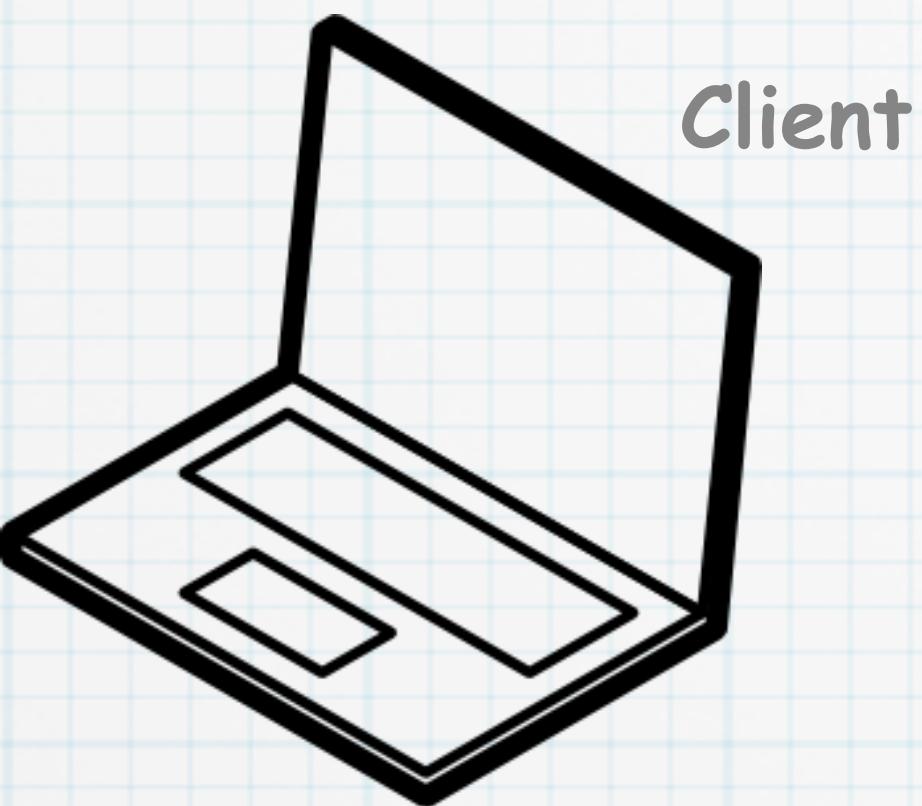
[haoopeng.github.io/echo-demo/](https://haoopeng.github.io/echo-demo/)

- Review
- Web structure
  - Degree
  - Bow-tie
  - Paths
- Homophily and topical locality

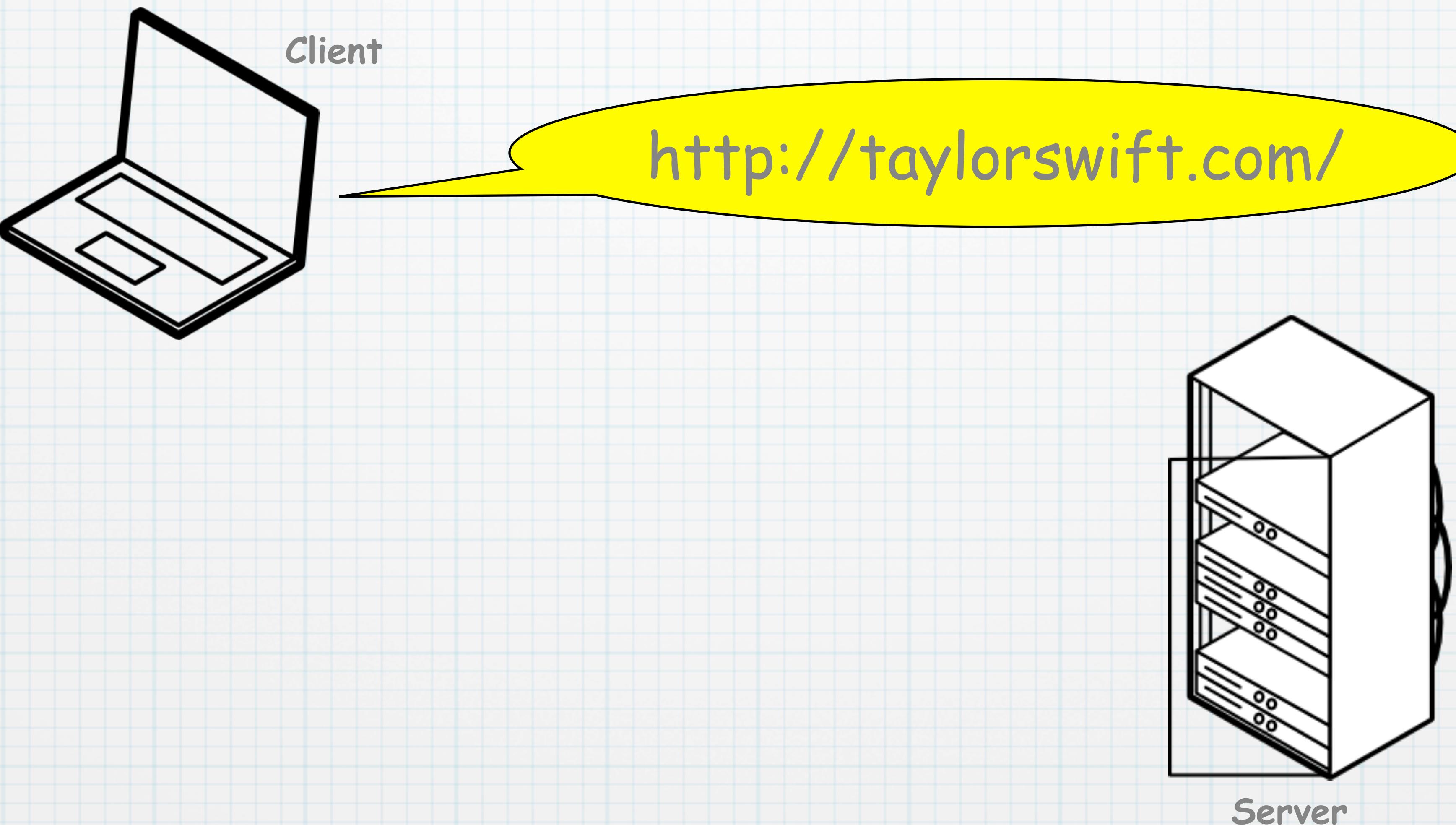
# Back to the Web

- \* How do we collect data?
- \* For the Wikipedia, it's all there for us to download and study
- \* For the Web at large, where is the data?
- \* How do we know?

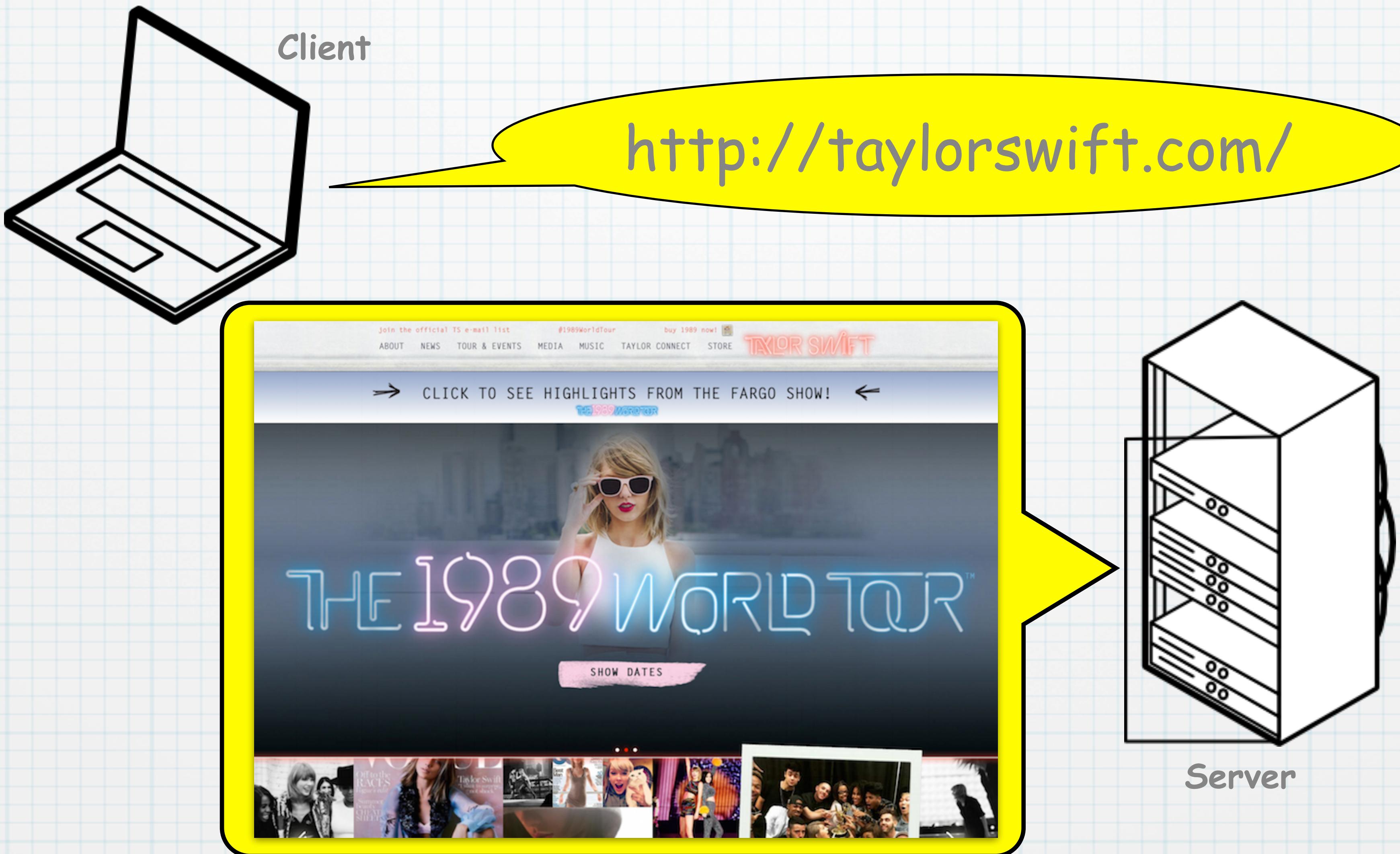
# How does the Web work?



# How does the Web work?

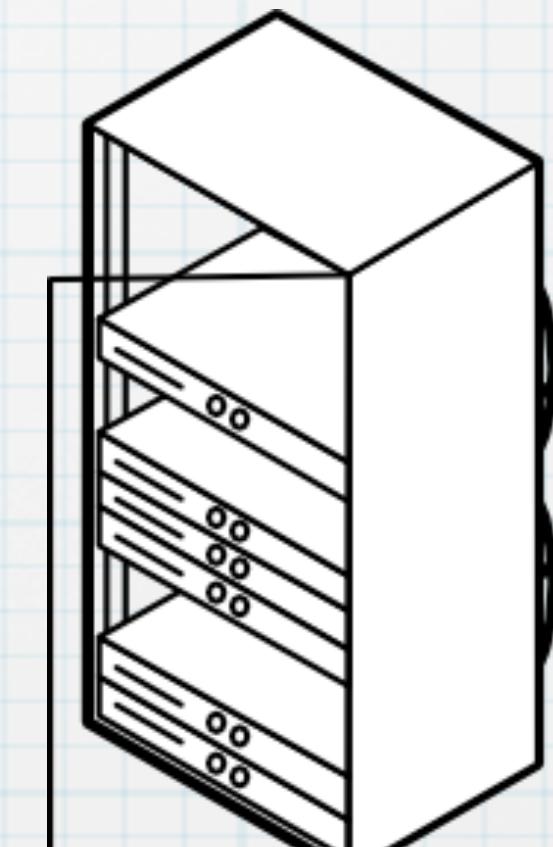
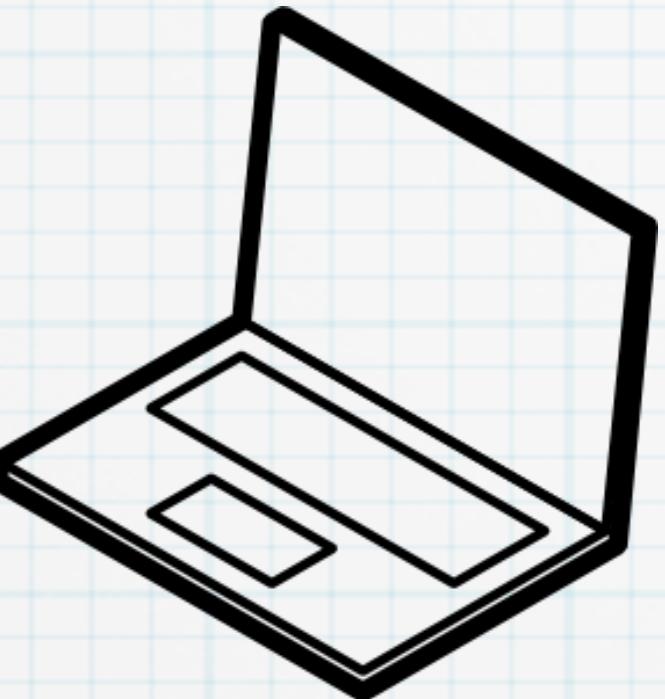


# How does the Web work?



# Let's try it the nerd way!

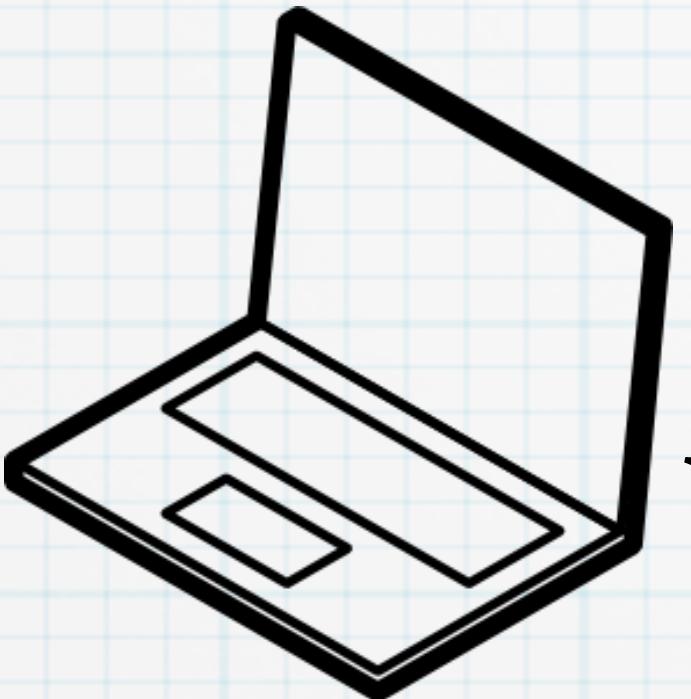
Client: "<http://taylorswift.com/>"



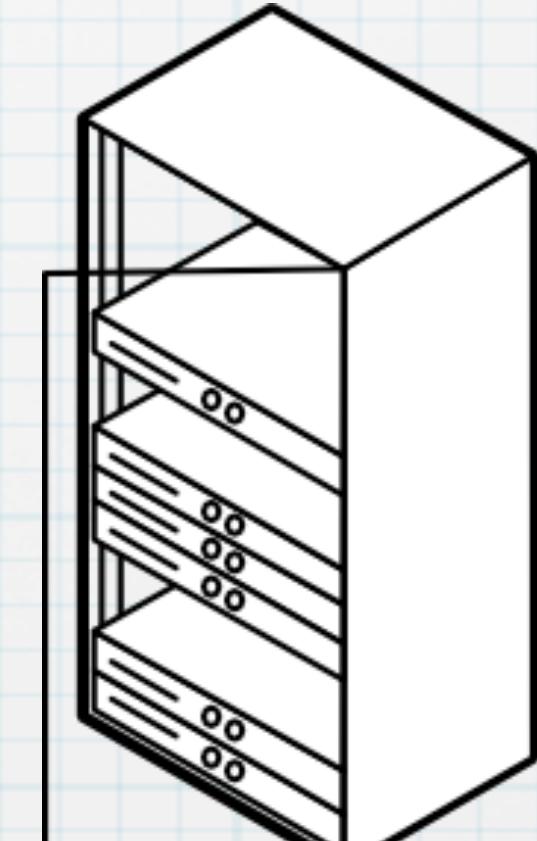
Server

# Let's try it the nerd way!

Client: "http://taylorswift.com/"



```
telnet taylorswift.com 80
Trying 2400:cb00:2048:1::6814:1a73...
Connected to taylorswift.com
Escape character is '^]'.
GET / HTTP/1.1
host: taylorswift.com
[return] (empty line means 'done')
```



Server

# Let's try it the nerd way!

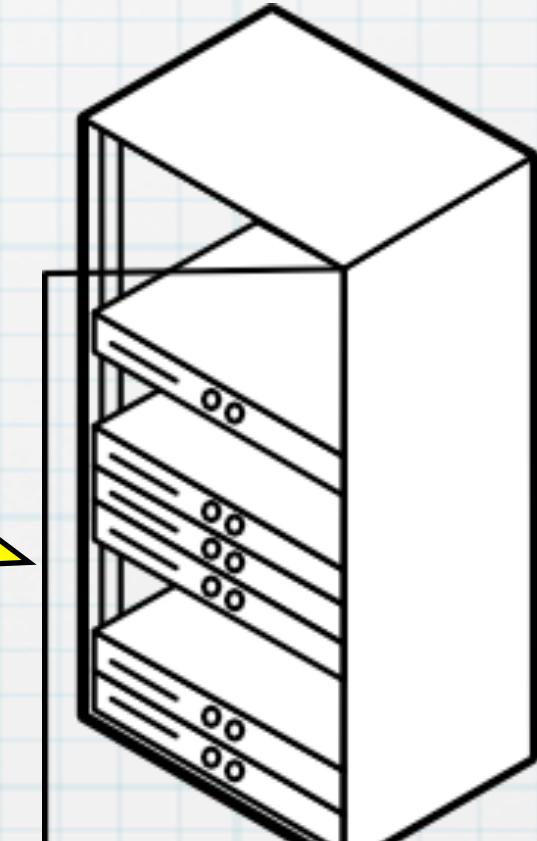
Client: "http://taylorswift.com/"



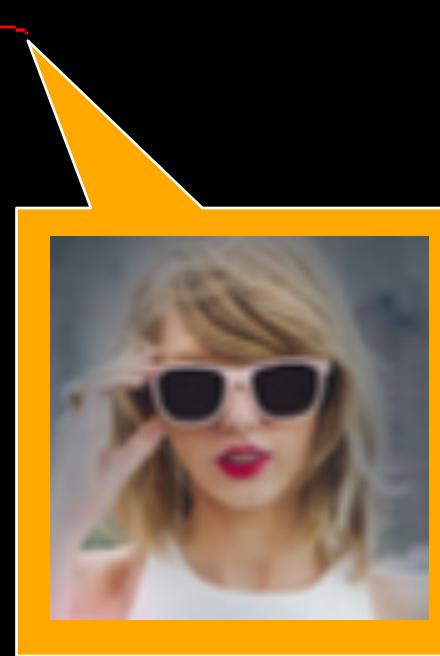
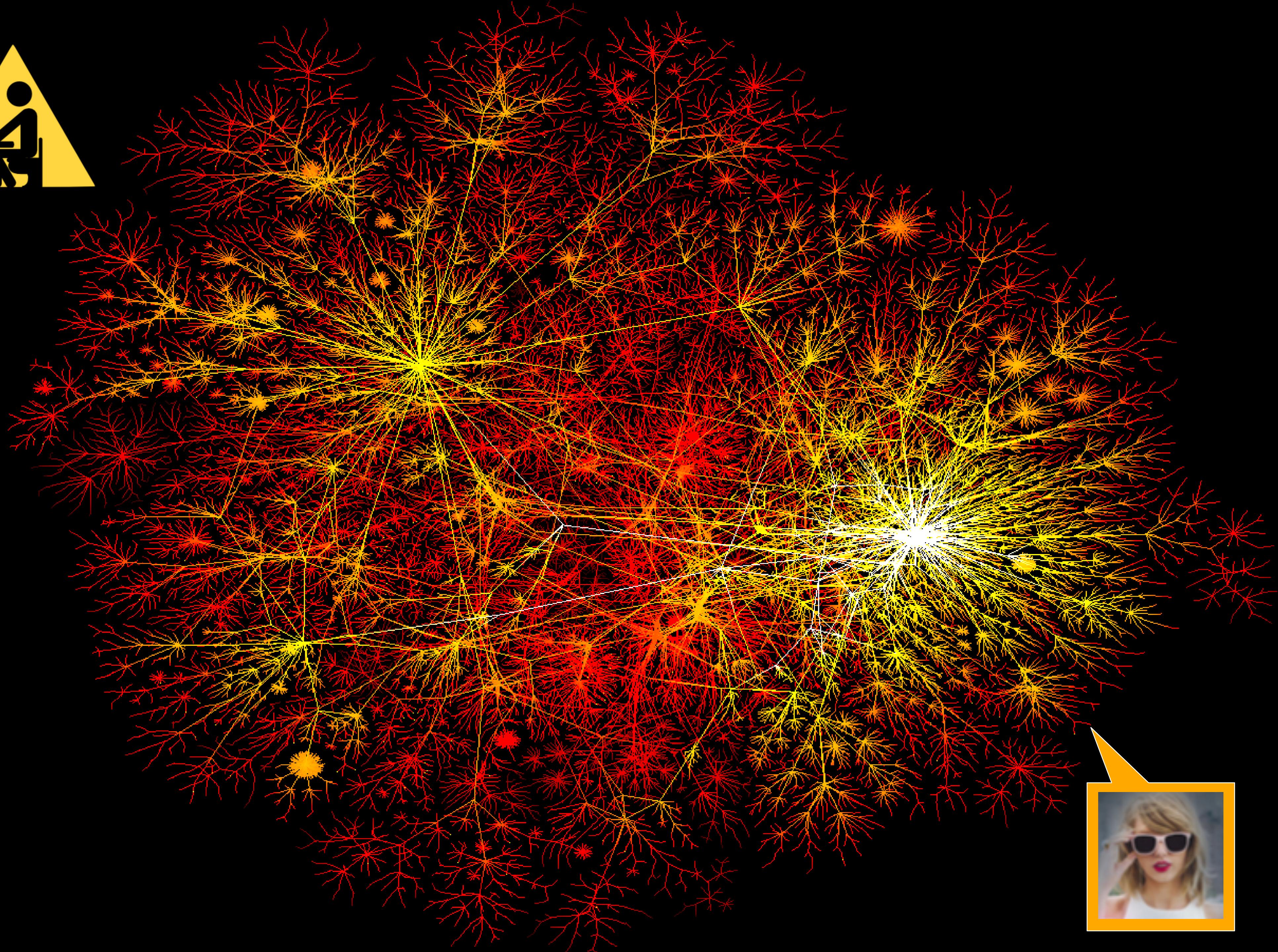
```
telnet taylorswift.com 80
Trying 2400:cb00:2048:1::6814:1a73...
Connected to taylorswift.com
Escape character is '^]'.
GET / HTTP/1.1
host: taylorswift.com
[return] (empty line means 'done')
```

```
HTTP/1.1 200 OK
Date: Thu, 15 Oct 2015 23:44:50 GMT
Content-Type: text/html; charset=UTF-8
...
Set-Cookie: __cfduid=d89379d...952690; expires=...; ... domain=.taylorswift...
...
Server: cloudflare-nginx

<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
  "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<!--[if lt IE 7]> <html class="no-js ie6 oldie ie" ... <![endif]-->
<head>
<meta content="text/html; charset=utf-8" http-equiv="Content-Type" />
<meta property="og:title" content="Taylor Swift" />
...
<title>Taylor Swift</title>
...
</body>
</html>
```



Server



Google swift

All Images News Maps Videos More Settings Tools

About 471,000,000 results (0.64 seconds)

**Swift - Apple Developer**  
<https://developer.apple.com/swift/> ▾  
 Swift is a powerful and intuitive programming language for macOS, iOS, watchOS and tvOS. Writing Swift code is interactive and fun, the syntax is concise yet ...  
[Swift Playgrounds](#) · [Swift Blog](#) · [Resources](#)

**Taylor Swift (@taylorswift13) · Twitter**  
<https://twitter.com/taylorswift13>

The #CallItWhatYouWant lyric video is out now! Watch it here: taylor.lk/CIWYWlyric  
 3 days ago · Twitter

"Call It What You Want" available now. @applemusic Pre-order #reputation: taylor.lk/reputation-iT pic.twitter.com/EWxoTpS...  
 4 days ago · Twitter

Call It What You Want. Midnight Eastern. pic.twitter.com/nTmlQUz...  
 4 days ago · Twitter

**Swift.org - Welcome to Swift.org**  
<https://swift.org/> ▾  
 Swift is now open source! We are excited by this new chapter in the story of Swift. After Apple unveiled the Swift programming language, it quickly became one of ...

**Top stories**

Diplo Goes After Taylor Swift Yet Again  
 Cosmopolitan · 19 hours ago

Taylor Swift's Boyfriend Joe Alwyn Debuts as a Prada Model  
 People · 2 days ago

ACLU Backs Blogger Facing Legal Action From Taylor Swift After Post Demanding She...  
 Stereogum · 24 mins ago

→ More for swift

**Swift**  
 Programming language

Swift is a general-purpose, multi-paradigm, compiled programming language developed by Apple Inc. for iOS, macOS, watchOS, tvOS, and Linux. [Wikipedia](#)

**Stable release:** 4.0 / September 19, 2017; 34 days ago

**Developer:** Apple Inc.

**First appeared:** June 2, 2014; 3 years ago

**OS:** Darwin, Linux, FreeBSD

**Typing discipline:** Static, strong, inferred

**License:** Apache License 2.0 (Swift 2.2 and later); Proprietary (up to Swift 2.2)

**People also search for** [View 15+ more](#)

Python Objective-C C Ruby Java

[Feedback](#)

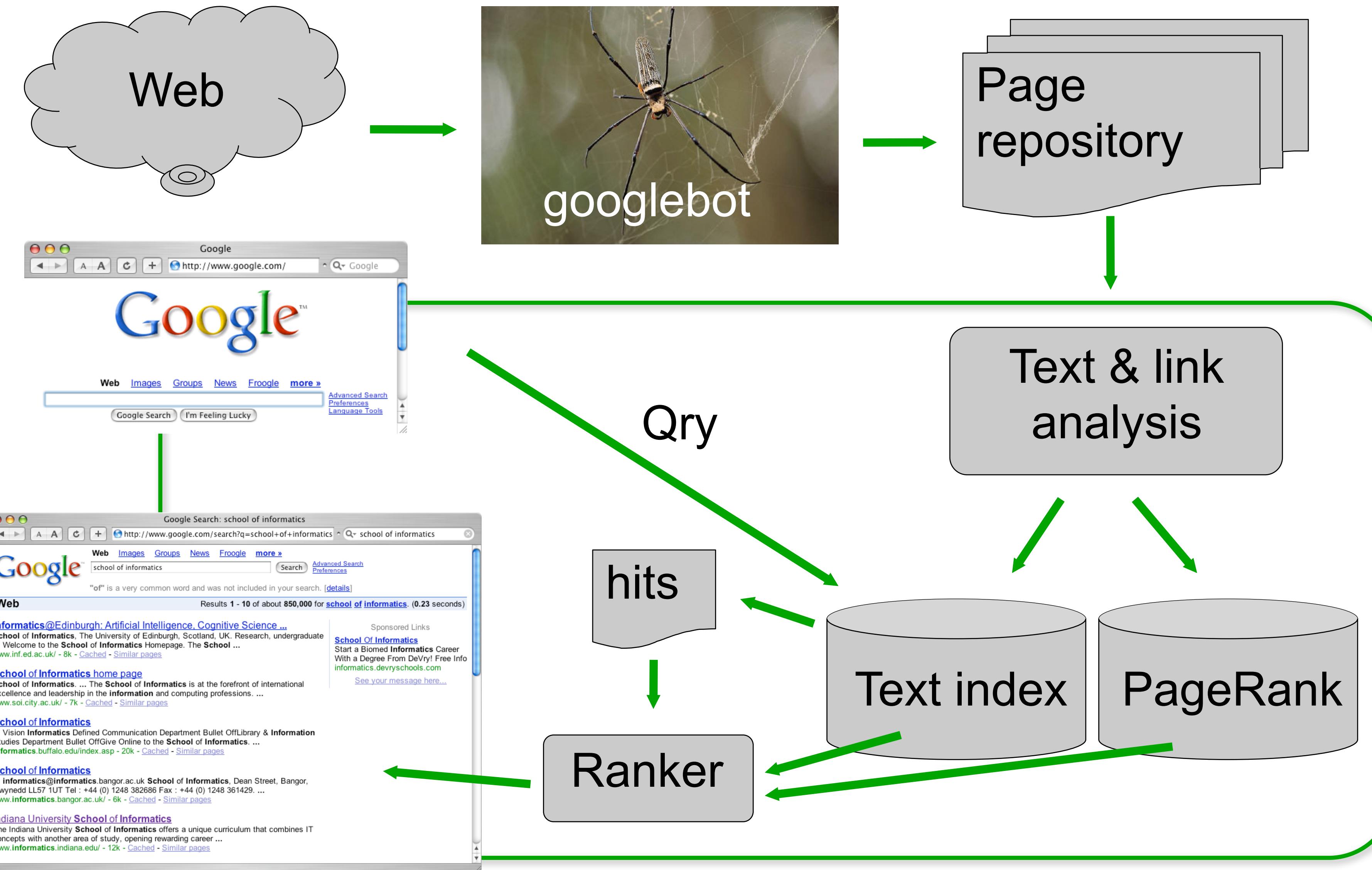
**See results about**

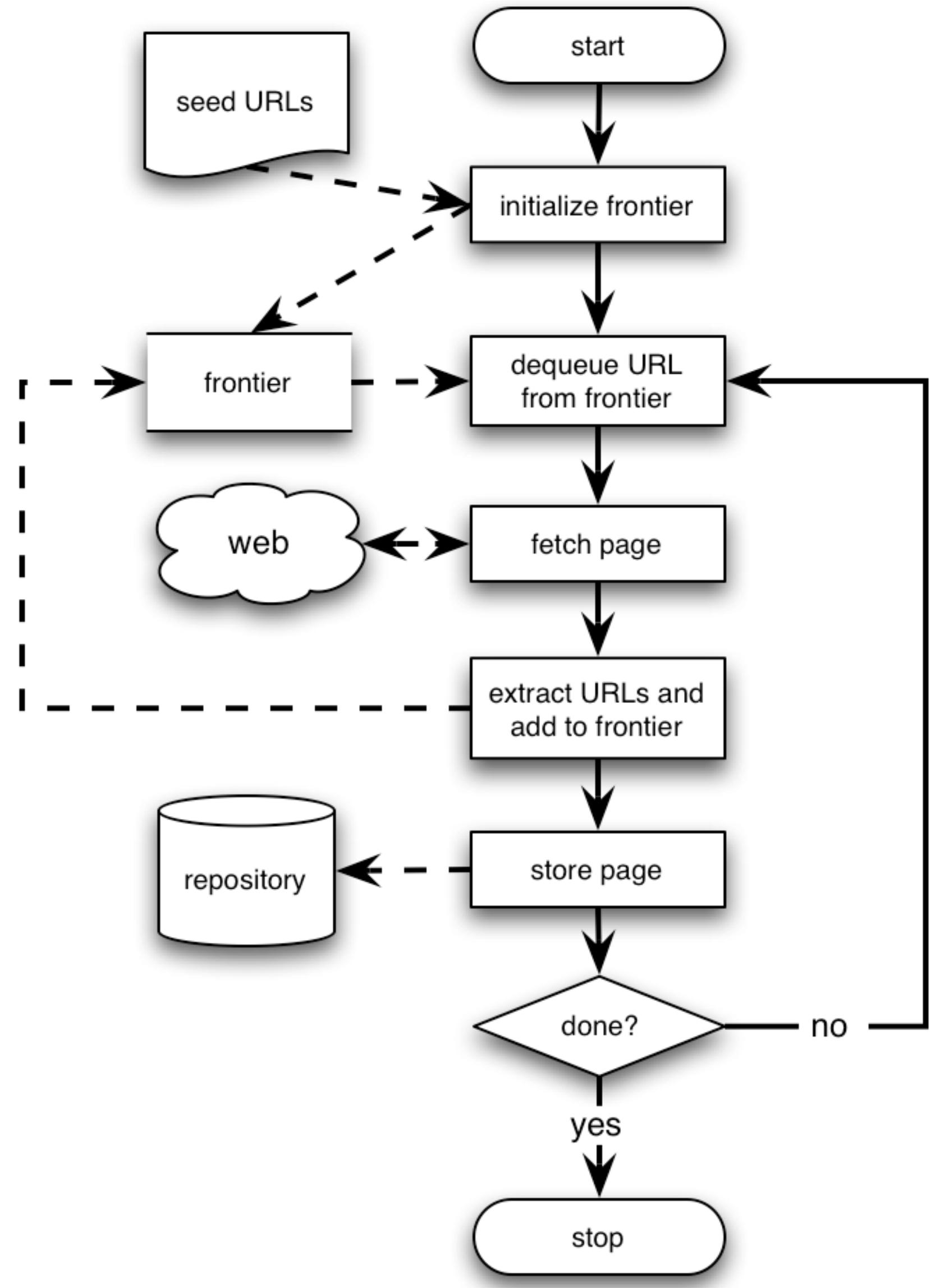
Swift Gamma-Ray Burst Mission (Program)  
 Start date: November 20, 2004  
 Rocket: Delta II

Society for Worldwide Interbank Financial Telecommunic...  
 The Society for Worldwide Interbank Financial Telecommunication provides a network that ...



# Anatomy of a search engine





# A basic crawler

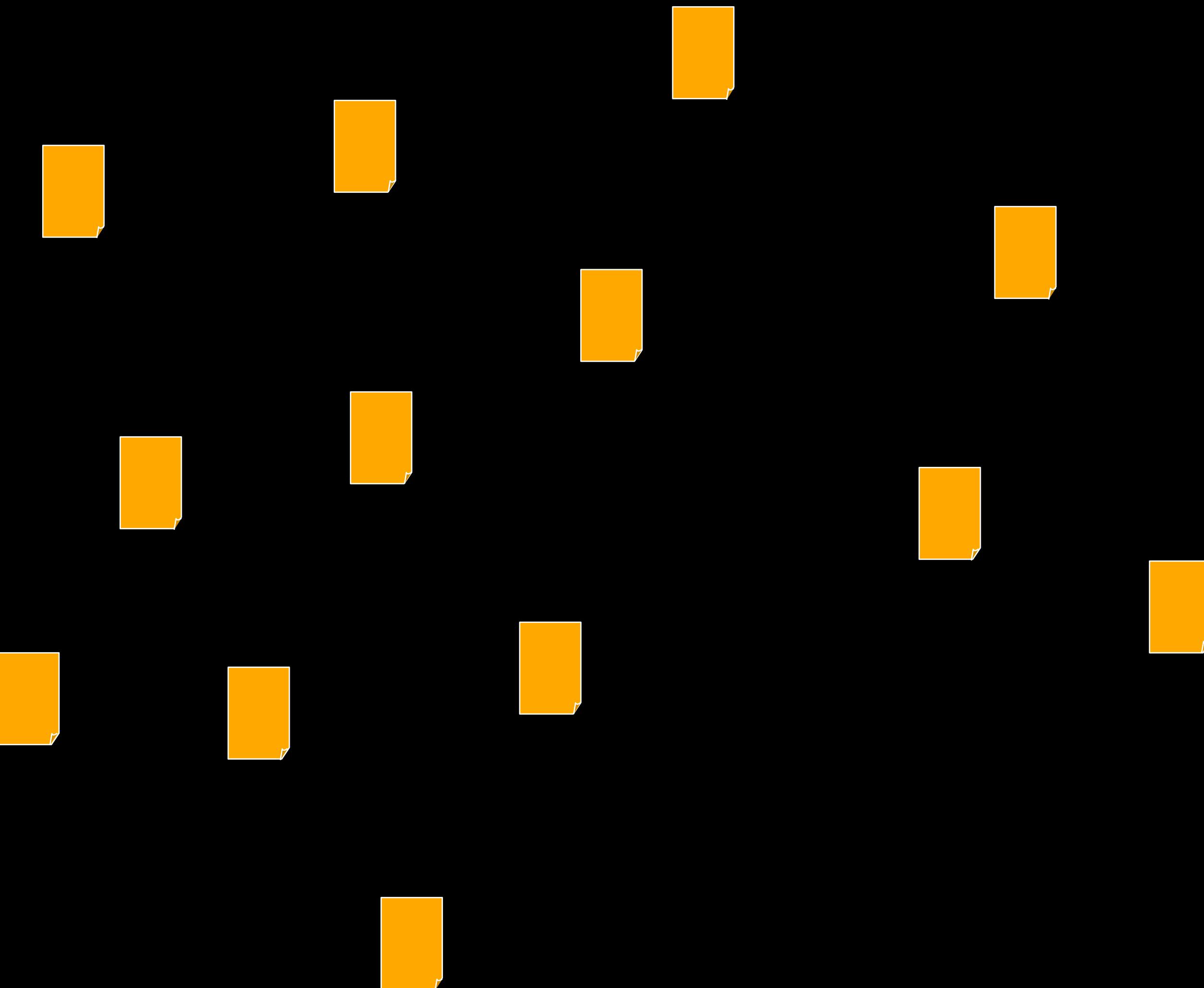
- This is a **sequential** crawler
- **Seeds** can be any list of starting URLs
- Order of page visits is determined by **frontier** data structure
- **Stop** criterion can be anything

# Googlebot & you

```
tcsh — #1  
homer:~% more /var/log/httpd/access_log  
129.217.55.111 - - [11/Sep/2004:04:36:24 -0500] "GET /~fil/Thanksgiving/1999/Pages/Image10.html HTTP/1.0" 200 302  
84.135.208.173 - - [11/Sep/2004:04:40:57 -0500] "GET /~fil/Max/2000/Fall/November/ HTTP/1.1" 404 320  
80.100.20.198 - - [11/Sep/2004:04:41:40 -0500] "GET /~fil/Max/2000/Fall/November/ HTTP/1.0" 404 308  
64.68.82.182 - - [11/Sep/2004:04:41:51 -0500] "GET /robots.txt HTTP/1.0" 404 290  
62.39.213.35 - - [11/Sep/2004:04:41:52 -0500] "GET /~fil/Max/2000/Fall/November/ HTTP/1.0" 404 308  
64.68.82.182 - - [11/Sep/2004:04:41:52 -0500] "GET /network/network.map HTTP/1.0" 200 3544  
129.217.55.111 - - [11/Sep/2004:04:41:58 -0500] "GET /~fil/Max/2003/Fall/Fall-Pages/Image3.html HTTP/1.0" 200 491  
129.217.55.111 - - [11/Sep/2004:04:42:01 -0500] "GET /~fil/Max/2002/Spring/Spring-Pages/Image6.html HTTP/1.0" 200 495  
129.217.55.111 - - [11/Sep/2004:04:42:03 -0500] "GET /~fil/Max/2002/Europe02/Crans-Montana/ HTTP/1.0" 200 6361  
129.217.55.111 - - [11/Sep/2004:04:42:36 -0500] "GET /~fil/Vacation/Europe02/Venezia/Pages/Image12.html HTTP/1.0" 200 352  
129.217.55.111 - - [11/Sep/2004:04:43:01 -0500] "GET /~fil/Thanksgiving/1999/Pages/Image9.html HTTP/1.0" 200 301  
129.217.55.111 - - [11/Sep/2004:04:43:43 -0500] "GET /~fil/Max/2003/Fall/Fall-Pages/Image2.html HTTP/1.0" 200 485  
129.217.55.111 - - [11/Sep/2004:04:43:45 -0500] "GET /~fil/Max/2002/Spring/Spring-Pages/Image5.html HTTP/1.0" 200 498  
129.217.55.111 - - [11/Sep/2004:04:43:48 -0500] "GET /~fil/Max/2002/Europe02/Bologna/ HTTP/1.0" 200 2469  
129.217.55.111 - - [11/Sep/2004:04:44:14 -0500] "GET /~fil/Vacation/Europe02/Venezia/Pages/Image11.html HTTP/1.0" 200 352  
129.217.55.111 - - [11/Sep/2004:04:44:49 -0500] "GET /~fil/Thanksgiving/1999/Pages/Image8.html HTTP/1.0" 200 301  
129.217.55.111 - - [11/Sep/2004:04:45:30 -0500] "GET /~fil/Max/2003/Fall/Fall-Pages/Image1.html HTTP/1.0" 200 485  
129.217.55.111 - - [11/Sep/2004:04:45:31 -0500] "GET /~fil/Max/2002/Spring/Spring-Pages/Image4.html HTTP/1.0" 200 501  
129.217.55.111 - - [11/Sep/2004:04:45:57 -0500] "GET /~fil/Vacation/Europe02/Venezia/Pages/Image10.html HTTP/1.0" 200 352  
129.217.55.111 - - [11/Sep/2004:04:46:25 -0500] "GET /~fil/Thanksgiving/1999/Pages/Image7.html HTTP/1.0" 200 301  
129.217.55.111 - - [11/Sep/2004:04:50:27 -0500] "GET /~fil/Max/2003/Fall/Fall-Pages/Image0.html HTTP/1.0" 200 495  
129.217.55.111 - - [11/Sep/2004:04:50:30 -0500] "GET /~fil/Max/2002/Spring/Spring-Pages/Image3.html HTTP/1.0" 200 501  
129.217.55.111 - - [11/Sep/2004:04:50:59 -0500] "GET /~fil/Vacation/Europe02/Venezia/Pages/Image9.html HTTP/1.0" 200 318  
129.217.55.111 - - [11/Sep/2004:04:51:32 -0500] "GET /~fil/Thanksgiving/1999/Pages/Image6.html HTTP/1.0" 200 301  
129.217.55.111 - - [11/Sep/2004:04:52:40 -0500] "GET /~fil/Max/2002/Spring/Spring-Pages/Image2.html HTTP/1.0" 200 522  
homer:~% host 64.68.82.182  
182.82.68.64.in-addr.arpa domain name pointer crawler14.googlebot.com.  
homer:~%
```



# Index





# Index

The screenshot shows a web browser window with the title "Official Britney Spears Biography". The main content area features a large photo of Britney Spears wearing a pink visor. To the left of the photo is a sidebar with the heading "ABOUT BRITNEY" and links for "Biography", "Awards", and "Photo Gallery". Below the sidebar is a section titled "ABOUT BRITNEY: BIOGRAPHY" with a small thumbnail image of her eye. The main text area begins with: "Britney Spears may have titled her new single "Me Against The Music," but she has rarely been more creatively in tune than she is right now. "I feel like I've hit a great new stride as an artist," she says with pride. "I've worked hard, and I feel like I've grown on so many levels." In truth, "Me Against The Music" is hardly about declaring war against grooves. "Actually, it's about the intensity that people approach music with," Britney shares. "It's about getting totally lost in the music and pushing yourself to the edge in every way you can imagine. I love thoroughly immersing myself in music, and I wanted to capture that intensity in a song." Britney's musical intensity and her evolution from a teen renegade into a provocative young woman are undeniable throughout "In The Zone," her fourth Jive Records collection. First and foremost, the project shows her flexing notably strong and mature songwriting muscles. She co-wrote 7 of the project's 12 sterling new compositions, collaborating with such heavy hitters as Red Zone ("Me Against The Music," "The Hook Up"), The Matrix ("Shadow"), Moby ("Early Mornin'), and Cathy Dennis ("Toxic," "Showdown"). Also contributing hit worthy material to the album is R. Kelly ("Outrageous"), Ying-Yang Twins on "(I Got That) Boom Boom." Perhaps most significant is the appearance of pop icon Madonna, who lends her voice to the single "Me Against The Music." Collaborating with one of her all-time greatest musical influences was a dream come true for Britney. "The experience was beyond words or description," she says. The two forged what has become a powerful bond while rehearsing for their now-notorious performance on the MTV Video Music Awards this fall. "As we were working together, there were moments when I simply could not believe that I was standing there on stage next to her. It was never even in the realm of fantasy for me." The musical union of Britney and Madonna within the taut, classic-funk groove of "Me Against The Music" is quite real, though, and it reveals each of them at their most kinetic and soulful. The song's accompanying video clip, directed by Paul Hunter, shows Madonna enticing Britney through a maze-like underground club, only to disappear into thin air when Britney gets close enough to touch her. The clip is rife with symbolic gestures of Madonna passing the baton/pop power to Britney — an image that the young artist finds exciting, humbling, and perhaps a bit premature.



# Index

Official Britney Spears Biography

BRITNEY SPEARS

HOME | NEWS & NOTES | ABOUT BRITNEY | MUSIC & VIDEO | LYNNE'S CORNER | FOUNDATIO

JOIN THE BRITNEY MAILING LIST  
TELL A FRIEND ABOUT THE SITE

ABOUT BRITNEY

Biography  
Awards  
Photo Gallery

ABOUT BRITNEY: BIOGRAPHY

**BRITNEY SPEARS IN THE ZONE**

Britney Spears may have titled her new single "Me Against The Music," but she has rarely been more creatively in "the zone" now. "I feel like I've really come into my own as an artist," says Britney, with pride. "I've worked hard, and I feel like I've grown on so many levels." In truth, "Me Against The Music" is hardly about declaring war against grooves. "Actually, it's about the intensity that people approach music with," Britney shares. "It's about getting totally lost in the music and pushing yourself to the edge in every way you can imagine. I love thoroughly immersing myself in music, and I wanted to capture that intensity in a song."

Britney's musical intensity and her evolution from a teen renegade into a provocative young woman are undeniable throughout "In The Zone," her fourth Jive Records collection. First foremost, the album shows her flexing notably strong and mature songwriting muscles. She co-wrote 7 of the project's 11 tracks, including such heavy hitters as Red Zone ("Me Against The Music," "The Hook Up"), The Matrix ("Shadow"), Moby ("Early Morning"), and Cathy Dennis ("Toxic," "Showdown"). Also contributing hit worthy material to the album is R. Kelly ("Outrageous"), Ying-Yang Twins on ("I Got That) Boom Boom."

Perhaps most significant is the appearance of pop icon Madonna, who lends her voice to the single "Me Against The Music." Collaborating with one of her all-time greatest musical influences was a dream come true for Britney. "The experience was beyond words or description," she says. The two forged what has become a powerful bond while rehearsing for their now-notorious performance on the MTV Video Music Awards this fall. "As we were working together, there were moments when I simply could not believe that I was standing there on stage next to her. It was never even in the realm of fantasy for me."

The musical union of Britney and Madonna within the taut, classic-funk groove of "Me Against The Music" is quite real, though, and it reveals each of them at their most kinetic and soulful.

Britney's accompanying video clip, directed by Paul Hunter, shows Madonna enticing Britney through a maze-like underground club, only to disappear into thin air when Britney gets close enough to touch her. The clip is rife with symbolic gestures of Madonna passing the baton of power to Britney — an image that the young artist finds exciting, humbling, and perhaps a bit premature.



# Index

Official Britney Spears Biography

BRITNEY SPEARS

HOME | NEWS & NOTES | ABOUT BRITNEY | MUSIC & VIDEO | LYNNIE'S CORNER | FOUNDATIO

ABOUT BRITNEY

Biography

Awards

Photo Gallery

ABOUT BRITNEY: BIOGRAPHY

**BRITNEY SPEARS IN THE ZONE**

Britney Spears may have titled her new single "Me Against The Music," but she has rarely been more creatively in sync with the music industry now. "I feel like I've come into my stride as an artist," she says with pride. "I've worked hard, and I feel like I've grown on so many levels." In truth, "Me Against The Music" is hardly about declaring war against grooves. "Actually, it's about the intensity that people approach music with," Britney shares. "It's about getting totally lost in the music and pushing yourself to the edge in every way you can imagine. I love thoroughly immersing myself in music, and I wanted to capture that intensity in a song."

Britney's musical intensity and her evolution from a teen renegade into a provocative young woman are undeniable throughout "In The Zone," her fourth Jive Records collection. First foremost, the album shows her flexing notably strong and mature songwriting skills. She co-wrote 7 of the project's 11 tracks, including such heavy hitters as Red Zone ("Me Against The Music," "The Hook Up"), The Matrix ("Shadow"), Moby ("Early Morning"), and Cathy Dennis ("Toxic"). The album also features some hit worthy material to come, including a duet with R. Kelly ("Outrageous") and a cover of Madonna's "Material Girl." Perhaps most significant is the appearance of Madonna on the album. "Working with Madonna, who lends her voice to the single "Me Against The Music," was one of the greatest musical experiences of my life," Britney says. "The experience was beyond words or

truth music hardly declaring war grooves intensity  
approach music britney shares totally lost music  
pushing edge imagine love thoroughly immersing music  
wanted capture intensity song



# Index



truth music hardly declaring war grooves intensity  
approach music britney shares totally lost music  
pushing edge imagine love thoroughly immersing music  
wanted capture intensity song

britney	1	<a href="http://www.britneyspears.com/about-biography.php">http://www.britneyspears.com/about-biography.php</a>
edge	1	<a href="http://www.britneyspears.com/about-biography.php">http://www.britneyspears.com/about-biography.php</a>
grooves	1	<a href="http://www.britneyspears.com/about-biography.php">http://www.britneyspears.com/about-biography.php</a>
intensity	2	<a href="http://www.britneyspears.com/about-biography.php">http://www.britneyspears.com/about-biography.php</a>
love	1	<a href="http://www.britneyspears.com/about-biography.php">http://www.britneyspears.com/about-biography.php</a>
music	4	<a href="http://www.britneyspears.com/about-biography.php">http://www.britneyspears.com/about-biography.php</a>
song	1	<a href="http://www.britneyspears.com/about-biography.php">http://www.britneyspears.com/about-biography.php</a>
truth	1	<a href="http://www.britneyspears.com/about-biography.php">http://www.britneyspears.com/about-biography.php</a>
war	1	<a href="http://www.britneyspears.com/about-biography.php">http://www.britneyspears.com/about-biography.php</a>

...

# Query & Retrieval

Google Search: music intensity truth britney war love song

Web Images Groups News Froogle more »

music intensity truth britney war love song  Advanced Search Preferences

**Web** Results 1 - 10 of about 1,590 for [music intensity truth britney war love song](#). (0.55 seconds)

[Official Britney Spears Biography](#)  
... In truth, "Me Against The Music" is hardly about declaring war ... I love thoroughly immersing myself in music, and I wanted to capture that Intensity in a song ...  
[www.britneyspears.com/about.php](http://www.britneyspears.com/about.php) - 41k - Jun 14, 2004 - [Cached](#) - [Similar pages](#)

[MuchMusic.com | Artists | Britney Spears](#)  
... and I feel like I've grown on so many levels." In truth, "Me Against ... I love thoroughly immersing myself in music, and I wanted to capture that Intensity in a ...  
[www.muchmusic.com/music/artists/bio.asp?artist=80](http://www.muchmusic.com/music/artists/bio.asp?artist=80) - 82k - [Cached](#) - [Similar pages](#)

[Biograpgie - Britney Spears](#)  
In truth, "Me Against The Music" is hardly about declaring war ... I love thoroughly immersing myself in music, and I wanted to capture that Intensity in a song ...  
[berg.heim.at/anden/422268/biography.html](http://berg.heim.at/anden/422268/biography.html) - 11k - [Cached](#) - [Similar pages](#)

[Ecity Entertainment: Music - Ecify Article - About Britney](#)  
... and I feel like I've grown on so many levels.". In truth, "Me Against ... I love thoroughly immersing myself in music, and I wanted to capture that Intensity in a ...  
[www.ecityentertainment.com/music\\_ecityarticle\\_aboutbritney.html](http://www.ecityentertainment.com/music_ecityarticle_aboutbritney.html) - 25k - [Cached](#) - [Similar pages](#)

[Results of searching for you chosen keyword - ©Mobile Resurrection](#)  
... and I feel like I've grown on so many levels." In truth, "Me Against ... I love thoroughly immersing myself in music, and I wanted to capture that Intensity in a ...  
[www.mobileresurrection.co.uk/find\\_results.asp?find=Britney%20Spears&WK=Britney](http://www.mobileresurrection.co.uk/find_results.asp?find=Britney%20Spears&WK=Britney) - 52k - [Cached](#) - [Similar pages](#)

[Shakira Quotation](#)  
... If someone is telling me the truth that is when I will give ... I want my music to transcend all the barriers ... day to be able to love with the same Intensity the way ...  
[www.absolutely.net/shakira/quote.htm](http://www.absolutely.net/shakira/quote.htm) - 34k - [Cached](#) - [Similar pages](#)

# Is text enough to rank hits?

A screenshot of a Google search results page for the query "spears". The results are displayed under the "Web" category, showing approximately 9,440,000 results. A red circle highlights the search volume statistic "Results 1 - 10 of about 9,440,000 for spears [definition]. (0.08 seconds)".

**Web** Results 1 - 10 of about 9,440,000 for spears [definition]. (0.08 seconds)

[News results for spears](#) - View today's top stories

[Knee Injury Closes Spears' Onyx Hotel](#) - Billboard - 3 hours ago  
[Britney Spears' tour is canceled](#) - San Diego Union Tribune - 8 hours ago  
[Spears launches new fragrance](#) - The Star - Jun 14, 2004

[Britney Spears :: The Official Web Site](#)  
The Official Web Site of Britney Spears. Your official source for all things Britney. ...  
Remember, proceeds benefit the Britney Spears Foundation. ...  
[www.britneyspears.com/](#) - 41k - Jun 14, 2004 - [Cached](#) - [Similar pages](#)

[Britney Spears - britney.com - Jive Records](#)  
iTunes. Real/Rhapsody. Napster. Under 11.  
[www.britney.com/](#) - 10k - [Cached](#) - [Similar pages](#)

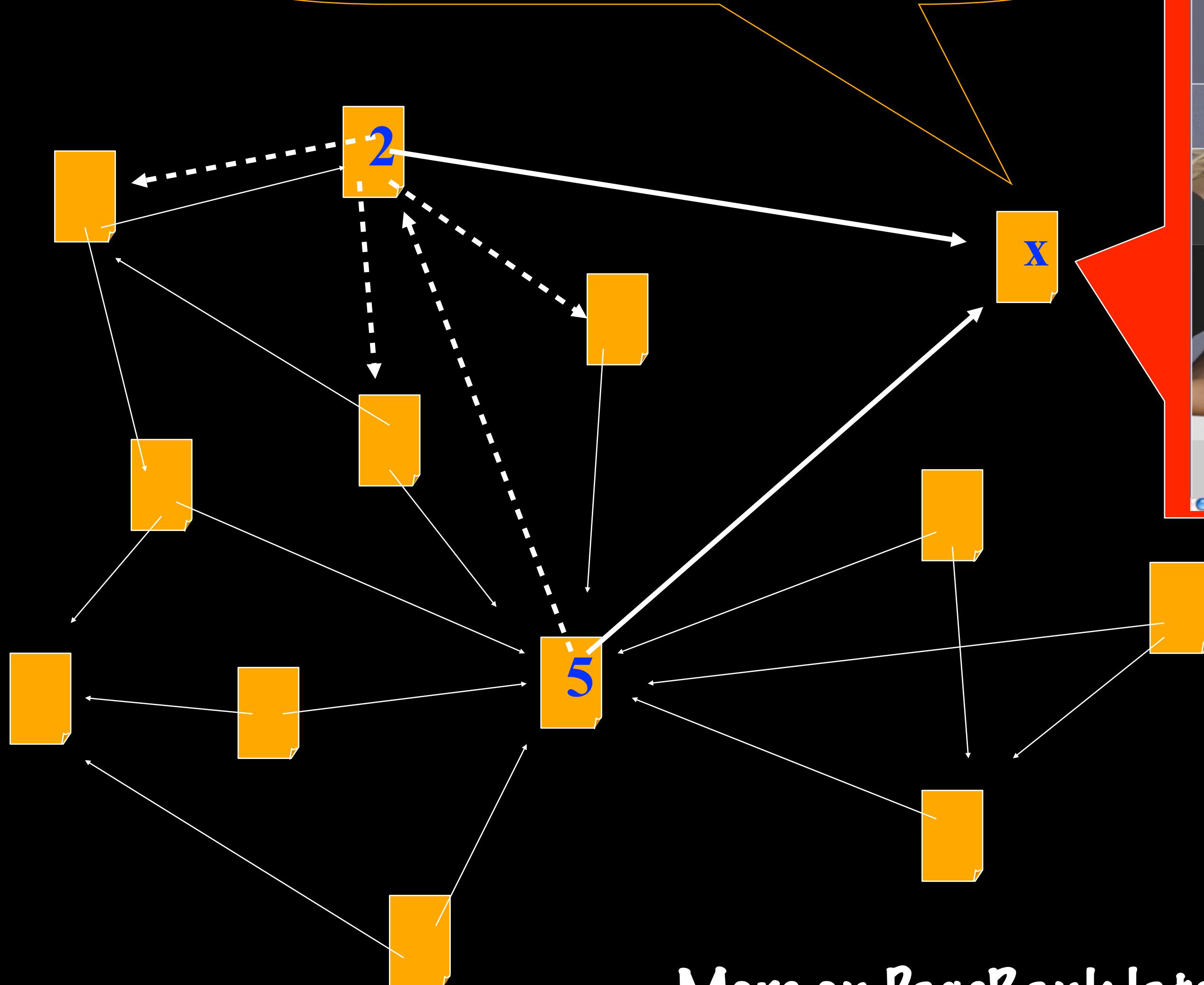
[Britney Spears guide to Semiconductor Physics: semiconductor ...](#)  
Britney Spears lectures on semiconductor physics, radiative and non-radiative transitions, edge emitting lasers and VCSELs. ...  
[britneyspears.ac/lasers.htm](#) - 13k - [Cached](#) - [Similar pages](#)

[Britney Spears Portal - pics, lyrics, MP3s and more!](#)  
Britney Spears pics, lyrics, MP3s, news, gossip, fan sites, forums, and much more!  
Britney Spears Portal, ... ); ');. Britney Spears Portal, ...  
[www.britney-spears-portal.com/](#) - 25k - [Cached](#) - [Similar pages](#)

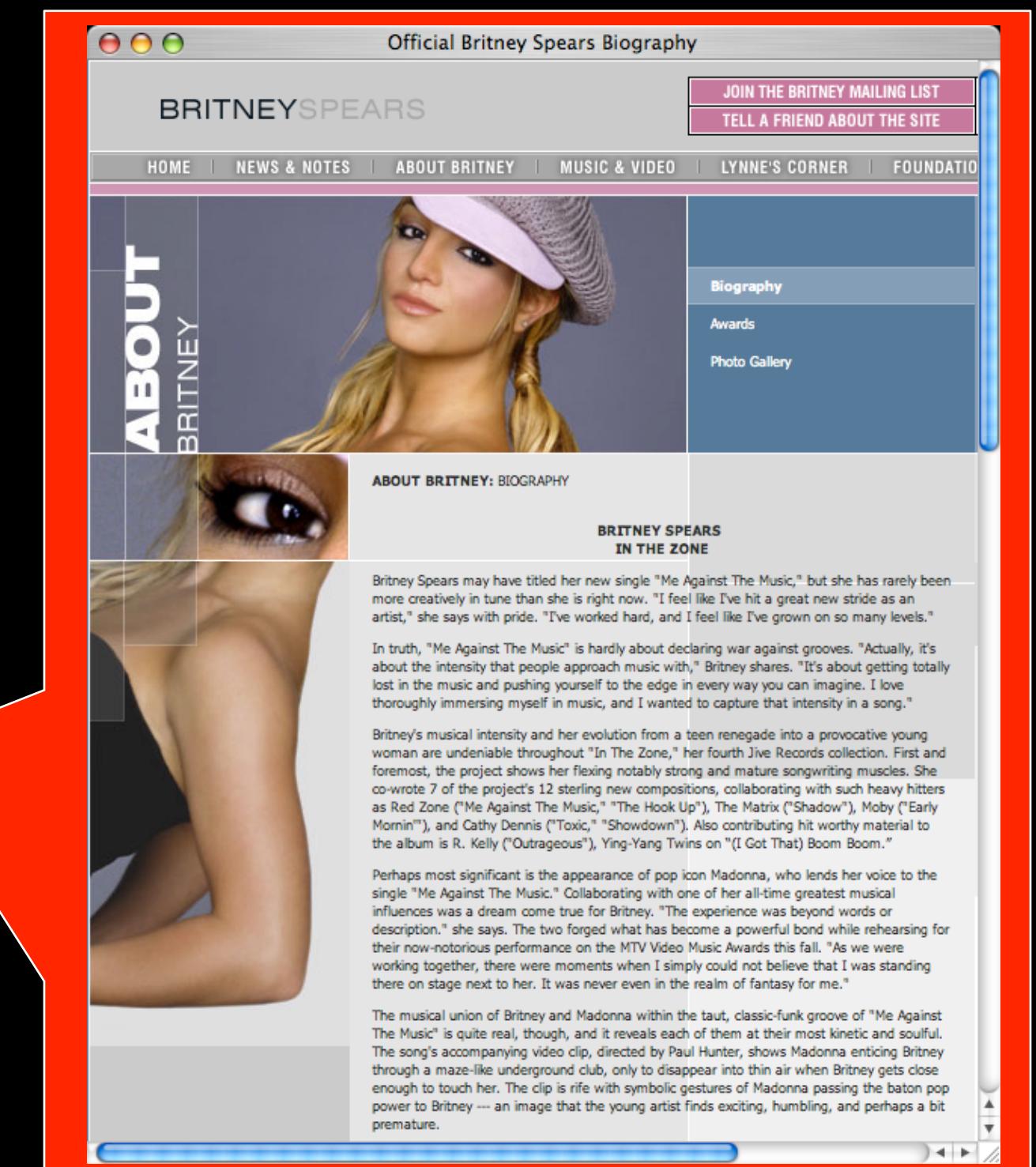
[BritneySpears.org: Your online guide to Britney!](#)  
A comprehensive Britney Spears fansite which pays tribute to Britney with the most active message board, daily news, many pictures, desktop media and more. ...  
[www.britneyspears.org/](#) - 78k - Jun 14, 2004 - [Cached](#) - [Similar pages](#)



$$\text{PageRank}(x) = 5/2 + 2/4 = 3$$



More on PageRank later..



# Measures of importance

- \* In the Web it is important to know what pages are more important, more prestigious, more reliable, more trustworthy, more central, etc... Why?
  - \* Search
  - \* Spam
  - \* ...?

# Measures of importance/prestige for the Web

- \* Graph-based notions of centrality
- \* In-degree, closeness centrality, betweenness
- \* No single measure is suited for all applications
- \* Google's PageRank (Brin and Page 1998)
- \* Measure of prestige for every Web page
- \* Analogous to a random walk with random jumps

# PageRank



# PageRank

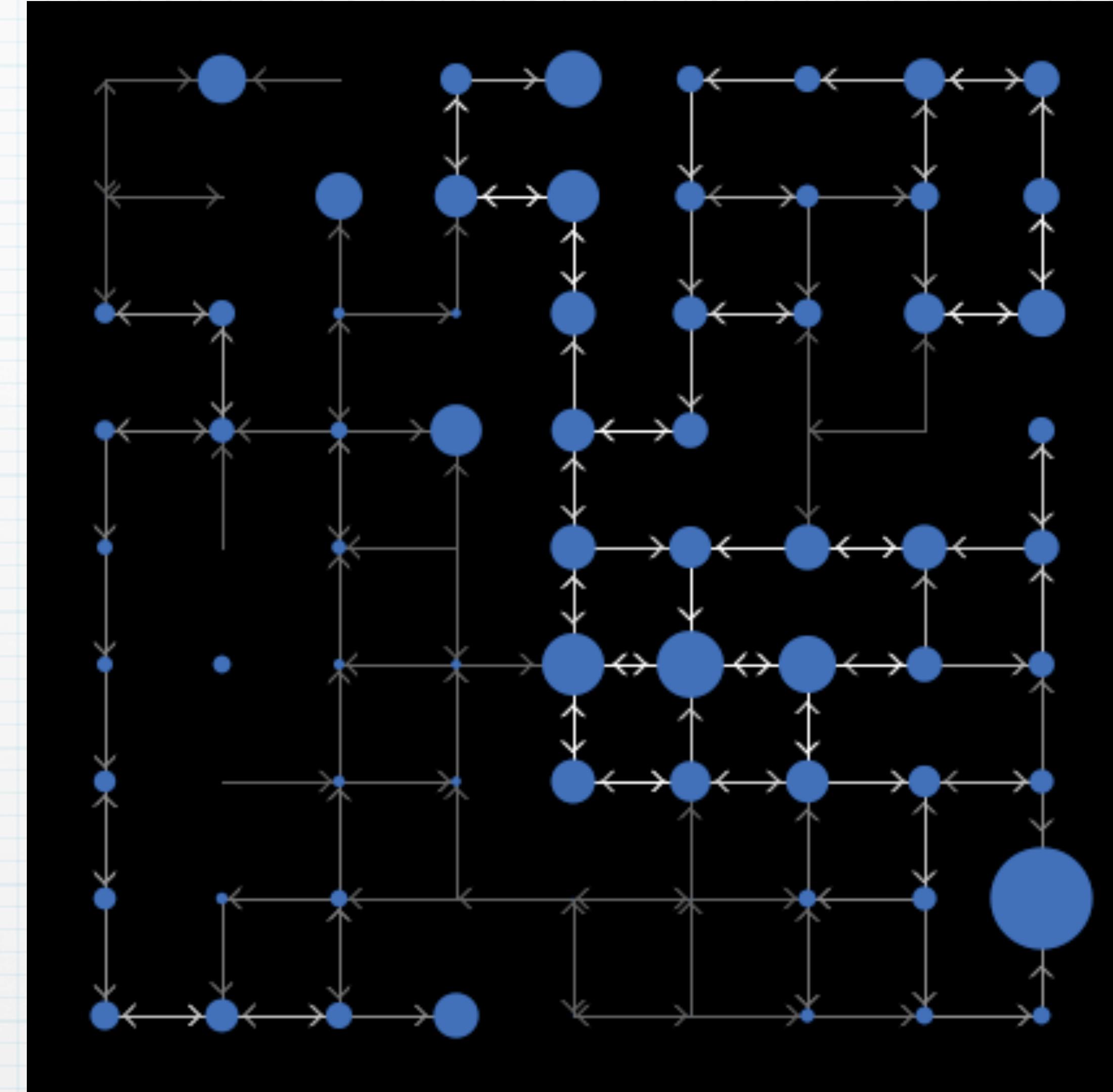
- \* Brin and Page introduced a measure of “page importance” or “prestige” called PageRank in their seminal Google paper (1998)
- \* Invented by Seeley in 1949 (few good ideas are original)
- \* Related to HITS authority (Kleinberg)
- \* Computed from hyperlink structure
- \* Today, PageRank is just one of many (>200) ranking criteria — but an important one!

# PageRank

- \* Brin and Page introduced a measure of “page importance” or “prestige” called PageRank in their seminal Google paper (1998)
- \* Invented by Seeley in 1949 (few good ideas are original)
- \* Related to HITS authority (Kleinberg)
- \* Computed from hyperlink structure
- \* Today, PageRank is just one of many (>200) ranking criteria — but an important one!
- \* Idea: how often will a random web surfer visit a page?
- \* Pages with more inlinks are visited more frequently and thus are more “important”
- \* Recursive definition: the importance of a page is given by the importance of the pages that link to it!

# NetLogo

- \* Diffusion on directed network
- \* Does it converge?

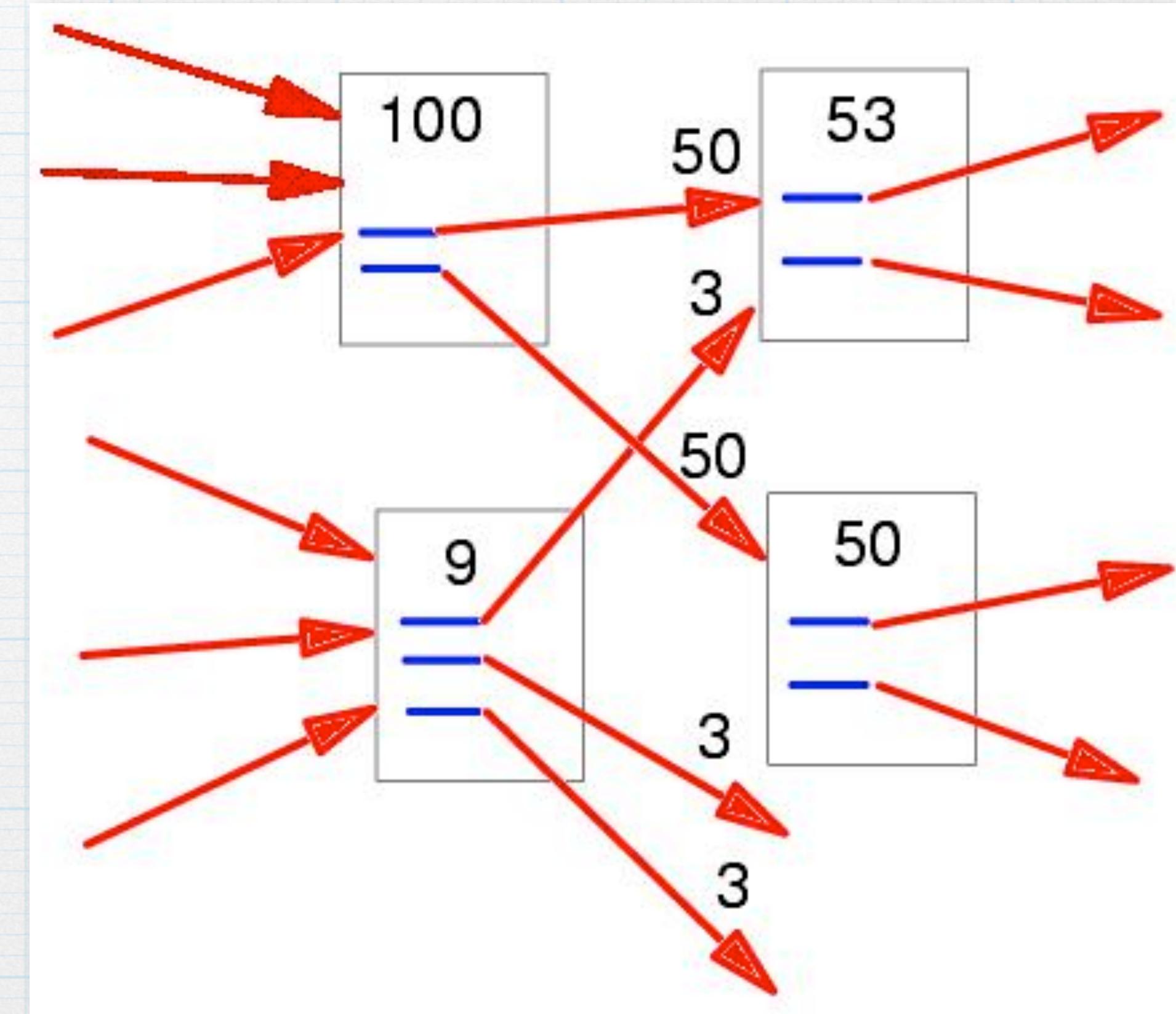


# Simplified PageRank definition

- PageRank is received via inlinks and distributed via outlinks
- PageRank is “conserved”: divided among outlinks

$$R(i) = \sum_{j:j \rightarrow i} \frac{R(j)}{k_{out}(j)}$$

- Note recursive definition!



# Power (iterative) computation of PR

- Initialize:  $R_0(i) = \frac{1}{N}$  (unimportant)
- Repeat for each time step  $t$ , until convergence:

$$R_{t+1}(i) = \sum_{j:j \rightarrow i} \frac{R_t(j)}{k_{out}(j)}$$

importance of page i

pages j that link to page i

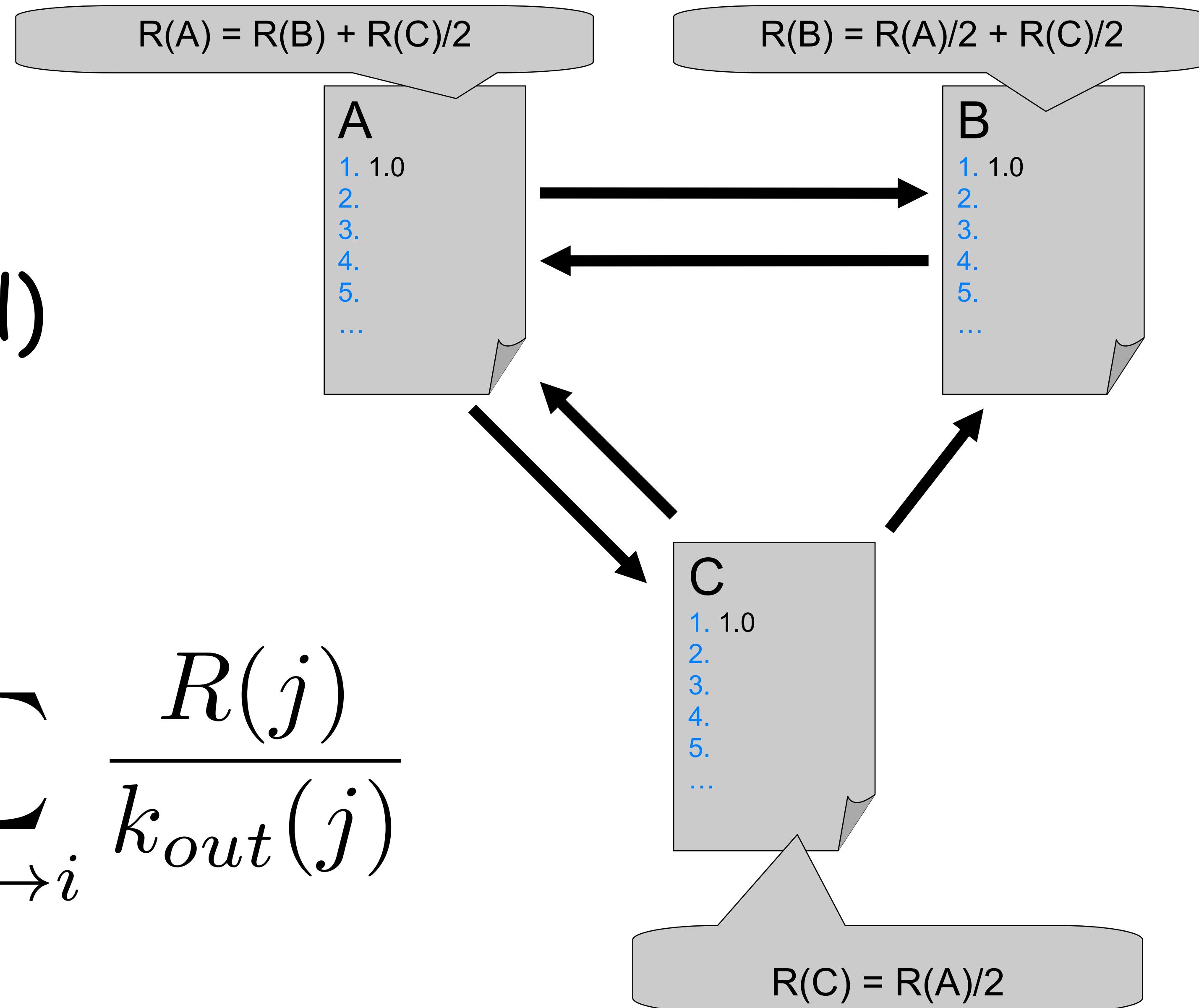
importance of page j

out-degree of page j



# (Simplified) PageRank exercise

$$R(i) = \sum_{j:j \rightarrow i} \frac{R(j)}{k_{out}(j)}$$



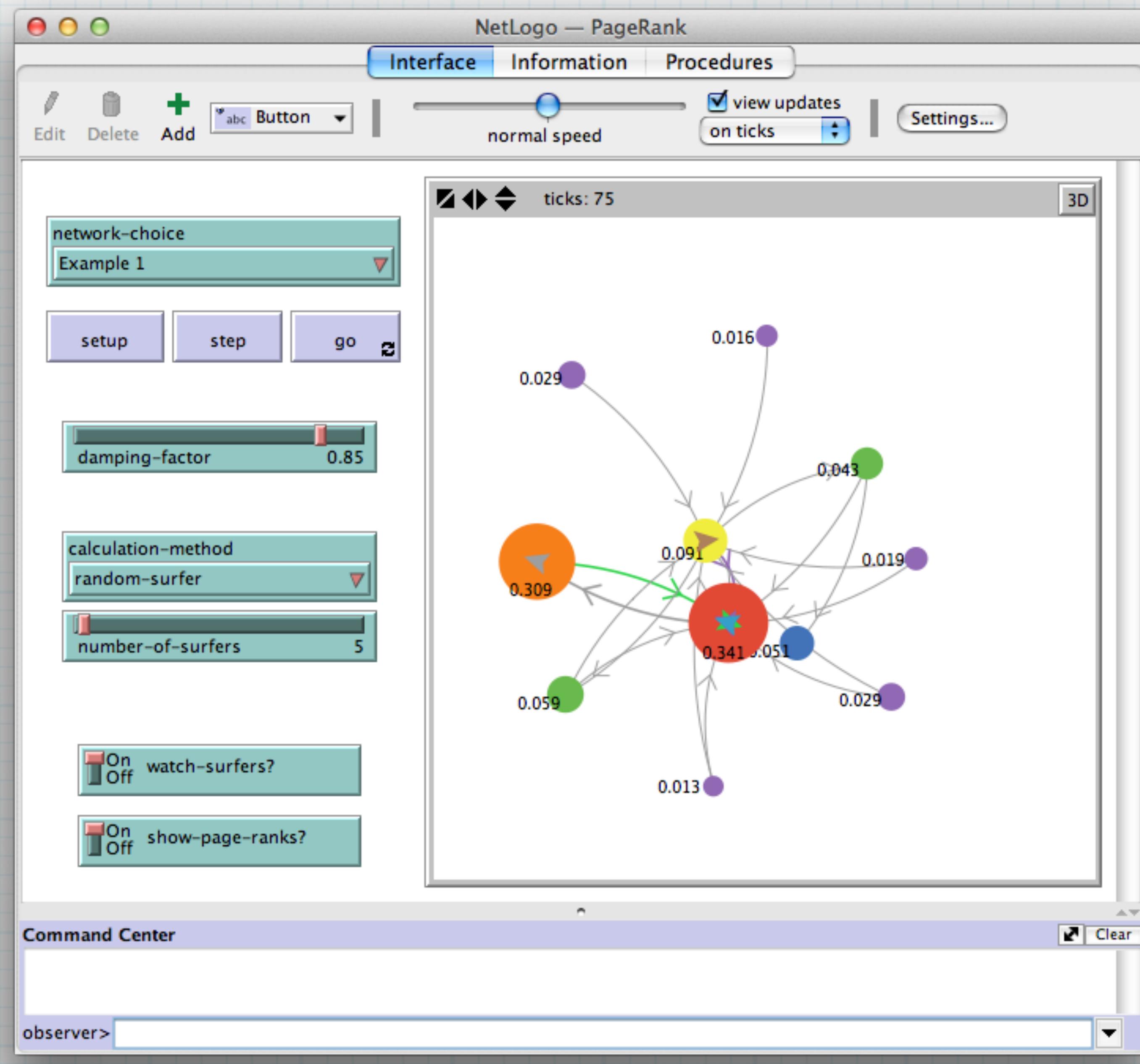
# Actual PageRank

- Problems:
  - Pages without inlinks would quickly converge to PR=0
  - Pages without outlinks (aka **dangling links**) do not share their PR, keep growing
  - Some pages cannot be reached from some others
- Solution: **random jumps (teleportation)**
  - Each page has a minimum PR, typically  $\alpha = 0.15$  (jump or teleportation factor)
  - Damping factor  $(1-\alpha)$ , typically 0.85
- Another issue:
  - Cliques of links to artificially inflate PR
  - Search engine use various measures to combat “spamdexing”

$$R(i) = \frac{\alpha}{N} + (1 - \alpha) \sum_{j:j \rightarrow i} \frac{R(j)}{k_{out}(j)}$$

# NetLogo

- \* PageRank model
- \* Note two methods:
  - \* random surfer
  - \* diffusion (power)



How does PR work  
for ranking search results?  
Let's try a toy version!

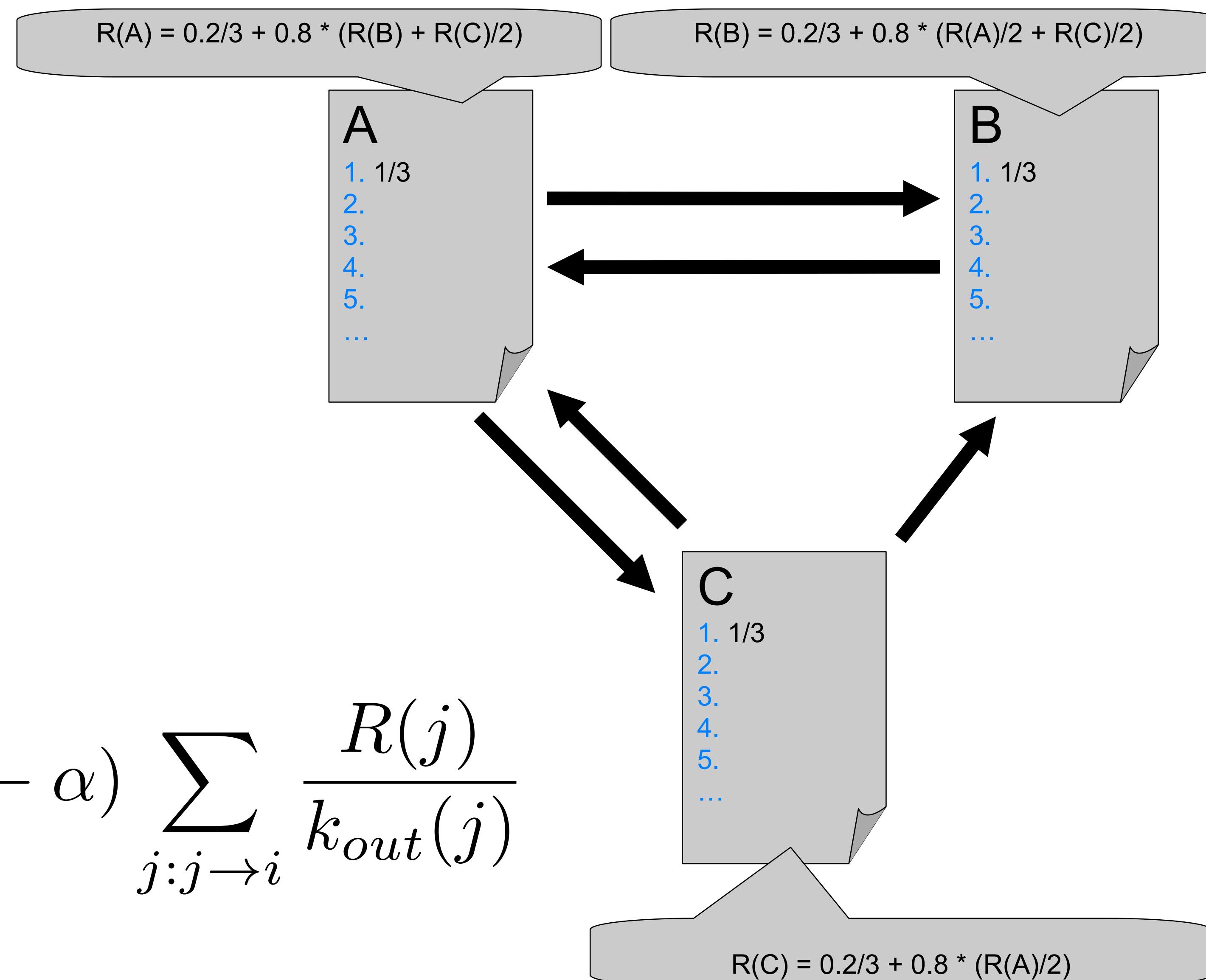
[carl.cs.indiana.edu/fil/cgi-bin/pagerank/](http://carl.cs.indiana.edu/fil/cgi-bin/pagerank/)  
[go.iu.edu/pagerank](http://go.iu.edu/pagerank)



## (Actual) PageRank exercise

$\alpha = 0.2$

$$R(i) = \frac{\alpha}{N} + (1 - \alpha) \sum_{j:j \rightarrow i} \frac{R(j)}{k_{out}(j)}$$

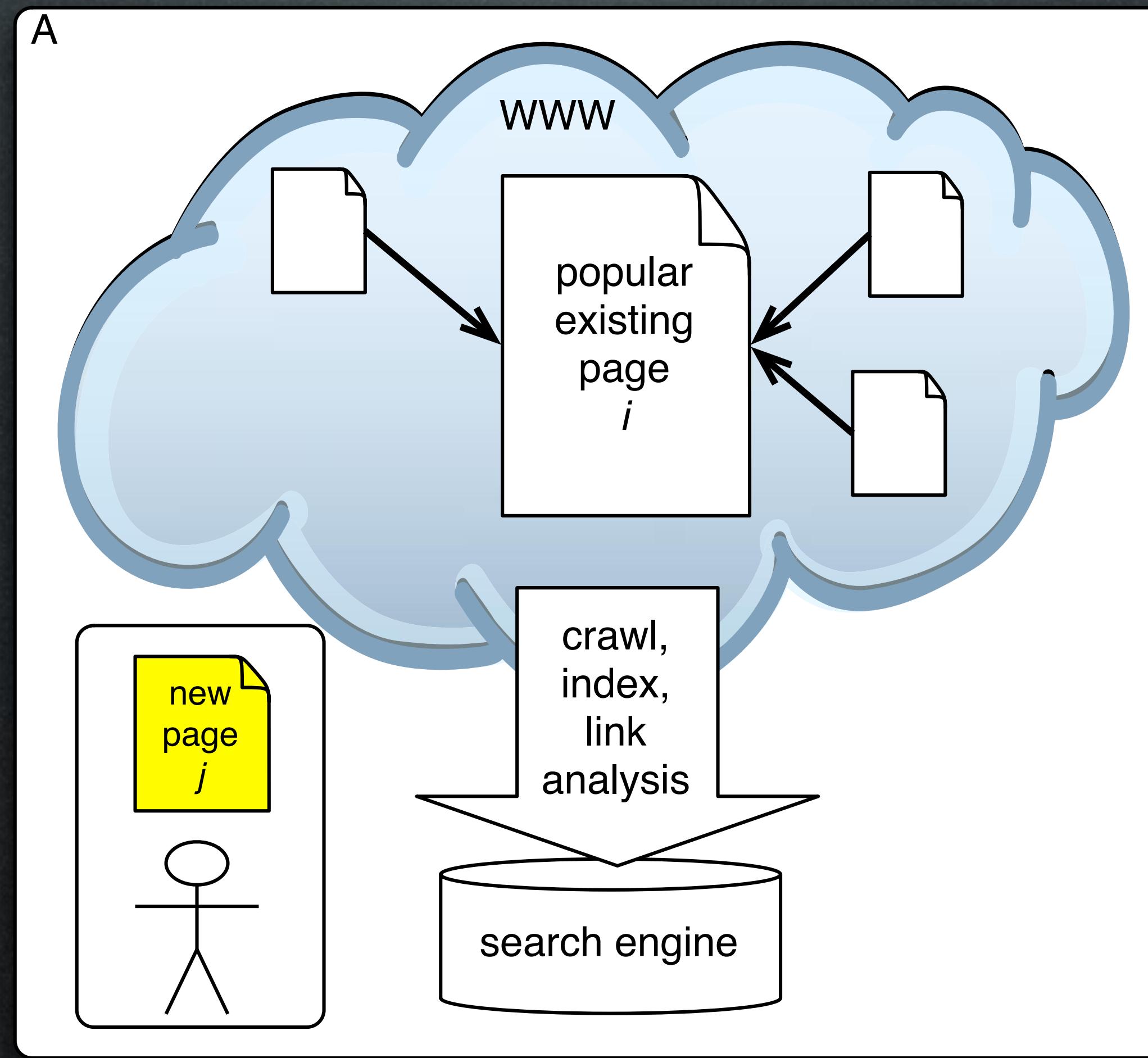


Also see the PR calculator at [http://www.webworkshop.net/pagerank\\_calculator.php](http://www.webworkshop.net/pagerank_calculator.php)

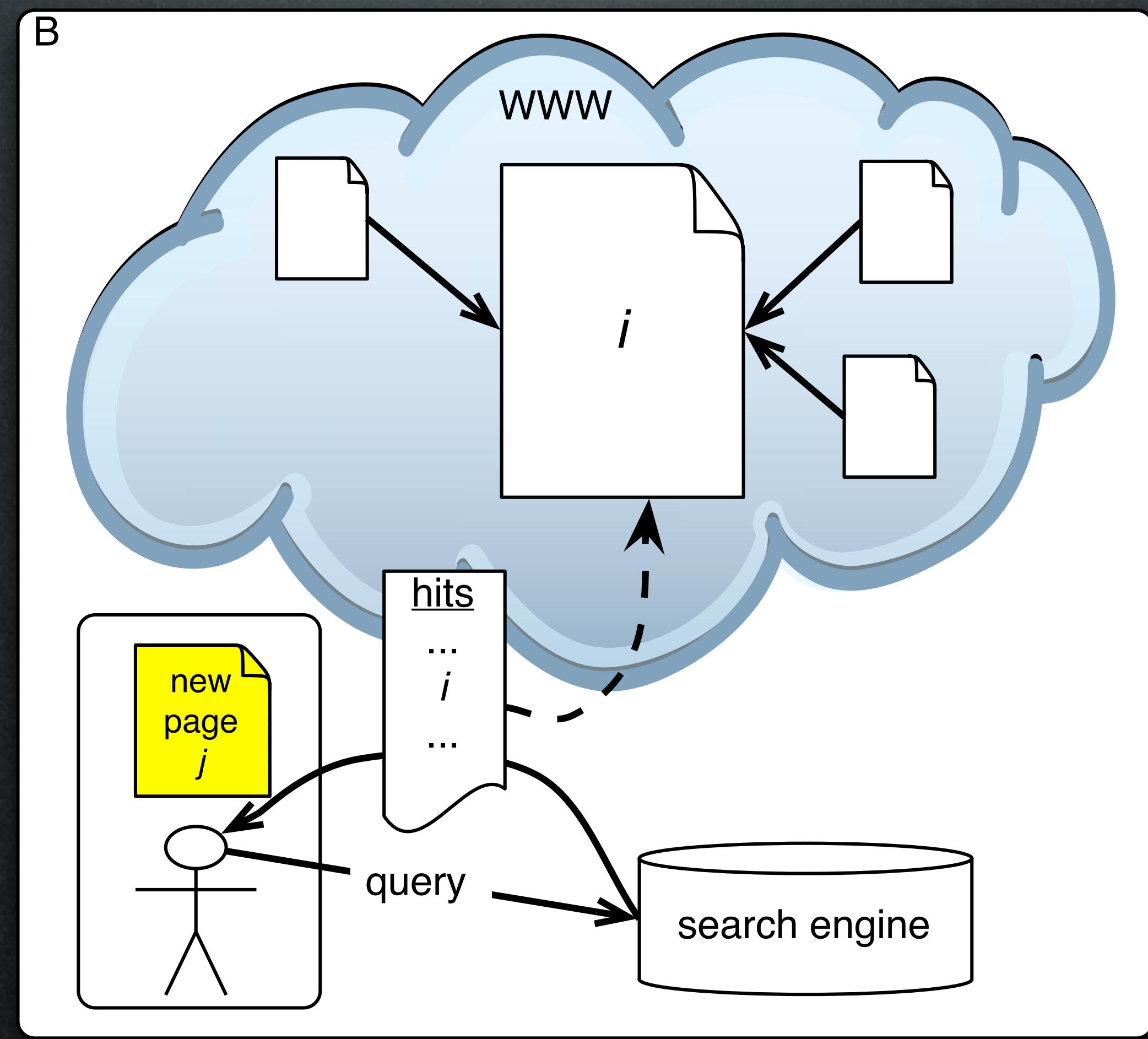
# PageRank competition

- \* Extra credit
- \* Go to  
[carl.cs.indiana.edu/fil/cgi-bin/pagerank/i368.cgi](http://carl.cs.indiana.edu/fil/cgi-bin/pagerank/i368.cgi)  
or [bit.ly/i368pagerank](http://bit.ly/i368pagerank)
- \* Several times until end of classes: top 3 students get some points
- \* Strategy!

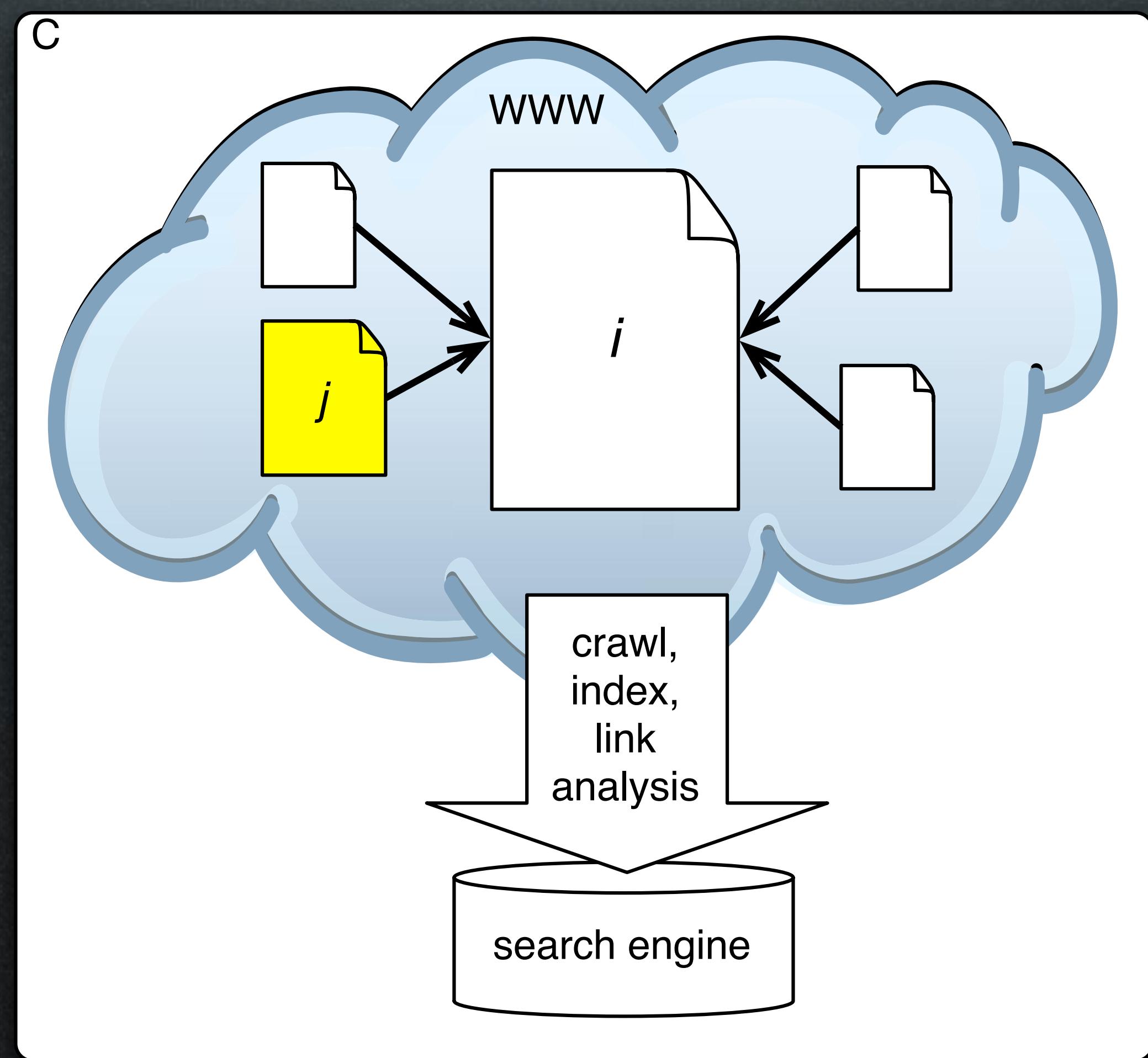
# Is there a popularity bias? ("Entrenchment", "Googlearchy")

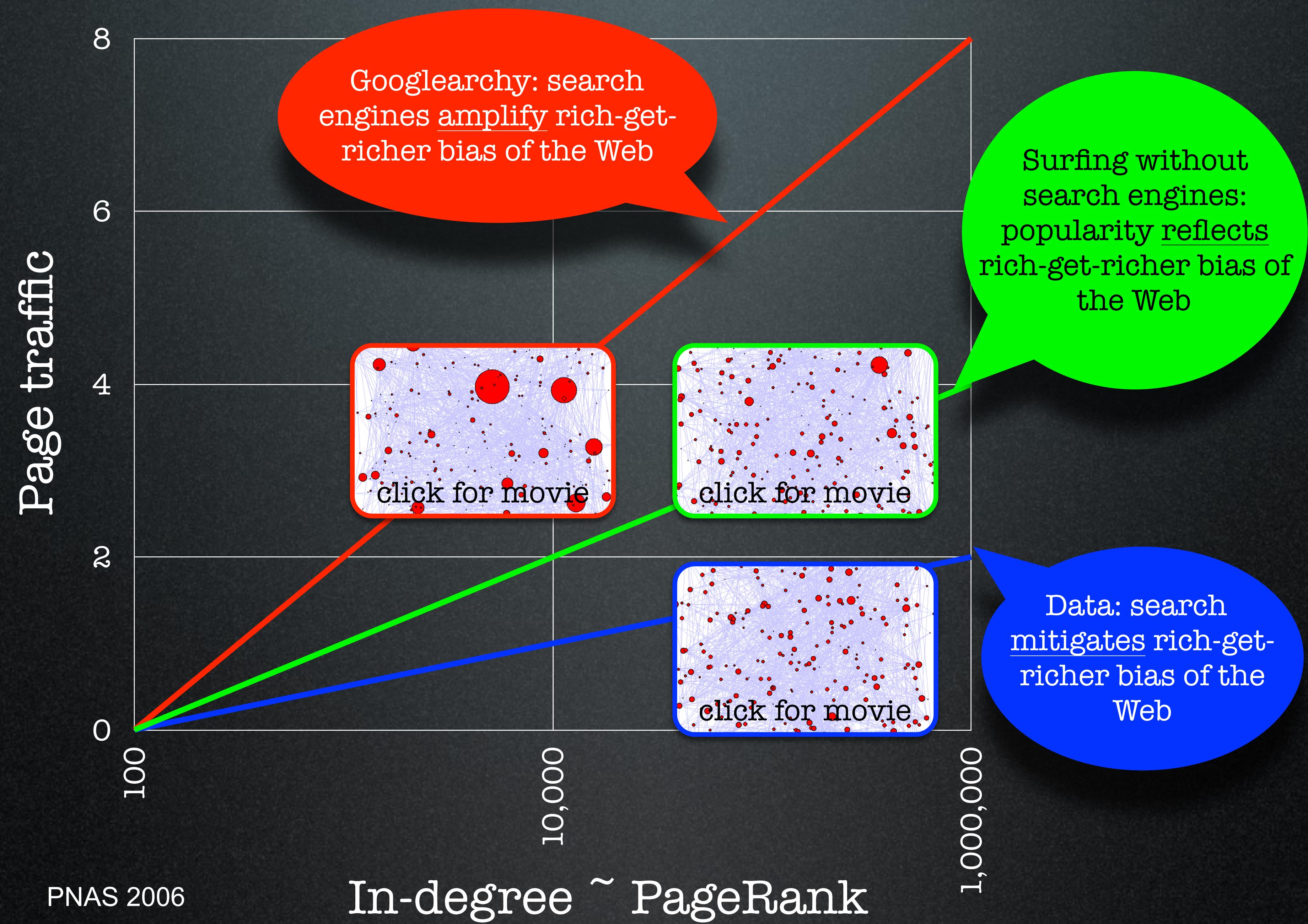


# Is there a popularity bias? ("Entrenchment", "Googlearchy")



# Is there a popularity bias? ("Entrenchment", "Googlearchy")



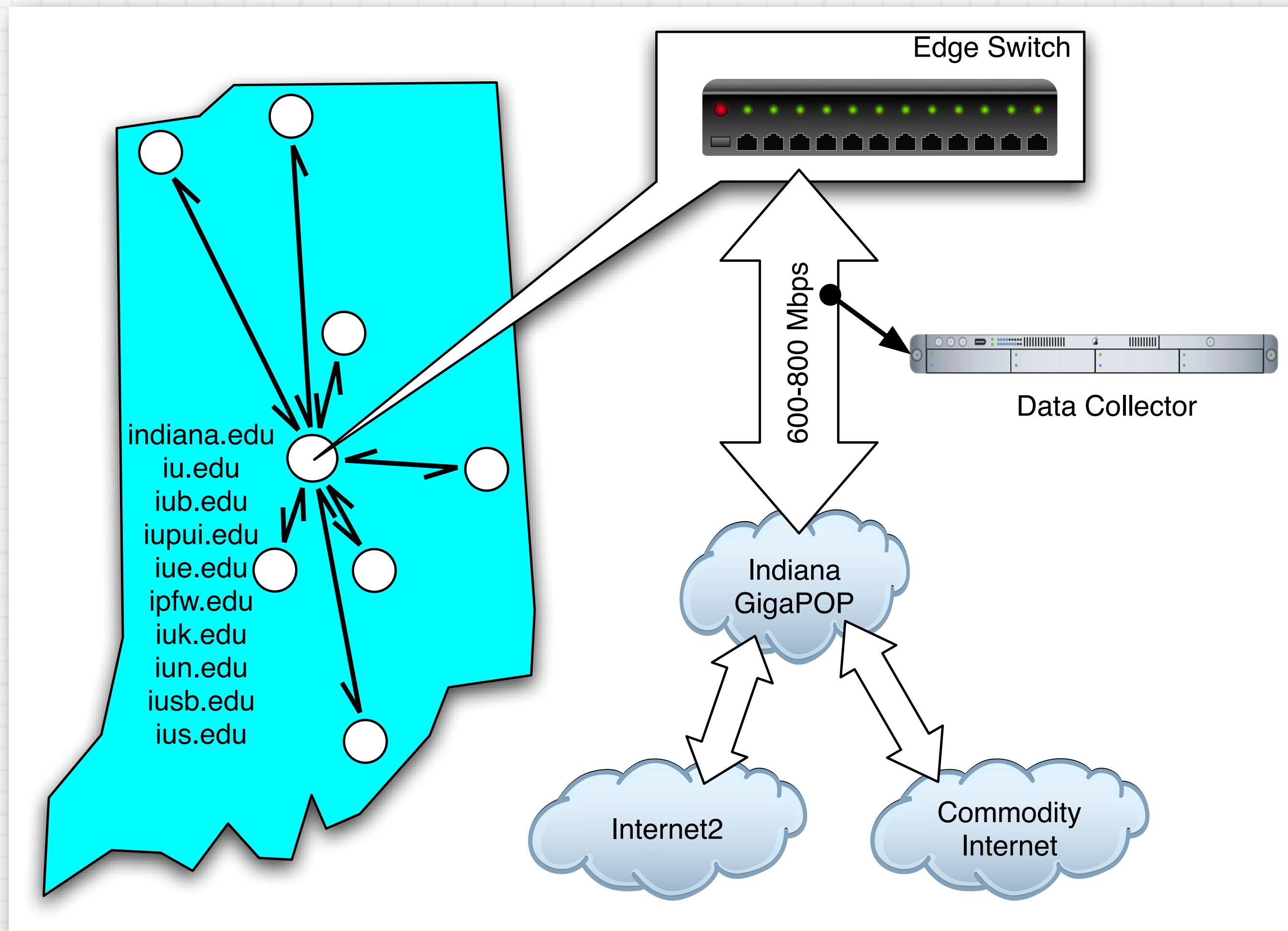


# Is there a popularity bias?

- Yes, in the sense that PageRank is biased by the scale-free structure of the Web graph
- But people's interests and therefore their queries are diverse and many are specific
- Search engines help find obscure pages that would be hard to find by browsing
- Therefore they mitigate the popularity bias of the Web

- PageRank competition!
- Review
  - Web protocol
  - Search engines components
  - Crawlers
  - Essential elements of search ranking
  - PageRank centrality

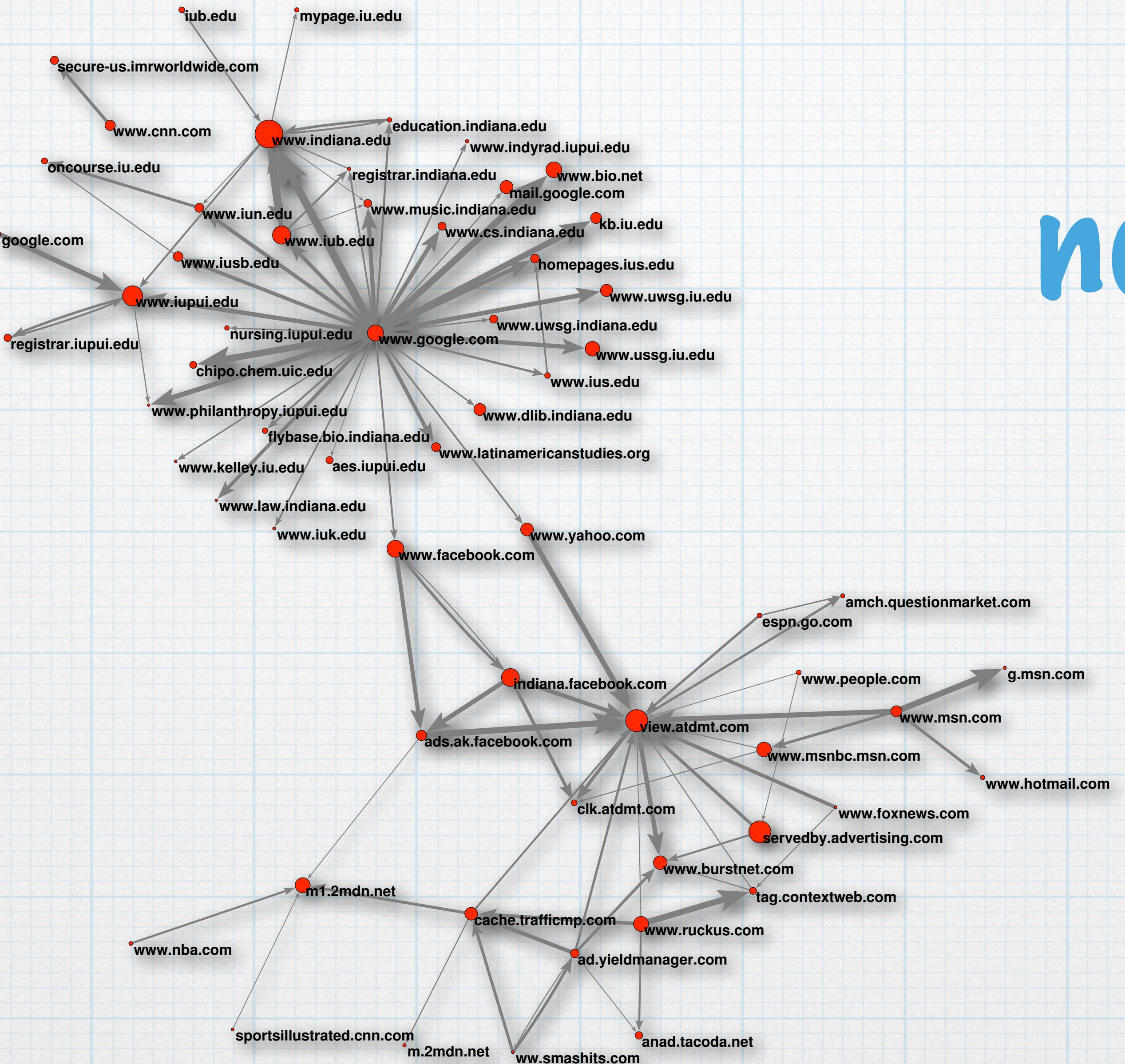
# IU Web traffic



# Traffic networks

If nobody clicks  
on a link, does  
it really exist?

- \* Weighted
- \* Directed



Clicky: Search Results

http://steinbeck.ucs.indiana.edu/~mmeiss/clickbrowser/ Google

## Clicky: Search Results for 'porn'

**7470 results found.**

Sorted by descending order of *instr*

Host	<i>k<sub>in</sub></i>	<i>k<sub>out</sub></i>	<i>s<sub>in</sub></i>	<i>s<sub>out</sub></i>
<a href="#">tgp.pornaccess.com</a>	626	106	112726	105097
<a href="#">free.freepornoofreeporn.com</a>	49	0	99907	0
<a href="#">thumbs.famouspornstars.com</a>	7	0	84016	0
<a href="#">html.freepornofiles.com</a>	81	6	48254	54338
<a href="#">photo.pornotube.com</a>	26	1	46747	76
<a href="#">www.totallypornstars.net</a>	13	0	43857	0
<a href="#">www.adultpornstarguide.com</a>	9	0	34965	0
<a href="#">www.pornpimps.com</a>	11	9	29752	30618
<a href="#">www.youngpornmovies.com</a>	56	123	27788	27179
<a href="#">www.wildpornreviews.com</a>	248	33	23781	17248
<a href="#">www.pornotube.com</a>	46	14	22619	54389
<a href="#">www.bravoporn.com</a>	17	13	21901	24446
<a href="#">www.megapornstarvids.com</a>	33	108	21895	27394
<a href="#">playah.itsyourporn.com</a>	3	4	21630	21405
<a href="#">products.adultlegalporn2go.com</a>	4	4	20894	26663
<a href="#">www.8teenporn.com</a>	33	88	19482	19535
<a href="#">img.porno-pics-free.com</a>	4	0	19125	0
<a href="#">www.pornstarfinder.net</a>	67	168	18540	18154
<a href="#">galleries2.porn365.com</a>	17	0	17339	0
<a href="#">www.pornmoviepage.com</a>	42	75	16852	17035
<a href="#">www.porno-vids.com</a>	80	13	15578	14265
<a href="#">www.porneskimo.com</a>	41	214	14464	29019
<a href="#">pic.freeporn.hu</a>	27	0	14220	0
<a href="#">www.jointheporn.com</a>	124	8	11560	10585
<a href="#">images.porninspector.com</a>	4	1	11353	44
<a href="#">www.tgpornstars.com</a>	22	105	11248	10831
<a href="#">www.duckyporn.com</a>	45	338	11213	13314
<a href="#">www.pornstarscope.com</a>	38	66	11206	13357
<a href="#">images.pornvddirect.com</a>	4	0	11019	0
<a href="#">www.hpornstars.com</a>	46	37	10874	10561
<a href="#">usemyporn.com</a>	2	19	10651	15706
<a href="#">www.famouspornstars.com</a>	63	287	10565	82240
<a href="#">scripts.sunporno.com</a>	4	0	10550	0

Clicky: Search Results

<http://steinbeck.ucs.indiana.edu/~mmeiss/clickbrowser/>

## Clicky: Search Results for 'porn'

7470 results found.

Sorted by descending order of *instr*

Host	k <sub>in</sub>	k <sub>out</sub>	s <sub>in</sub>	s <sub>out</sub>
<a href="#">tgp.pornaccess.com</a>	626	106	112726	105097
<a href="#">free.freepornoofreeporn.com</a>	49	0	99907	0
<a href="#">thumbs.famouspornstars.com</a>	7	0	84016	0
<a href="#">html.freepornofiles.com</a>	81	6	48254	54338
<a href="#">photo.pornotube.com</a>	26	1	46747	76
<a href="#">www.totallypornstars.net</a>	13	0	43857	0
<a href="#">www.adultpornstarguide.com</a>	9	0	34965	0
<a href="#">www.pornpimps.com</a>	11	9	29752	30618
<a href="#">www.youngpornmovies.com</a>	56	123	27788	27179
<a href="#">www.wildpornreviews.com</a>	248	33	23781	17248
<a href="#">www.pornotube.com</a>	46	14	22619	54389
<a href="#">www.bravoporn.com</a>	17	13	21901	24446
<a href="#">www.megapornstarvids.com</a>	33	108	21895	27394
<a href="#">playah.itsyourporn.com</a>	3	4	21630	21405
<a href="#">products.adultlegalporn2go.com</a>	4	4	20894	26663
<a href="#">www.8teenporn.com</a>	33	88	19482	19535
<a href="#">img.porno-pics-free.com</a>	4	0	19125	0
<a href="#">www.pornstarfinder.net</a>	67	168	18540	18154
<a href="#">galleries2.porn365.com</a>	17	0	17339	0
<a href="#">www.pornmoviepage.com</a>	42	75	16852	17035
<a href="#">www.porno-vids.com</a>	80	13	15578	14265
<a href="#">www.porneskimo.com</a>	41	214	14464	29019
<a href="#">pic.freeporn.hu</a>	27	0	14220	0
<a href="#">www.jointheporn.com</a>	124	8	11560	10585
<a href="#">images.porninspector.com</a>	4	1	11353	44
<a href="#">www.tgpornstars.com</a>	22	105	11248	10831
<a href="#">www.duckyporn.com</a>	45	338	11213	13314
<a href="#">www.pornstarscope.com</a>	38	66	11206	13357
<a href="#">images.porndvddirect.com</a>	4	0	11019	0
<a href="#">www.hpornstars.com</a>	46	37	10874	10561
<a href="#">usemyporn.com</a>	2	19	10651	15706
<a href="#">www.famouspornstars.com</a>	63	287	10565	82240
<a href="#">scripts.sunporno.com</a>	4	0	10550	0

Clicky: Viewing Node 'tgp.pornaccess.com'

[\(Do another search.\)](#)

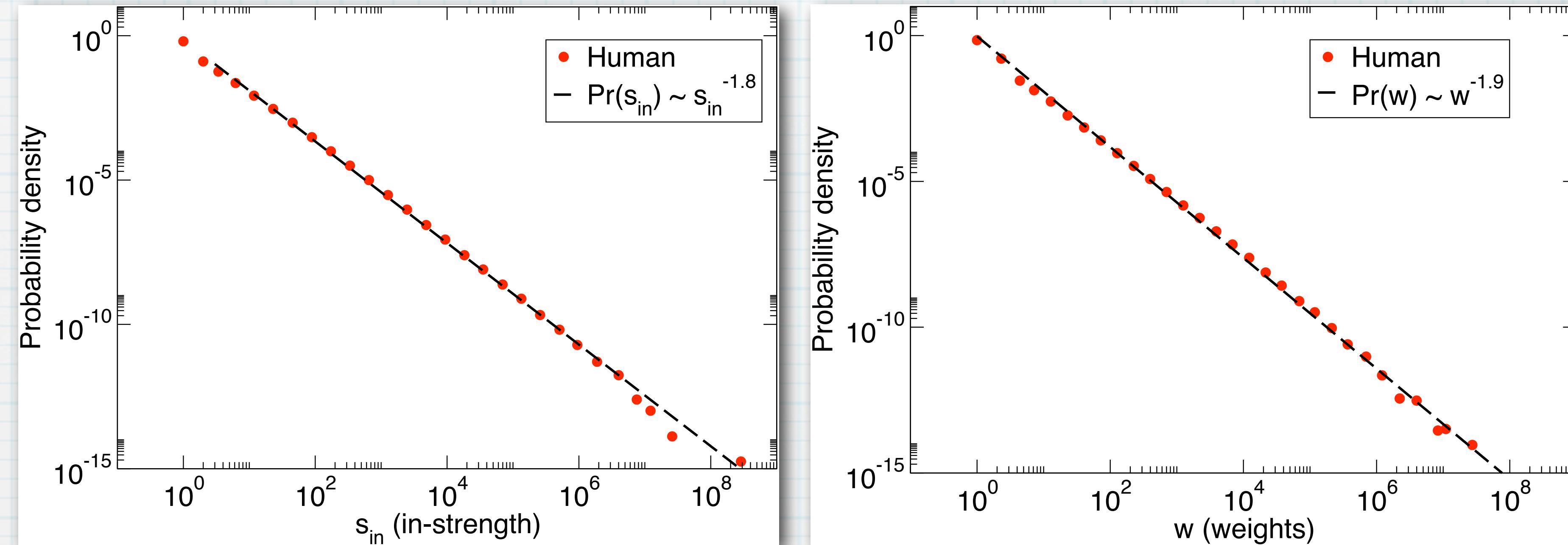
Incoming Links for [tgp.pornaccess.com](#): (112726 clicks from 626 hosts)

Host	Weight
<a href="#">tgp.pornaccess.com</a>	100533
-	5240
<a href="#">www.tiava.com</a>	386
<a href="#">www.giga galleries.com</a>	305
<a href="#">www.empire18.com</a>	247
<a href="#">www.xnxx.com</a>	195
<a href="#">frogsex.com</a>	187
<a href="#">www.miamatures.com</a>	165
<a href="#">www.catlist.com</a>	158
<a href="#">www.fuckkk.com</a>	153
<a href="#">tiava.com</a>	151
<a href="#">www.searchgalleries.com</a>	142
<a href="#">www.porno-pics-free.com</a>	123
<a href="#">search.askjolene.com</a>	114
<a href="#">www.onlymovies.com</a>	114
<a href="#">www.easygals.com</a>	108
<a href="#">www.boneme.com</a>	95
<a href="#">www.altavista.com</a>	94
<a href="#">www.lolitampegs.com</a>	87
<a href="#">www.milfmovs.com</a>	74
<a href="#">www.movietitan.com</a>	71
<a href="#">www.miateens.com</a>	69
<a href="#">www.freeones.com</a>	66
<a href="#">goatlist.com</a>	65
<a href="#">www.theboobsmovies.com</a>	64
<a href="#">www.searchvids.com</a>	62
<a href="#">www.gaymoviedome.com</a>	61
<a href="#">www.sex oasis.com</a>	61
<a href="#">www.lodita.com</a>	59
<a href="#">www.searchgals.com</a>	54
<a href="#">bigtits-cinema.com</a>	47
<a href="#">www.shaggle.com</a>	41

Outgoing Links for [tgp.pornaccess.com](#): (105097 clicks to 106 hosts)

Host	Weight
<a href="#">tgp.pornaccess.com</a>	100533
<a href="#">tgpvideos.pornaccess.com</a>	3499
<a href="#">sweetliltranny.promo.pornaccess.com</a>	610
<a href="#">track.pornaccess.com</a>	132
<a href="#">vip.pornaccess.com</a>	74
<a href="#">www.pornaccess.com</a>	69
<a href="#">track.qaypornaccess.com</a>	12
<a href="#">shemalemovies.pornaccess.com</a>	11
<a href="#">amateurmovies.test.pornaccess.com</a>	8
<a href="#">allnylonmovies.test.pornaccess.com</a>	7
<a href="#">bustyisland.pornaccess.com</a>	4
<a href="#">1001ultimatetits.pornaccess.com</a>	4
<a href="#">matureaboo.pornaccess.com</a>	4
<a href="#">momshardvideo.promo.pornaccess.com</a>	4
<a href="#">allnylonmovies.pornaccess.com</a>	4
<a href="#">bustymomvideos.pornaccess.com</a>	3
<a href="#">security-updater.com</a>	3
<a href="#">sluttylittlebabysitters.pornaccess.com</a>	3
<a href="#">momshardvideo.test.pornaccess.com</a>	3
<a href="#">teachersandteenagers.pornaccess.com</a>	3
<a href="#">grannyridesagain.pornaccess.com</a>	3
<a href="#">em.gad-network.com</a>	2
<a href="#">shemalessurprise.pornaccess.com</a>	2
<a href="#">olderchicksfuckingyoungerdicks.pornaccess.com</a>	2
<a href="#">247latexsex.pornaccess.com</a>	2
<a href="#">sexinpublicplaces.pornaccess.com</a>	2
<a href="#">joggs.pornaccess.com</a>	2
<a href="#">babysitters.pornaccess.com</a>	2
<a href="#">teacherspet.pornaccess.com</a>	2
<a href="#">officesexmovies.pornaccess.com</a>	2
<a href="#">hornyoldernymphos.pornaccess.com</a>	2
<a href="#">dadsandtwinks.gaypornaccess.com</a>	2

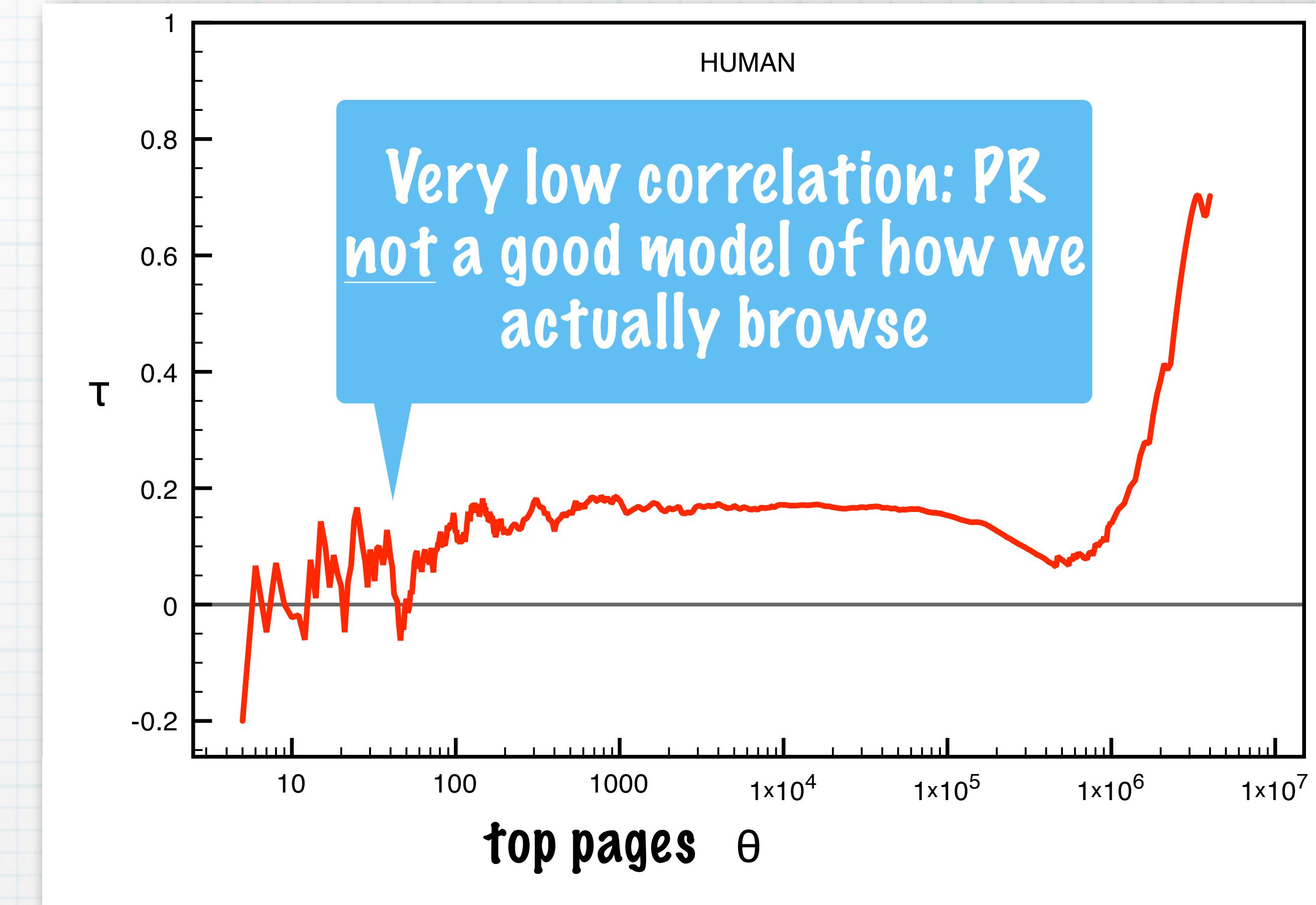
# Web traffic properties



# PageRank revisited

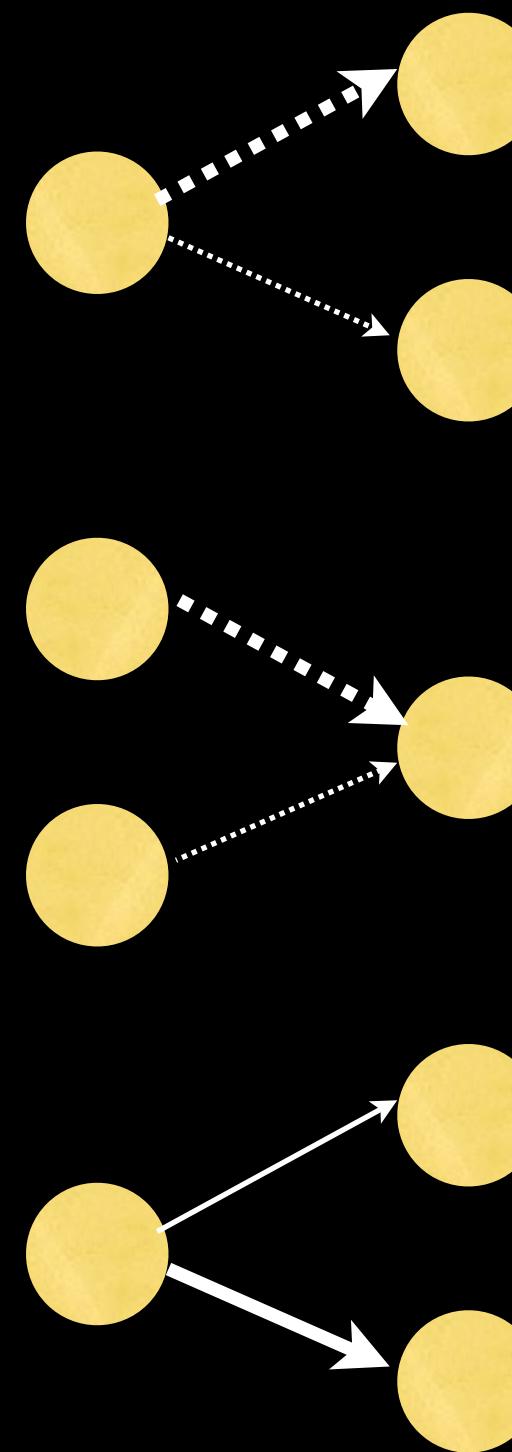
- \* PR as a model of Web navigation: stationary distribution of visit frequency by a modified random walk (with jumps) on the Web graph
- \* Is this a good model of Web traffic?
- \* From an application perspective, we care about the ranking of sites
- \* Compare with actual site traffic (in-strength) by measuring correlation between ranking by PR and ranking by s

# Kendall's rank correlation



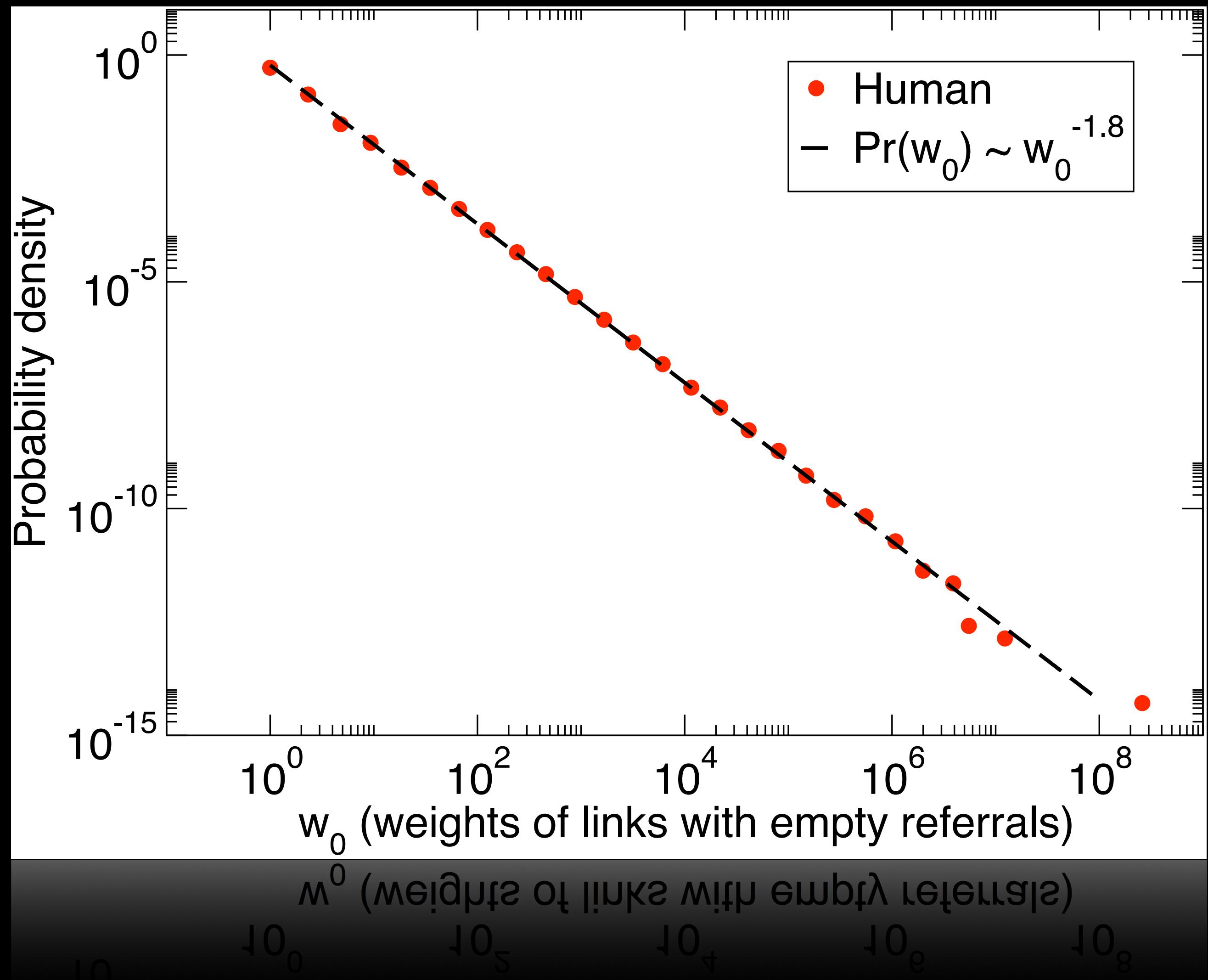
# PageRank assumptions

1. Equal probability of teleporting to each of the nodes
2. Equal probability of teleporting from each of the nodes
3. Equal probability of following each link from any given node

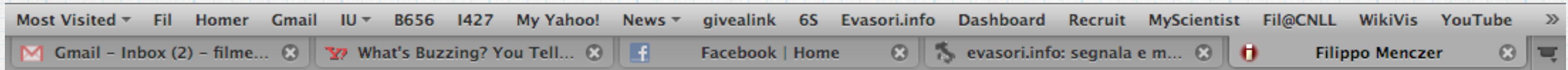


$$PRW(j) = \frac{\alpha}{N} + (1 - \alpha) \sum_{i:w_{ij} \neq 0} \frac{w_{ij}}{s_{out}(i)} PRW(i)$$

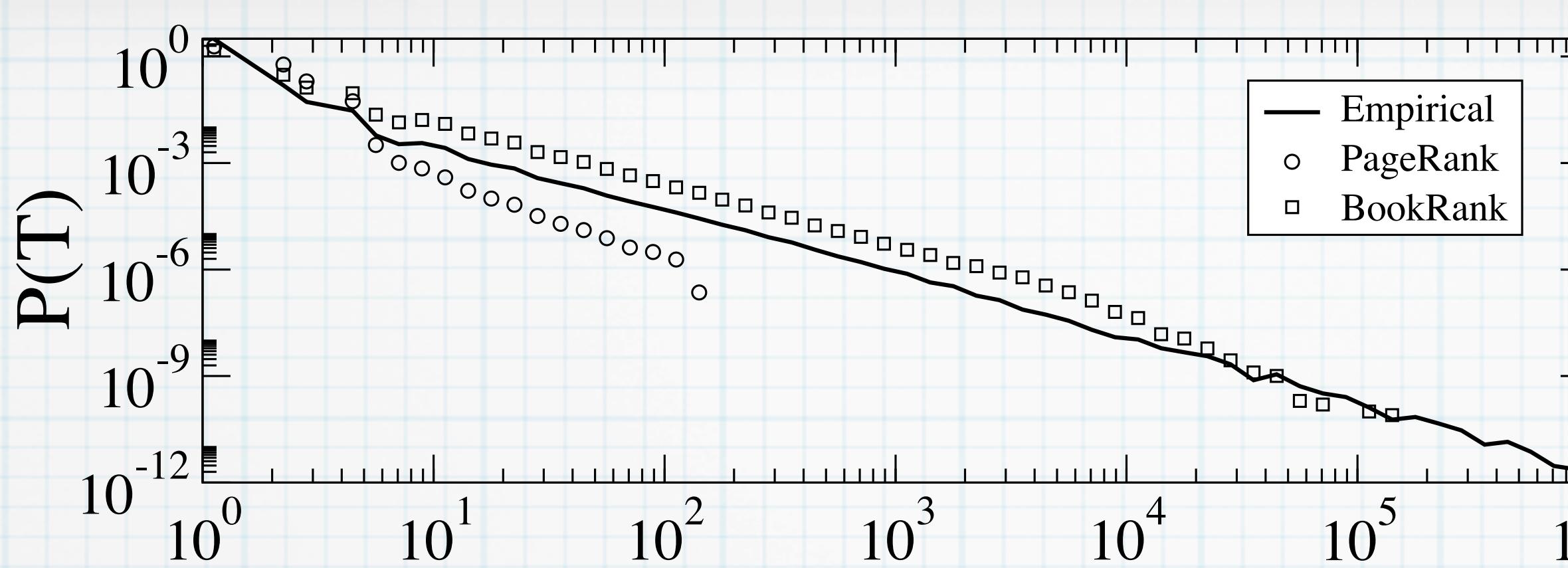
# Teleportation target heterogeneity



# Ingredients for an agent-based model

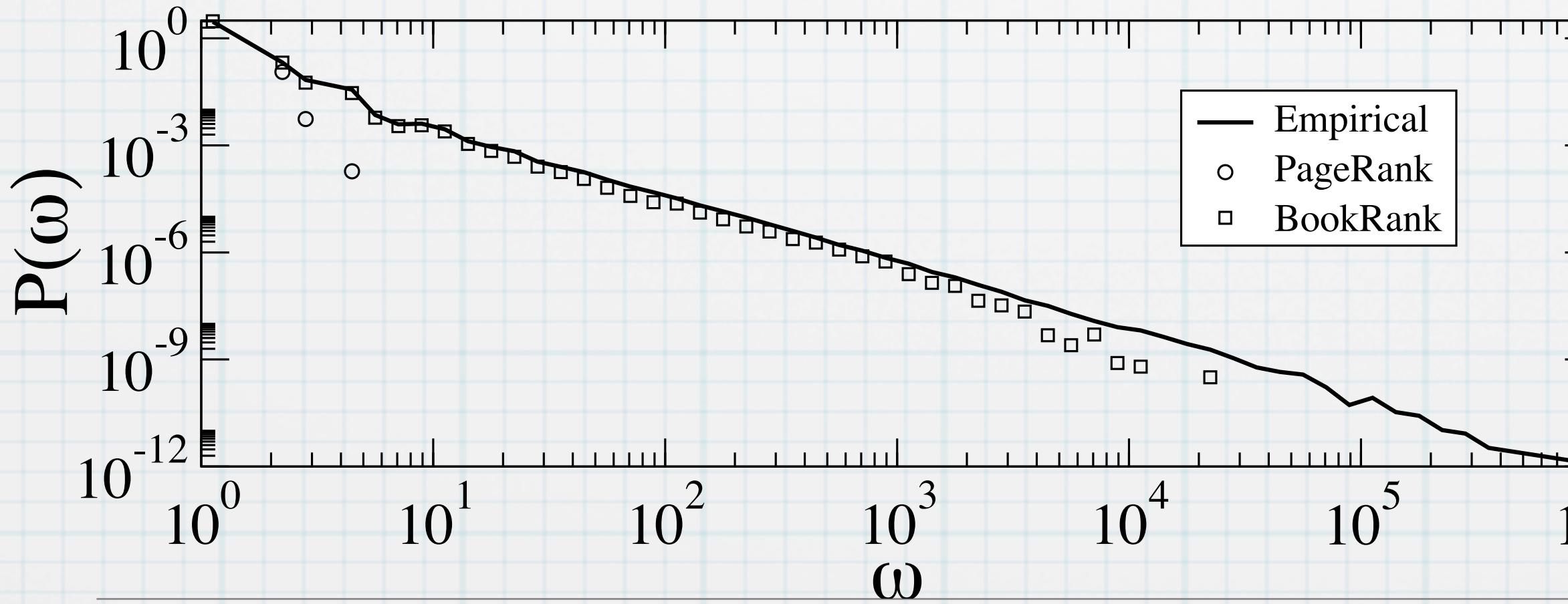


1. Bookmarks (memory): we remember the pages we like and are more likely to start browsing from those

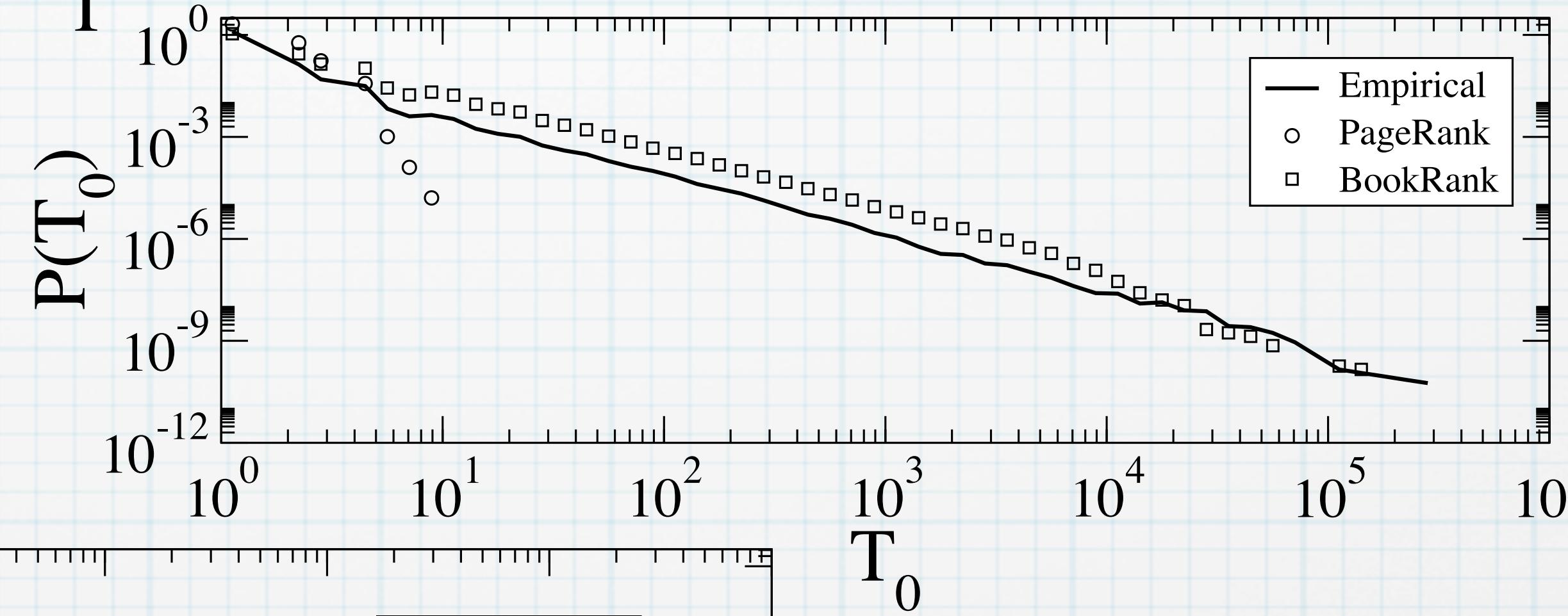


# Page Traffic

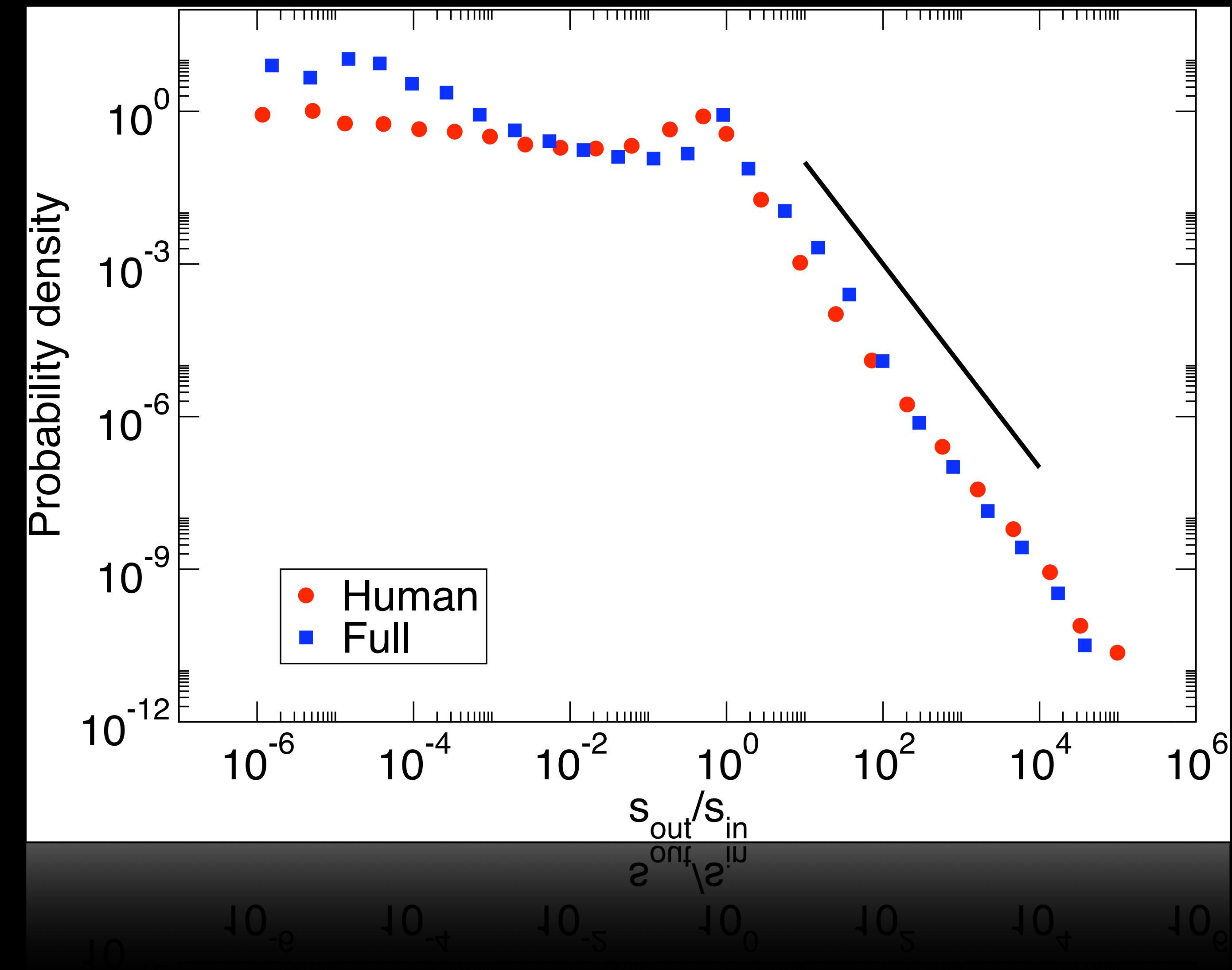
## Jump Traffic



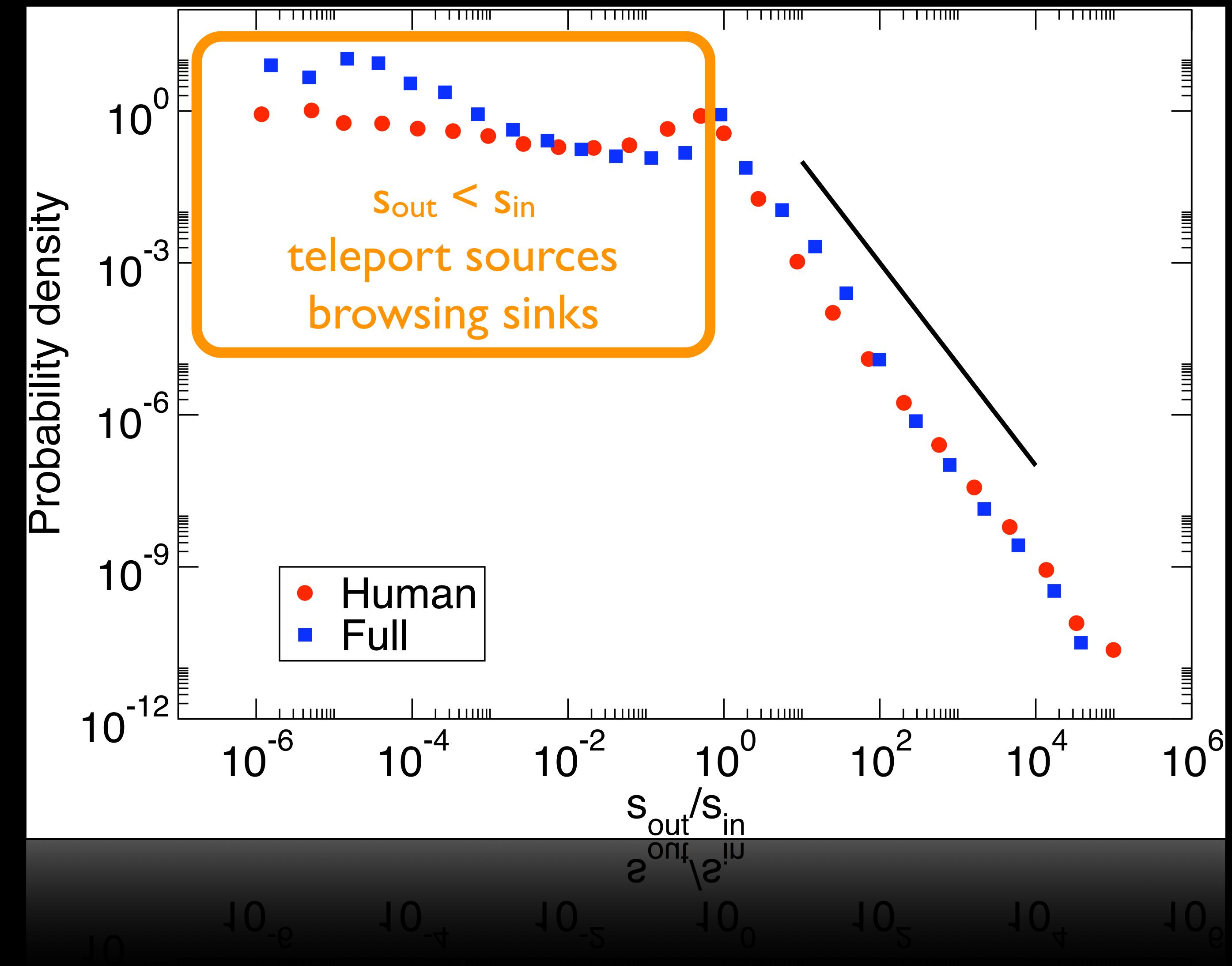
## Link Traffic



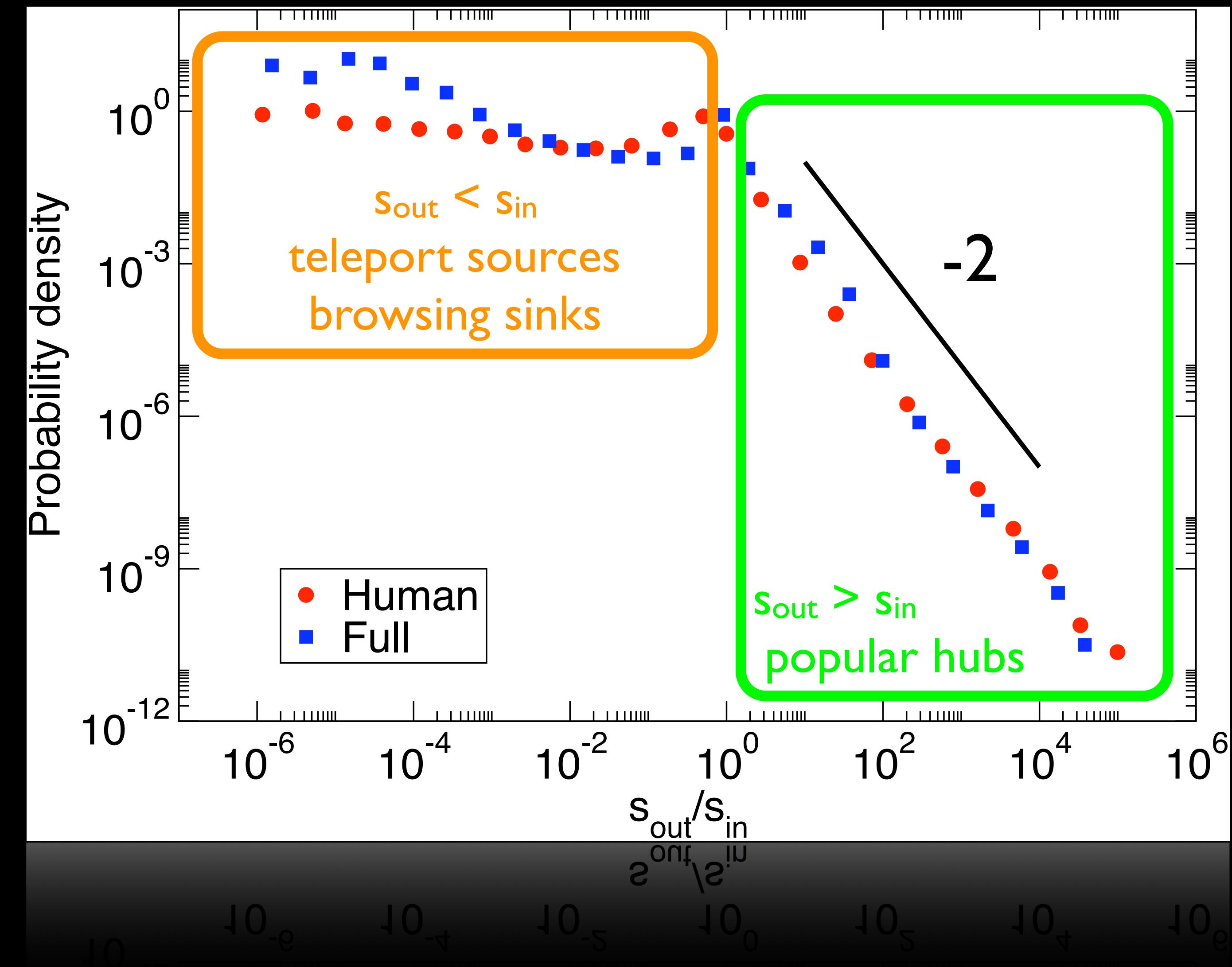
Teleportation  
source  
heterogeneity:  
“hubness”



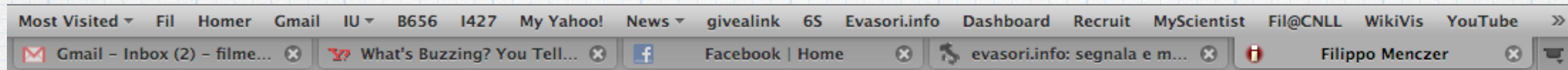
# Teleportation source heterogeneity: “hubness”



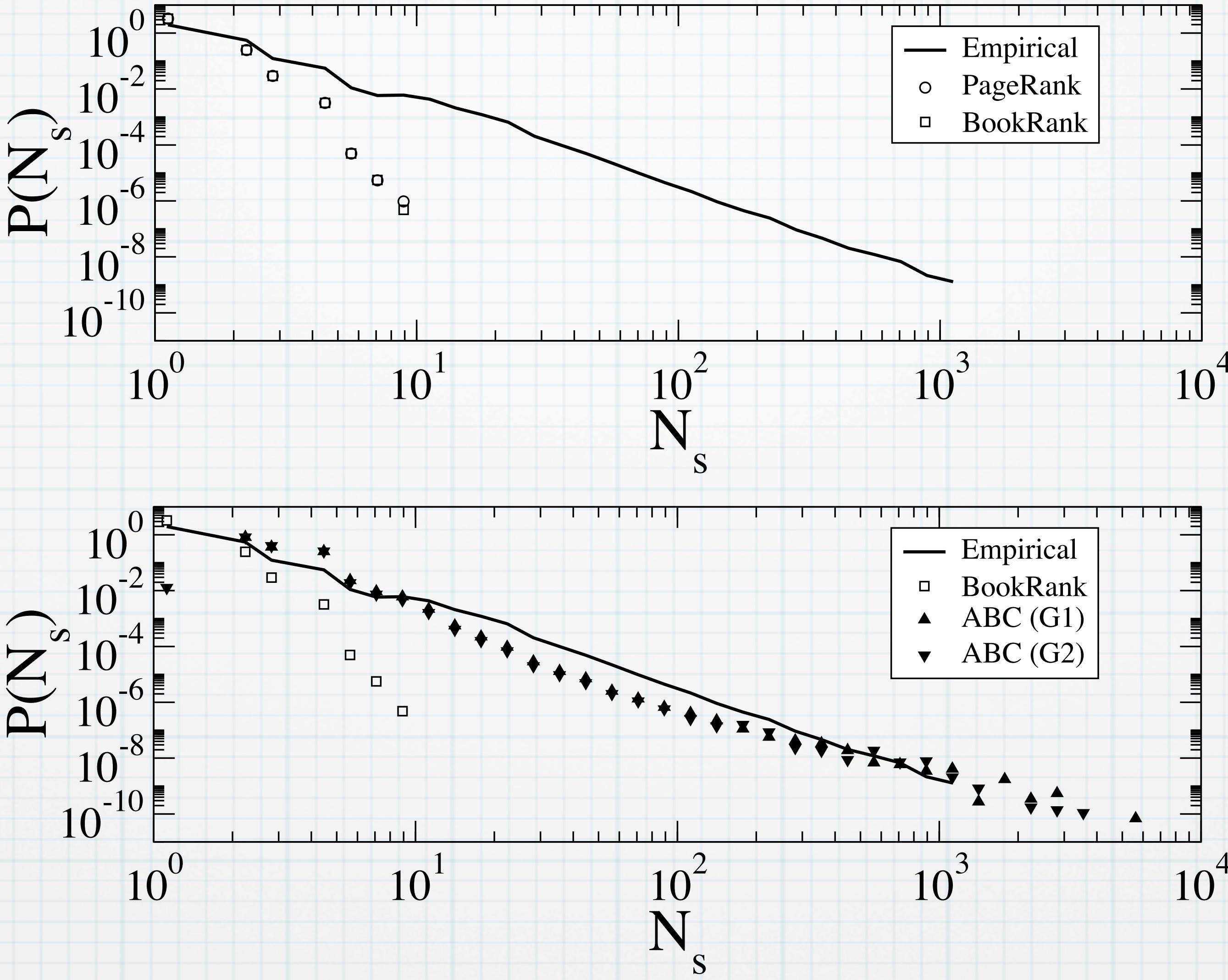
# Teleportation source heterogeneity: “hubness”



# Ingredients for an agent-based model

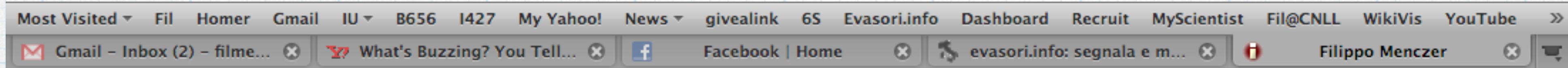


1. Bookmarks (memory)
2. Backtracking (or tabs)
3. Topicality (sessions)



# Session length ( $\varepsilon$ depth)

# Summary

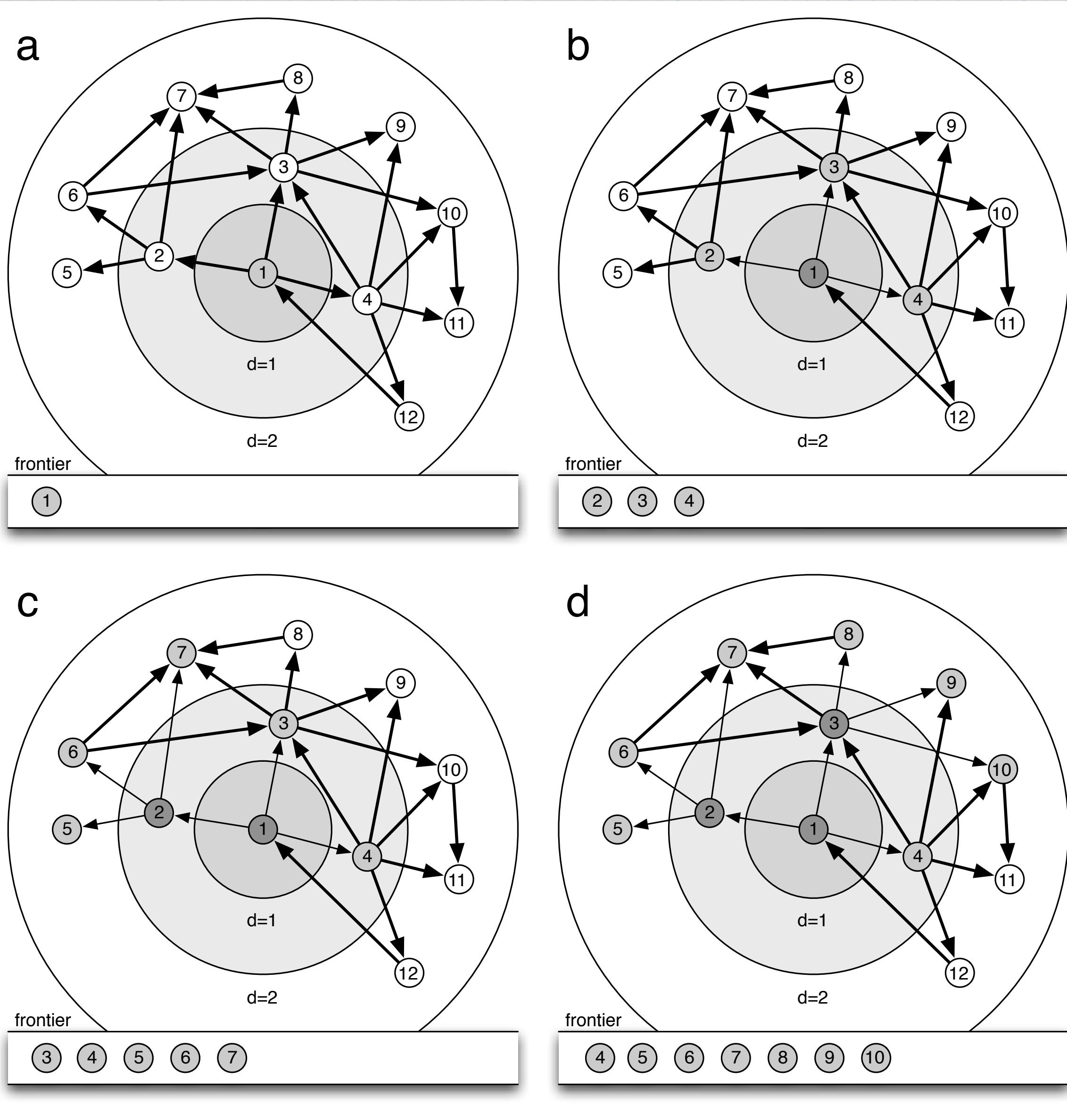


- \* Memory/bookmarks allow to reproduce page, jump, and link traffic
- \* To capture full user browsing behavior, including session length, the model needs to also account for backtracking and topicality

- PageRank competition!
- Review
- Web traffic
- PageRank as a model of Web traffic

# Search & navigation in networks

- \* How do we find a short path between two actors to play six degrees of Kevin Bacon?
- \* How do we determine a scholar's Erdős Number?
- \* How do Web crawlers find every page?
- \* How did the subjects in the Migram and Watts experiments forward the messages?
- \* How do we find a file in a peer network?

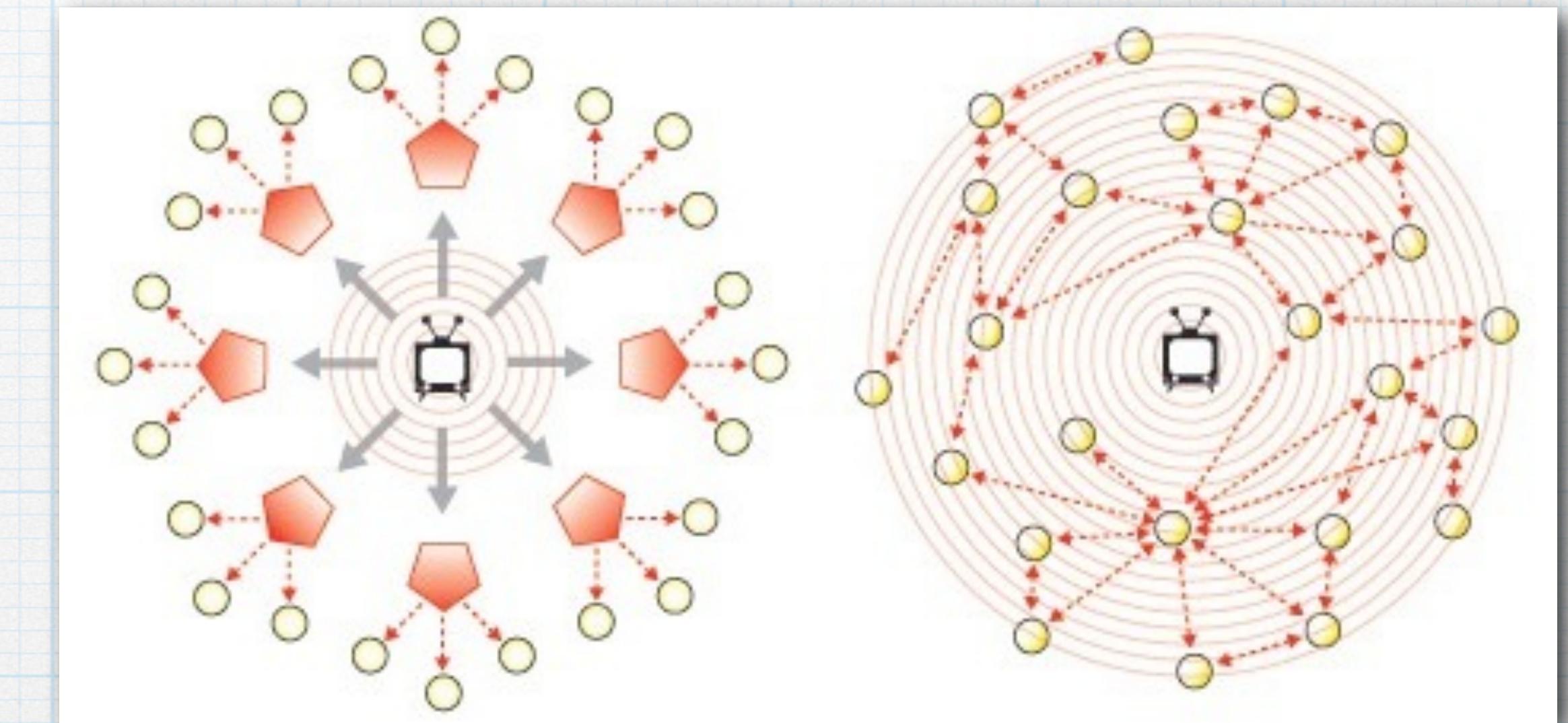


# Remember BFS...

- \* Why not just use breadth-first search?

# Search

- \* Exhaustive (BFS) or centralized search or flooding
- \* Like broadcast TV marketing
- \* Web crawlers
- \* Computer viruses
- \* Gnutella p2p network
- \* Directed or distributed search
- \* Like viral marketing
- \* Greedy by popularity (hubs) or by homophily (same profession, geography, etc.)
- \* Structured peer networks, such as BitTorrent



# Efficient directed search

Theory: In a small world network, short paths (logarithmic length in size of network) should exist between any two nodes.

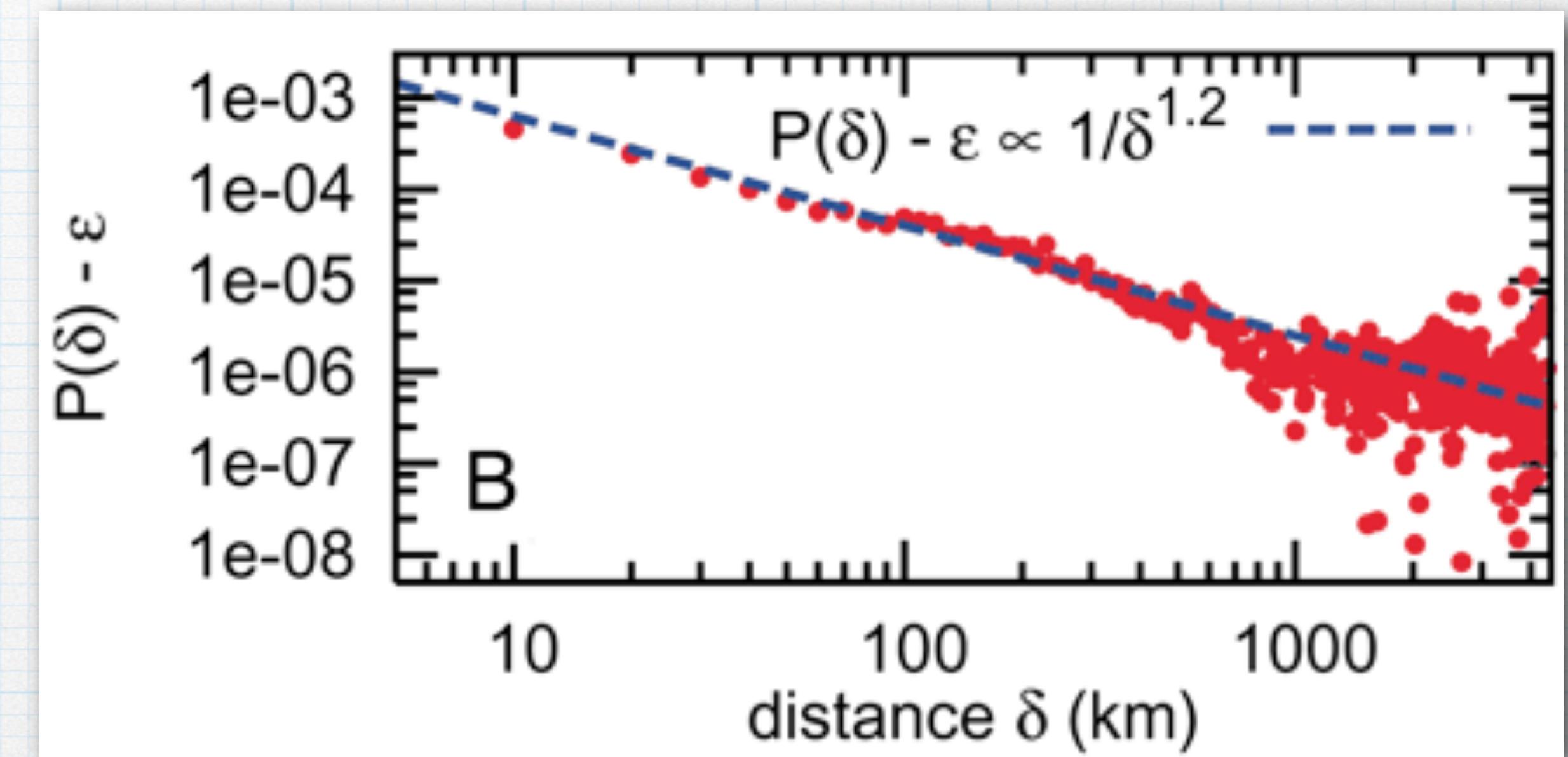
Practice: hard to find them!

Exceptions:

1. Greedy algorithms based on location in geographical small world networks
2. Greedy algorithms based on depth in hierarchical networks
3. Greedy algorithms based on degree in power law networks (go to hubs)

# 1. Geography

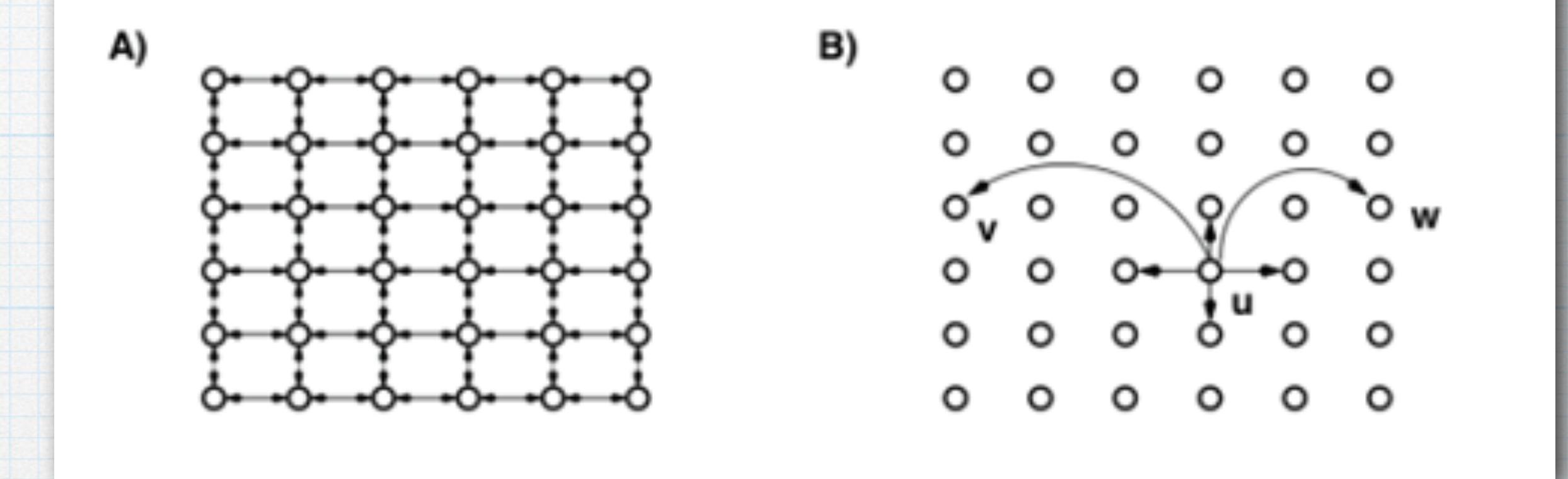
- \* People are more likely to be friends if they live nearby
- \* The probability decreases in a regular way (power law) with distance
- \* Suggests strategy: go “closer” to target if possible



Liben-Nowell et al. 2005 10.1073/pnas.0503018102

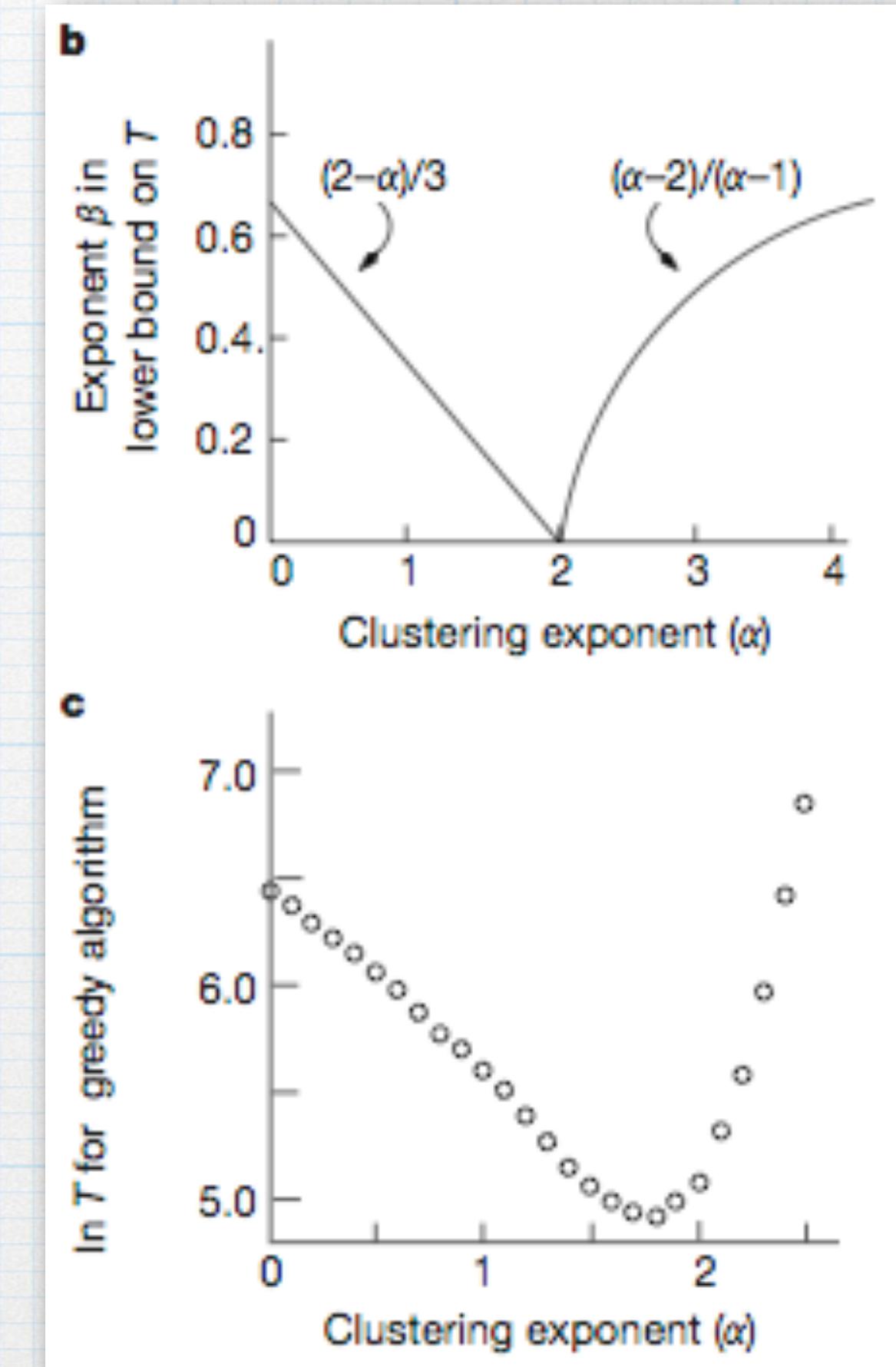
# Locality and geography

- \* Jon Kleinberg (2000): Assume we have information about position
- \* Local links to all lattice neighbors
- \* Long-range link probability distribution:  
power law  $\Pr(r) \sim r^{-\alpha}$
- \*  $r$ : lattice (Manhattan) distance
- \*  $\alpha$ : clustering exponent



# Locality and geography

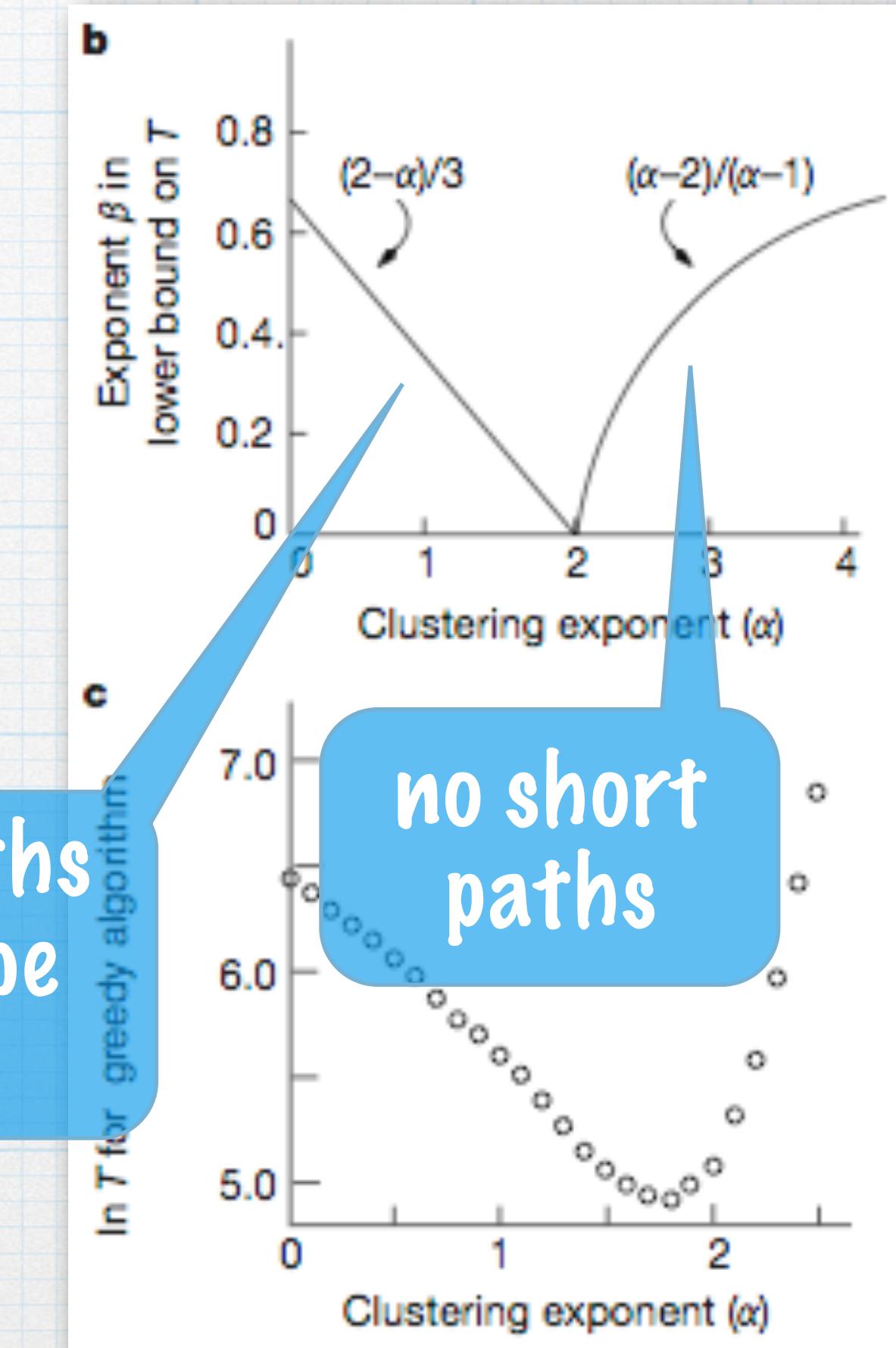
- \* Jon Kleinberg (2000): Assume we have information about position/location/distance
- \* Greedy algorithms: pick neighbor node closest to target
  - \*  $L \sim \log^2 N$  if  $\alpha=2$  (critical exponent = dimension of lattice)
  - \* Else  $L \sim N^\beta$



# Locality and geography

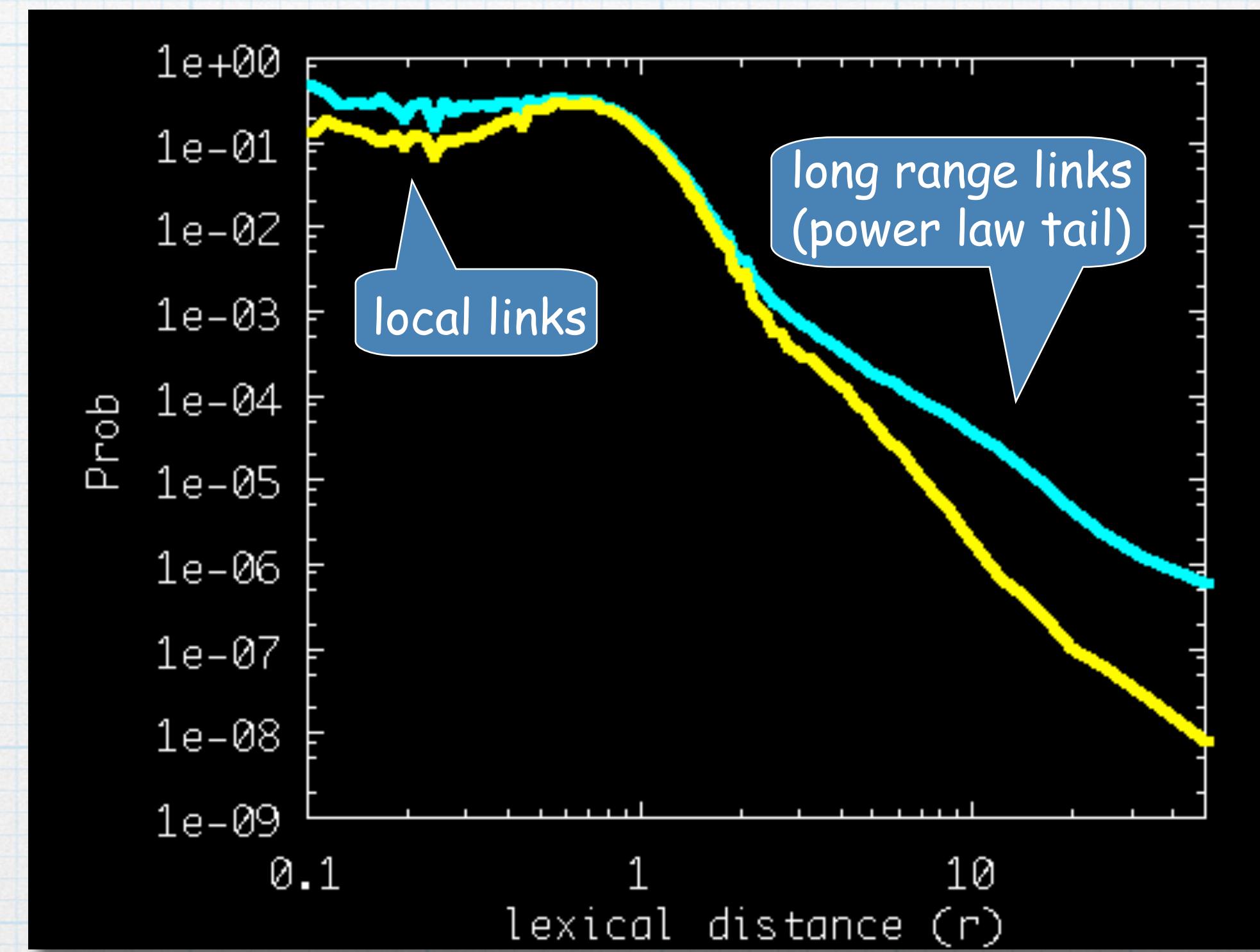
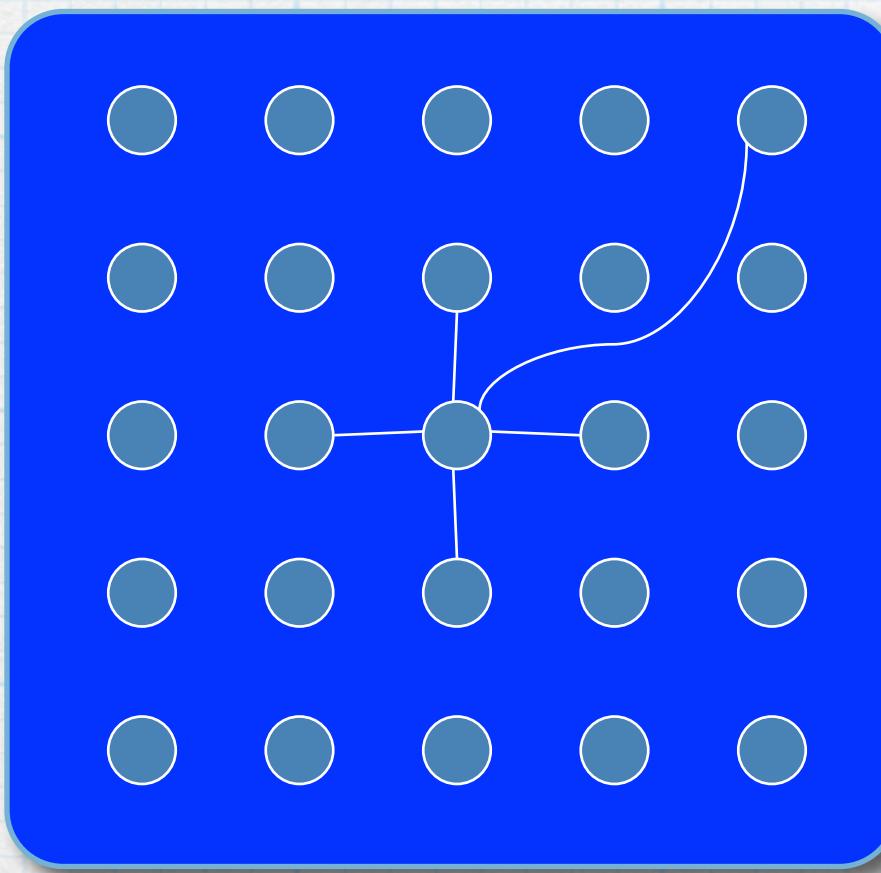
- \* Jon Kleinberg (2000): Assume we have information about position/location/distance
- \* Greedy algorithms: pick neighbor node closest to target
  - \*  $L \sim \log^2 N$  if  $\alpha=2$  (critical exponent = dimension of lattice)
  - \* Else  $L \sim N^\beta$

short paths  
cannot be  
found



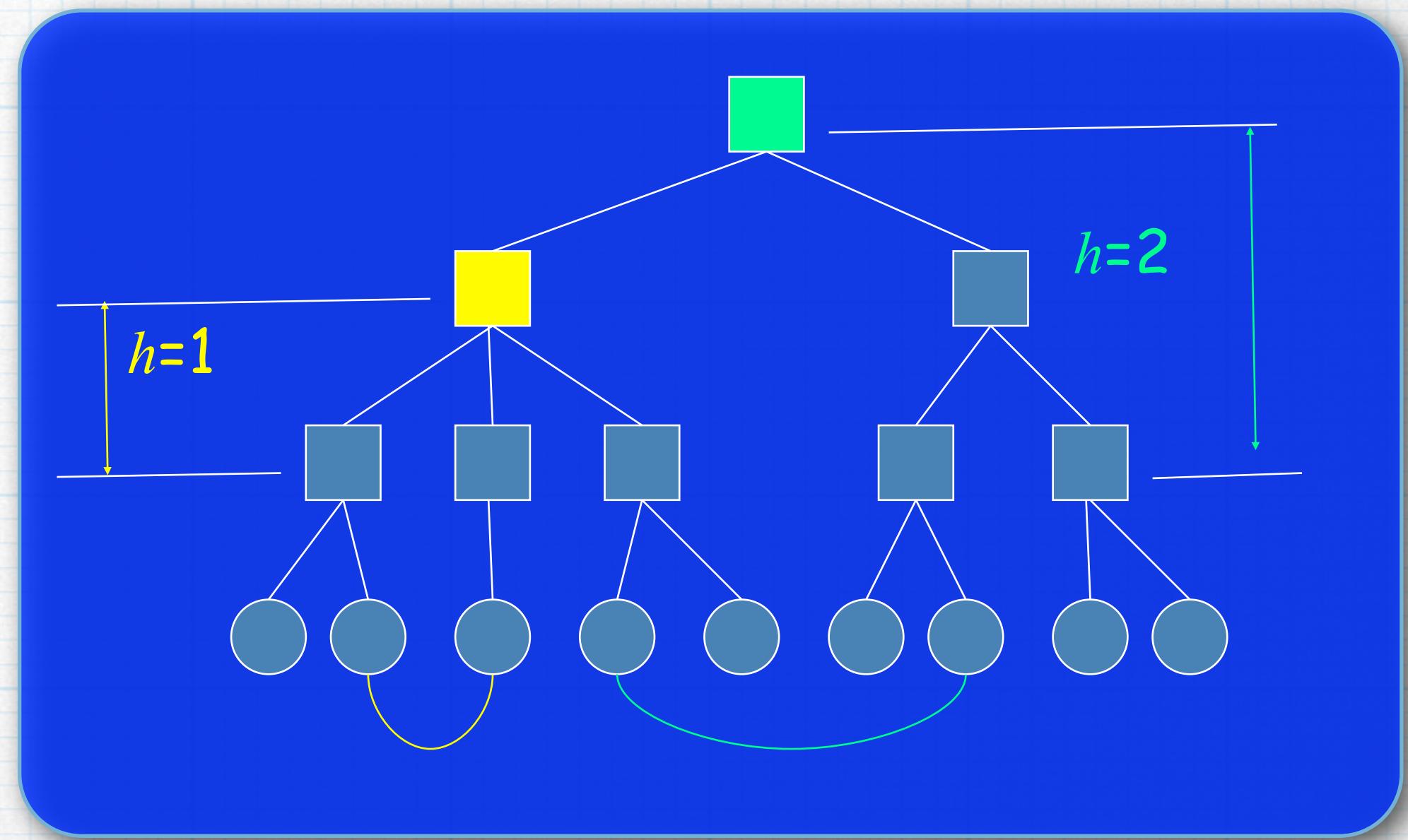
# Topical Locality & Searchability

- \* It turns out that the Web is a special case of “searchable” (local) network if we use “lexical” in place of geographic distance



# 2. Hierarchy

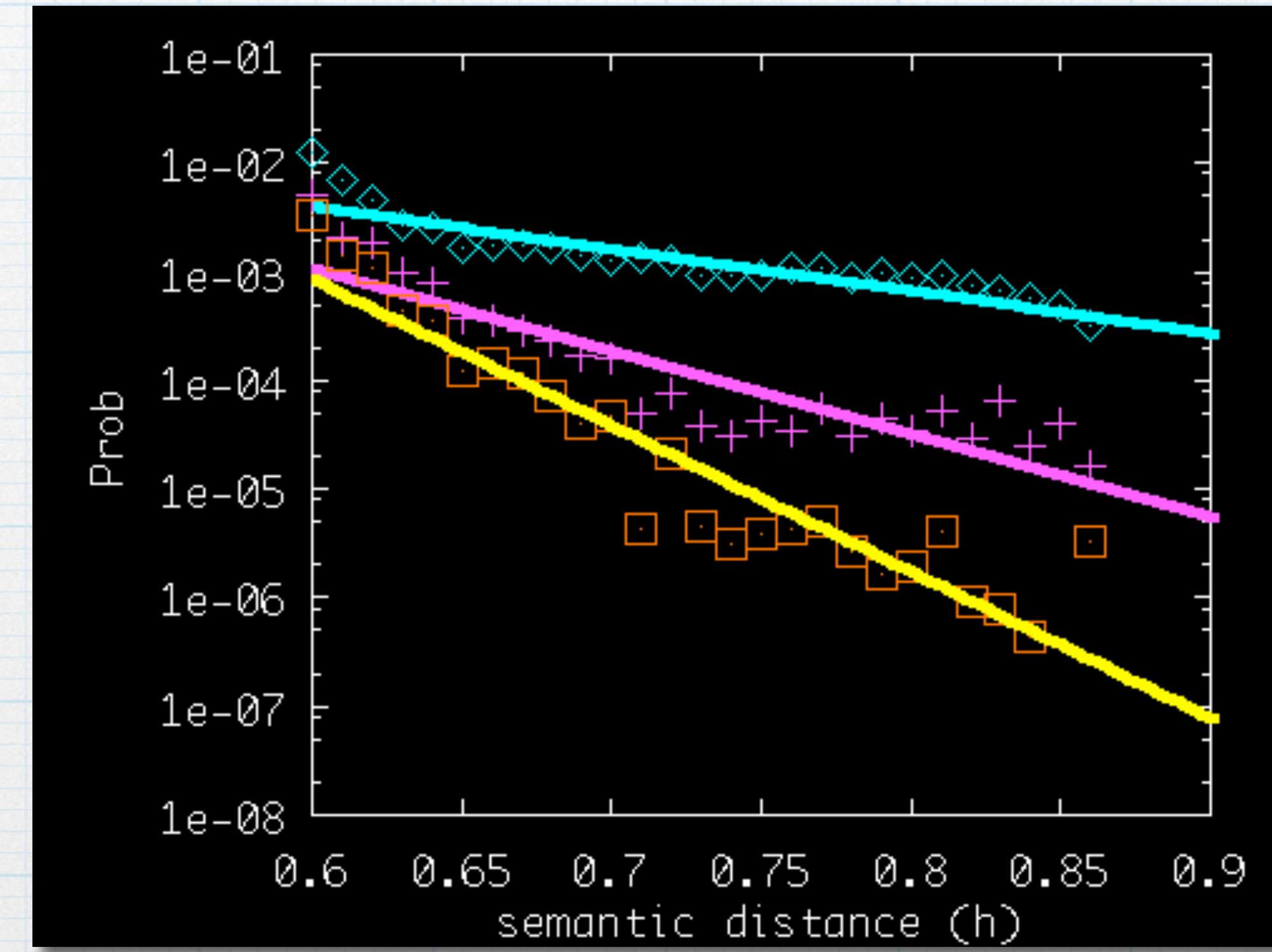
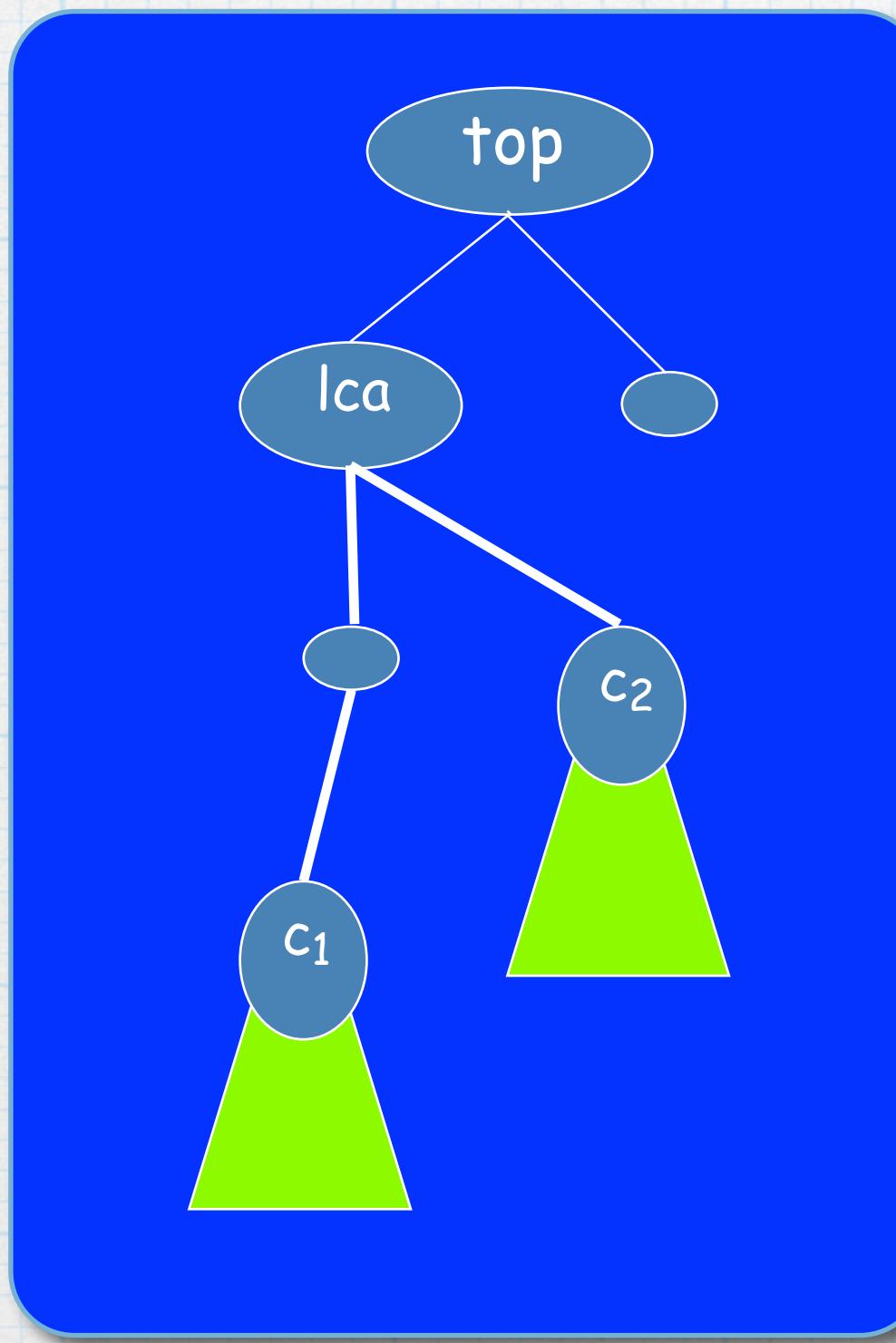
- \* Kleinberg (2002),  
Watts & al. (2002)
- \* Nodes are classified at the leaves of tree
- \* Eg, by topical interests, profession, etc.
- \* Link probability distribution:  
exponential  $\Pr(h) \sim e^{-h}$
- \*  $h$ : tree distance  
(height of lowest common ancestor)



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

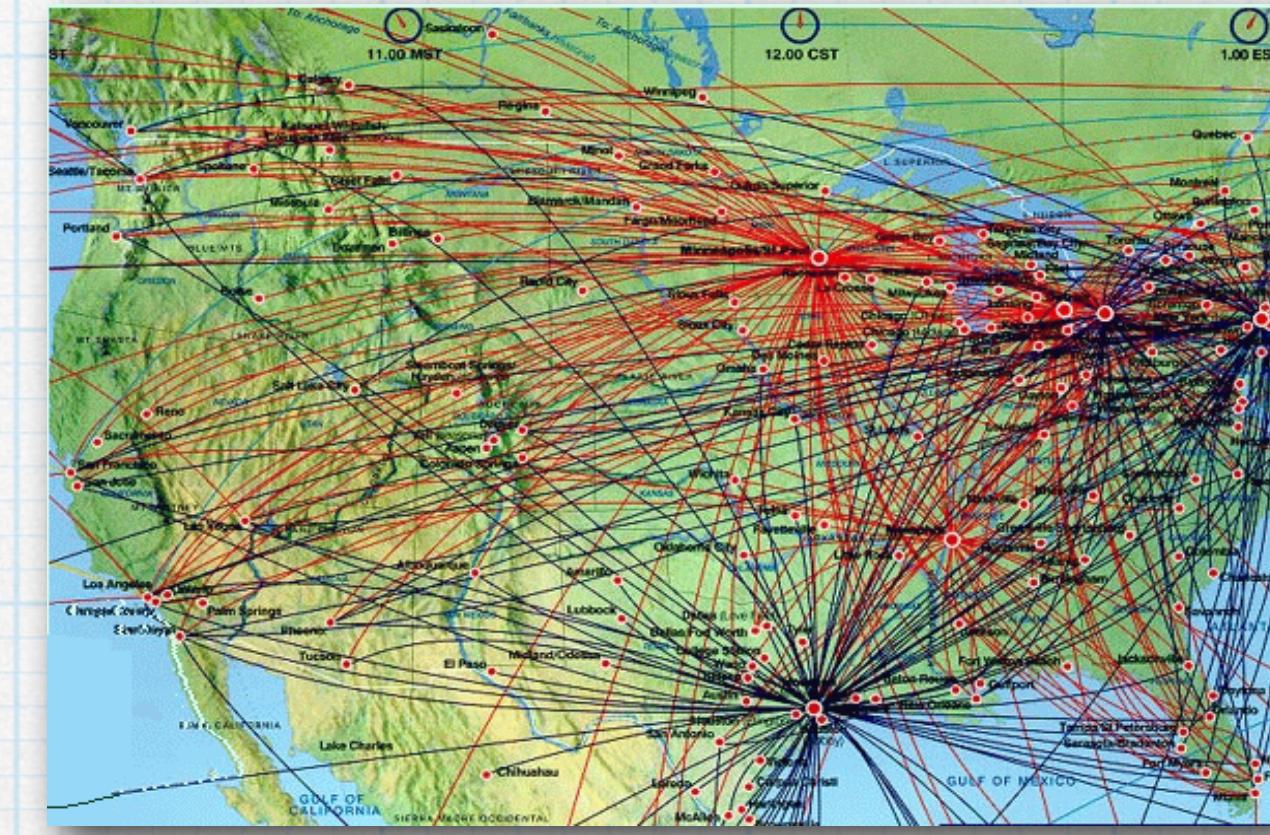
# Topical Hierarchy & Searchability

\* It turns out that the Web is a special case of “searchable” (hierarchical) network



# 3. Hubs

- \* Lada Adamic and Bernardo Huberman (2001): move to the neighbor with highest degree (hub)
- \* Air route approach
- \* Quickly “cover” entire P2P network if we have huge hubs
- \* Not good to search the Web at query time: too many links to explore! But crawlers implicitly leverage hubs, then information reused many times by search engines



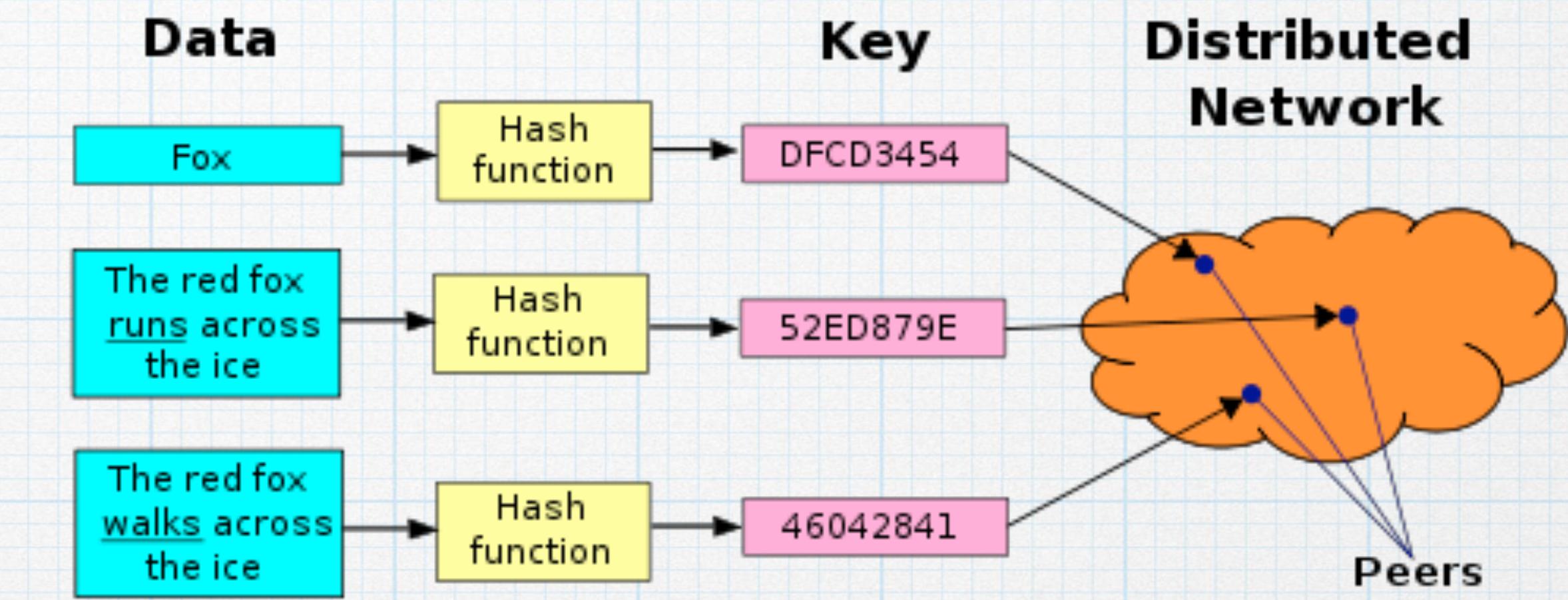
# Search in peer networks

- \* Gnutella (unstructured network): flood
  - \* When you need a file, ask all your neighbors
  - \* Each neighbor does the same...
  - \* Too many messages!
  - \* Peers busy forwarding messages most of the time...



# Search in peer networks

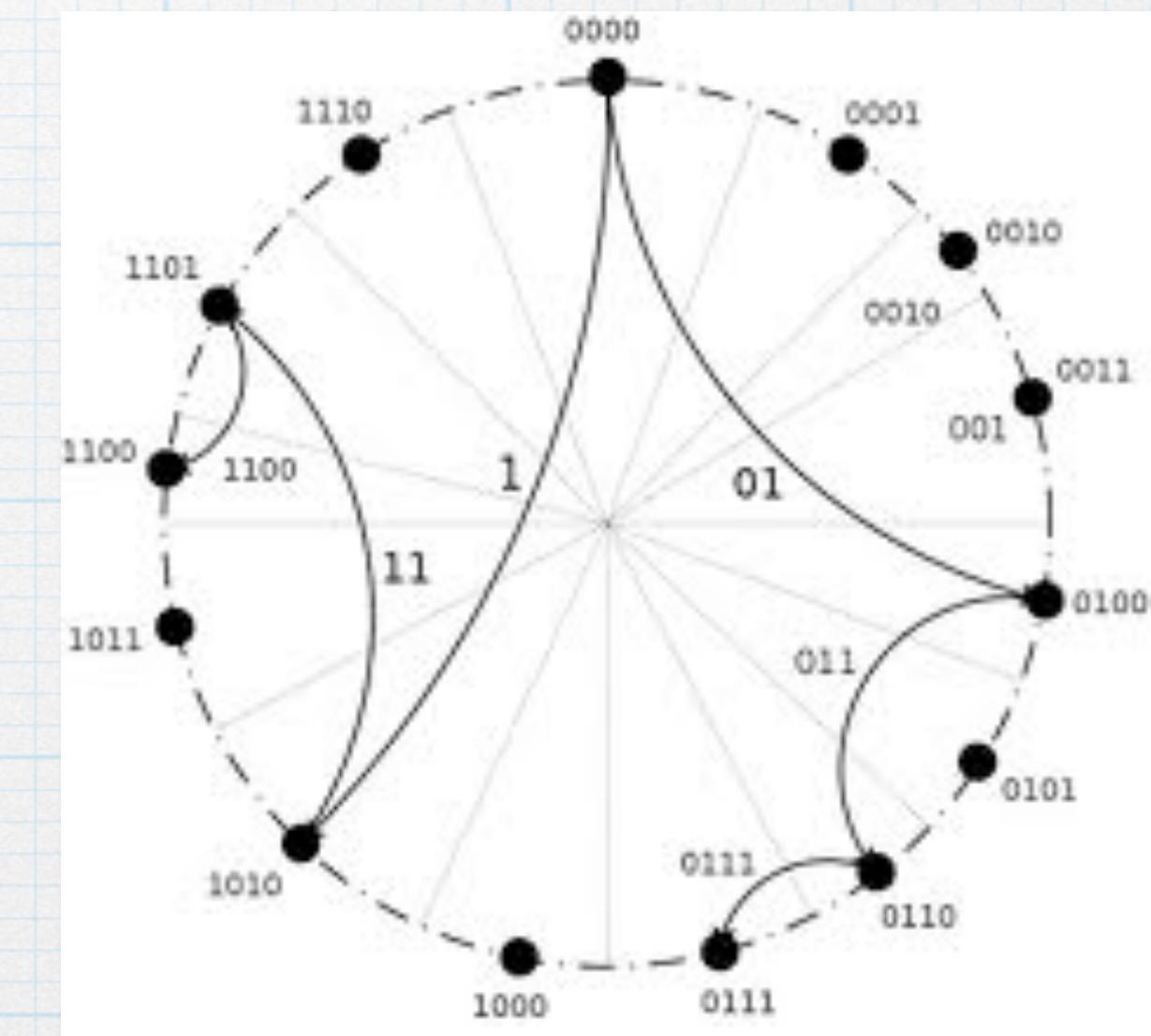
- \* Structured networks, distributed hash tables (DHT)



- \* Overlay network: Each node maintains a set of links to its neighbors (routing table). A node picks its neighbors according to a network's topology (KAD, CHORD, BitTorrent...)

# Search in peer networks

- \* DHT topological property: for any key  $k$ , each node either has a node ID that owns  $k$  or has a link to a node whose node ID is closer to  $k$
- \* Greedy routing algorithm:
  - \* At each step, forward the message to the neighbor whose ID is closest to  $k$ . If no such neighbor, done!
- \* Tradeoff: the more neighbors per node, the shorter the paths



# The End