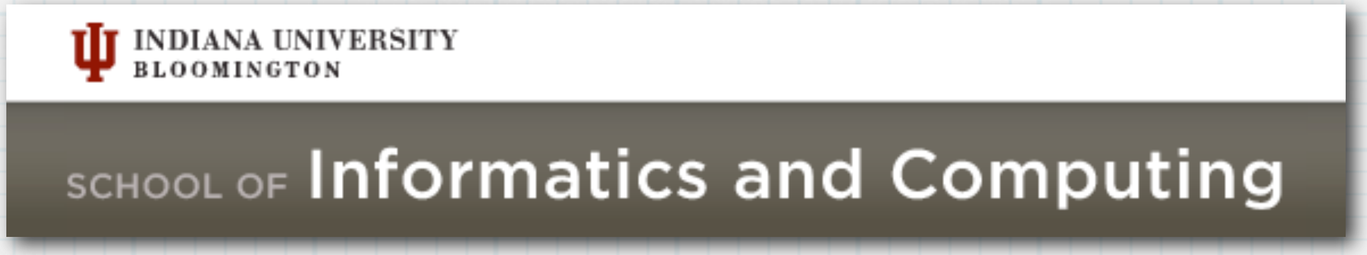
INFO 368
Introduction to network science

Filippo Menczer



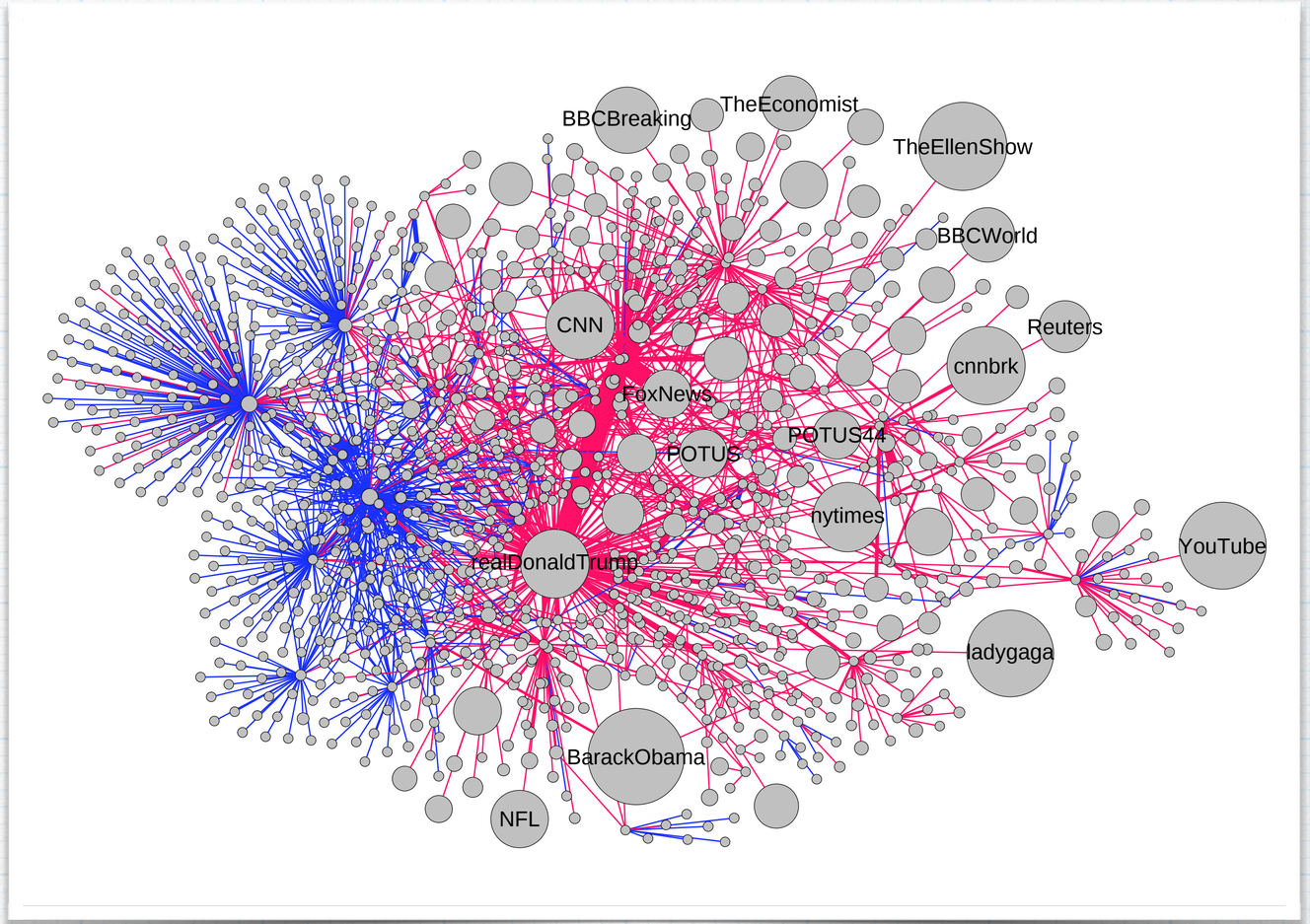
- Review
  - SIR vs threshold models of diffusion
  - Network role in spreading processes
  - Halting an epidemic
  - Selective immunization

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

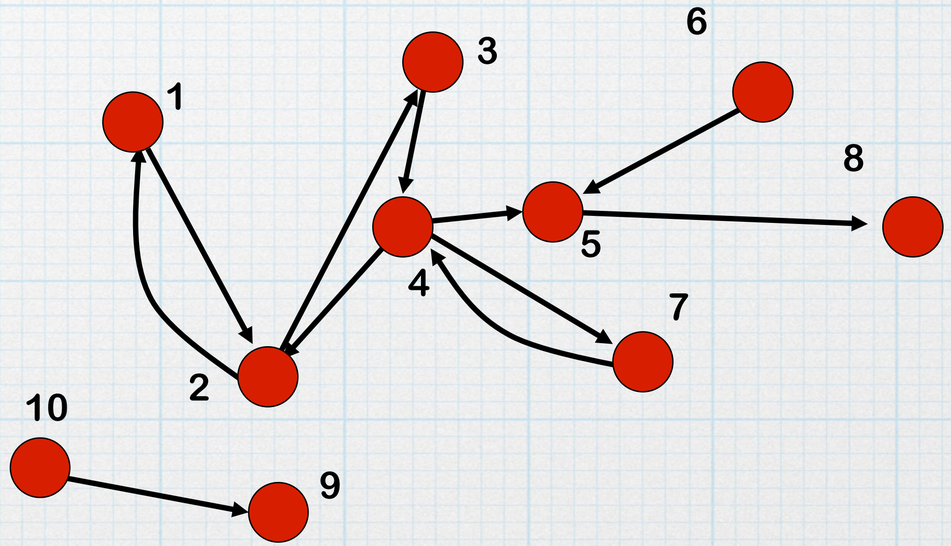
* Each link has a weight
* Examples?
* How do we extend the notion of degree to the weighted case?



* 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

directed_matrix - No...

File Edit Format View Help

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

Untitled - Notepad

File Edit Format View Help

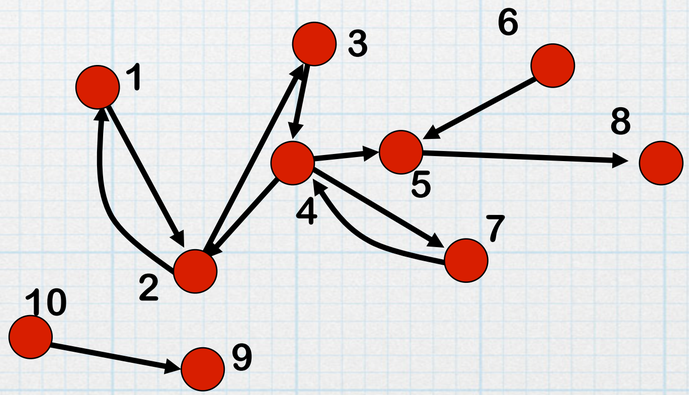
```
1 -> 2
2 -> 1 3
3 -> 4
4 -> 2 5 7
5 -> 8
6 -> 5
7 -> 4
10-> 9
```

binary_dir...

File Edit Format View Help

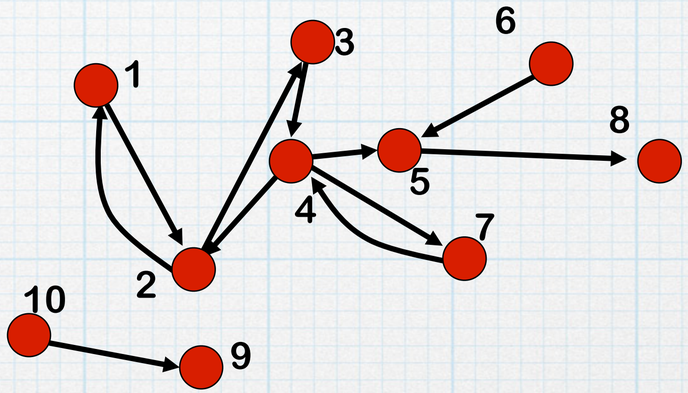
```
1  2
2  1
2  3
3  4
4  2
4  5
4  7
5  8
6  5
7  4
10 9
```

* Representations:
  * Adjacency matrix
  * Adjacency list
  * Edge list



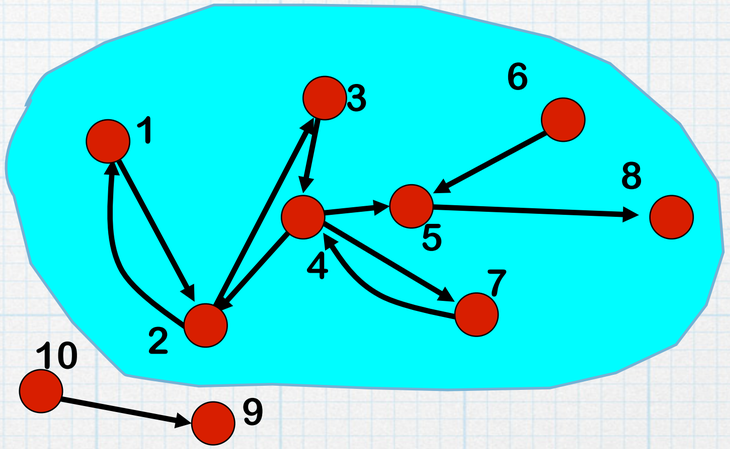
Directed networks

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



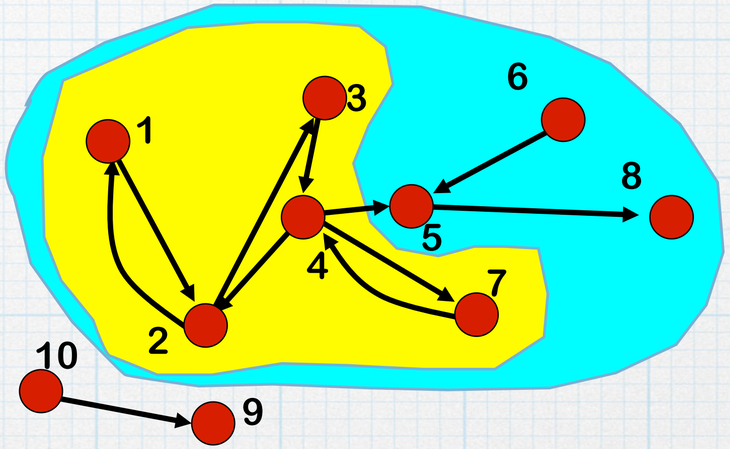
Directed networks

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



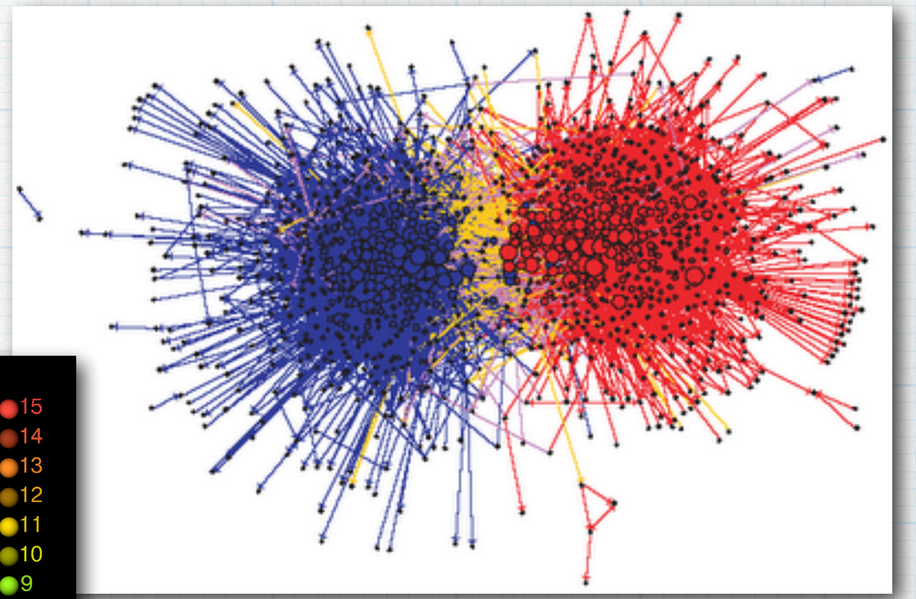
Directed networks

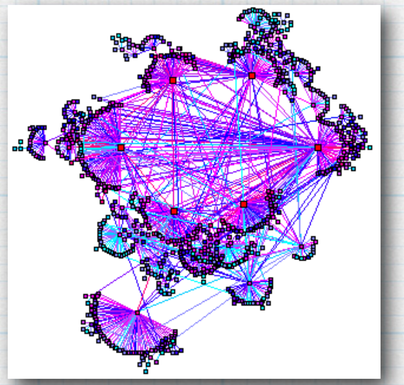
* Paths are directed
* There may be a path from A to B but not from B to A
* Strongly and weakly connected components

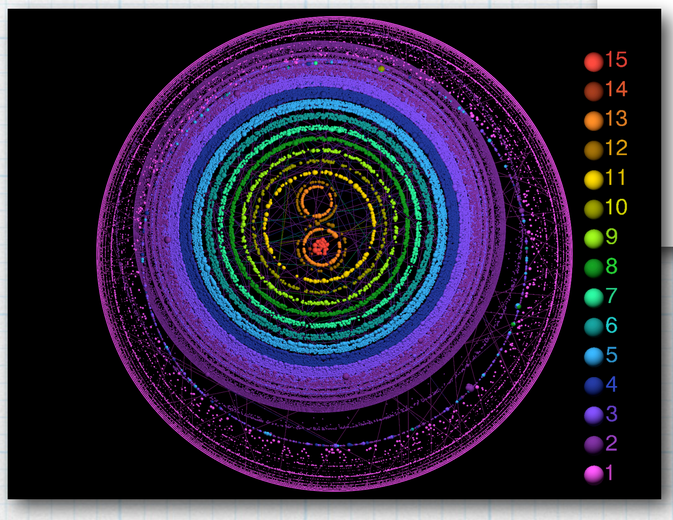


Information networks: Web, Wikipedia, blogs...

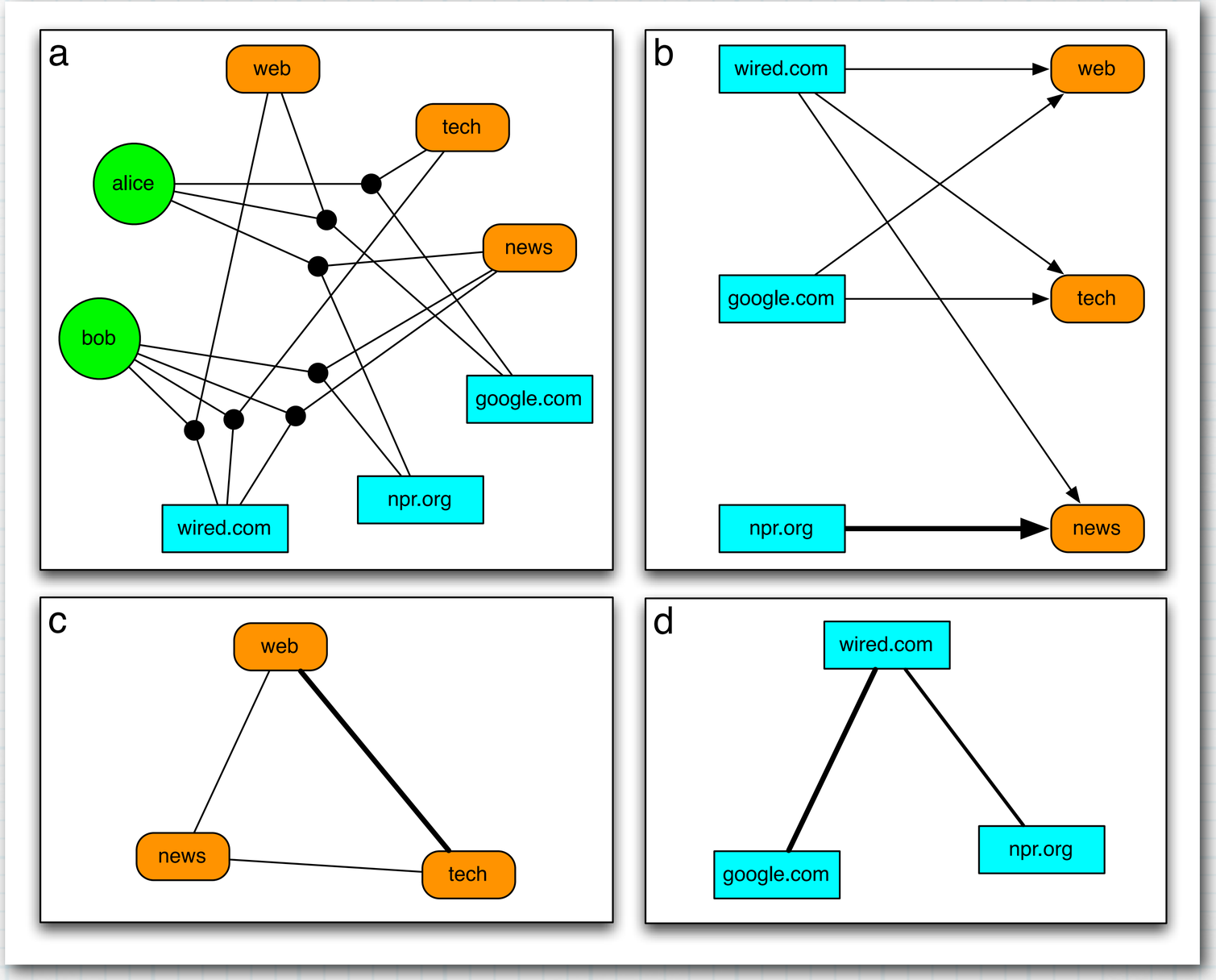
* What are the nodes?
* What are the edges?
* How do they form?



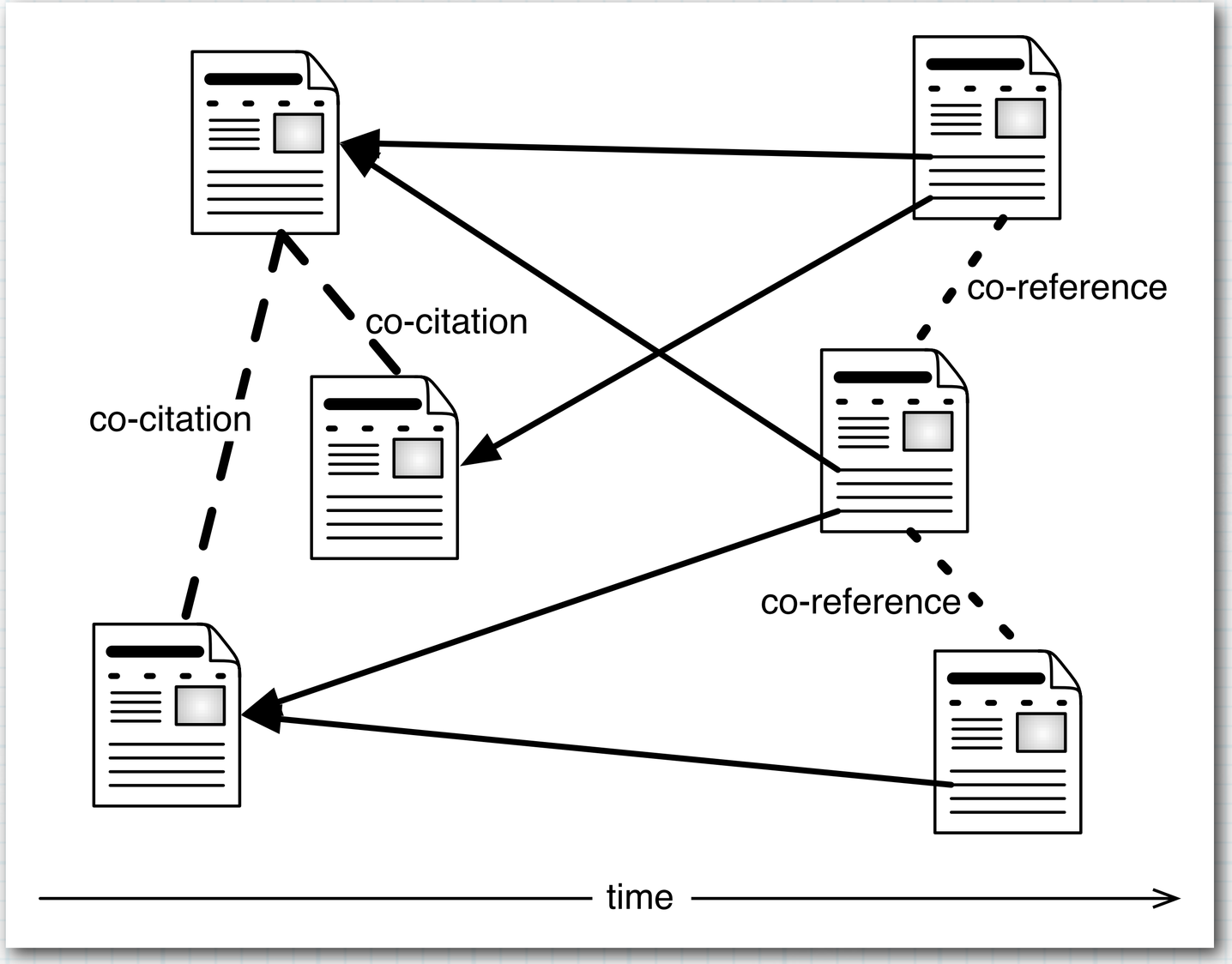




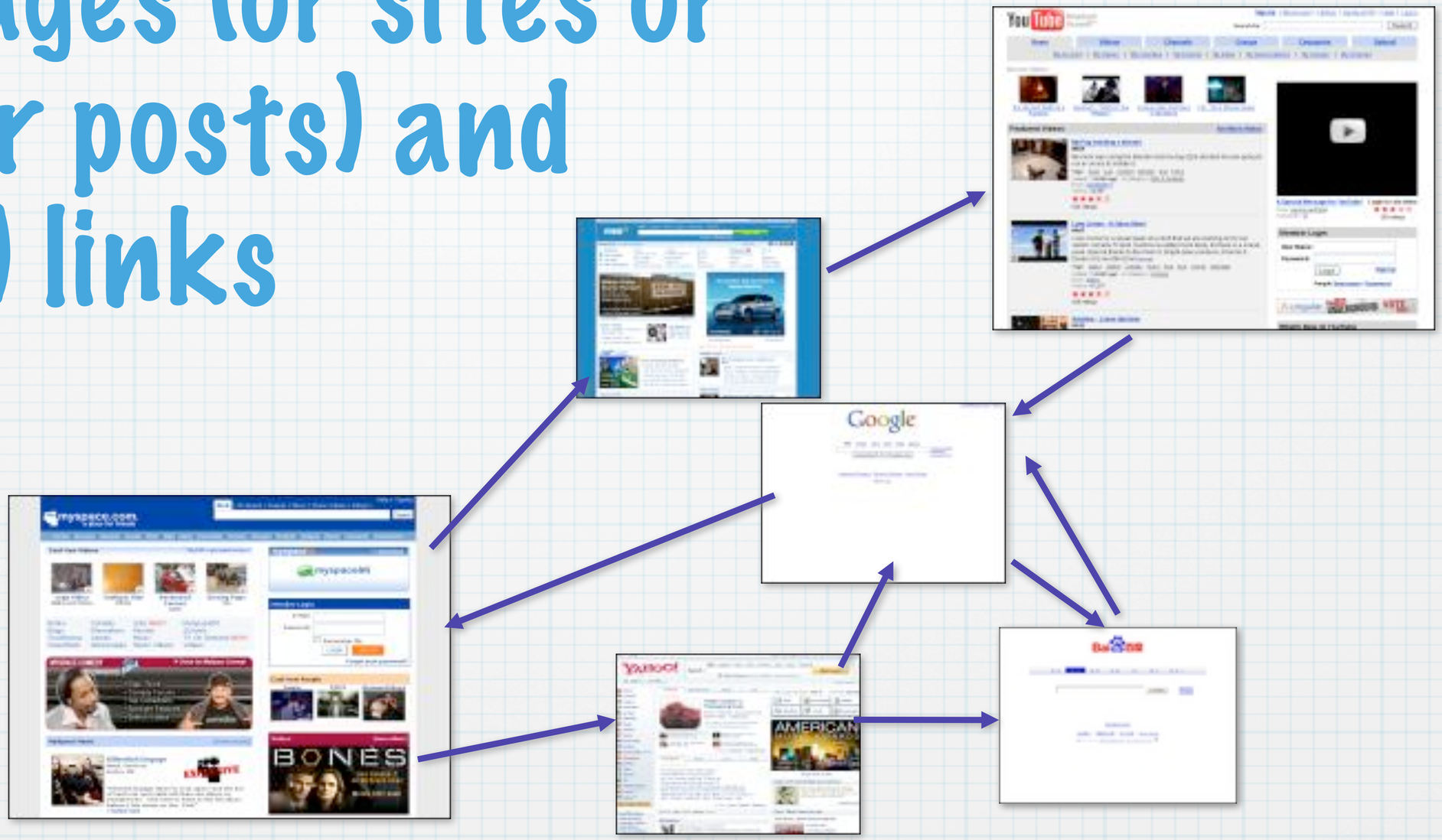
Folksonomies & tag networks



Citations



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

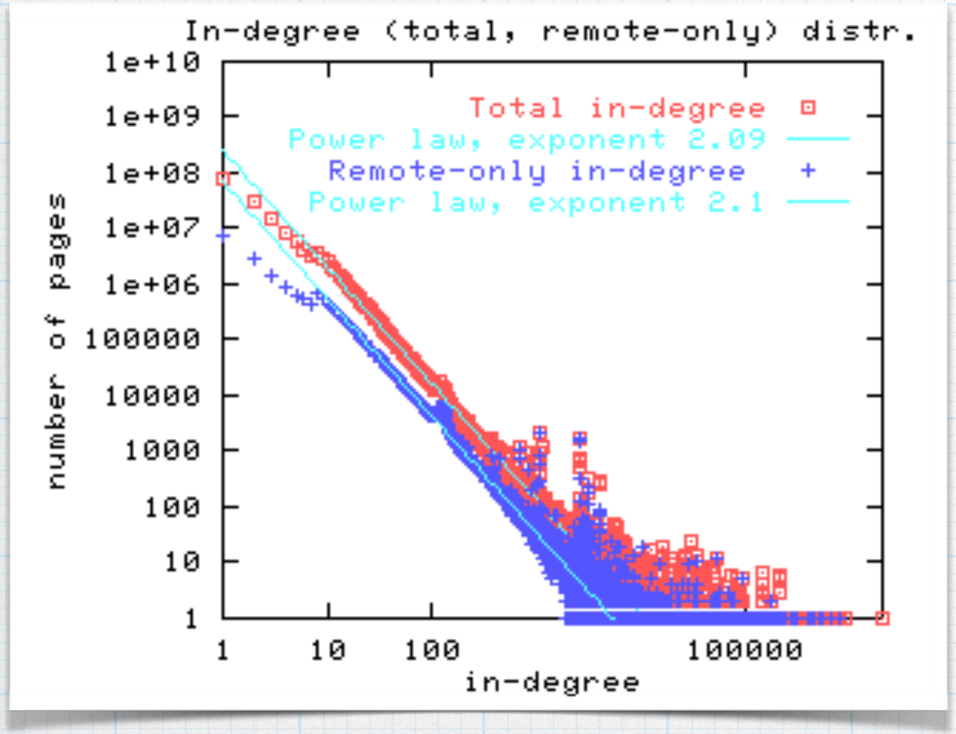


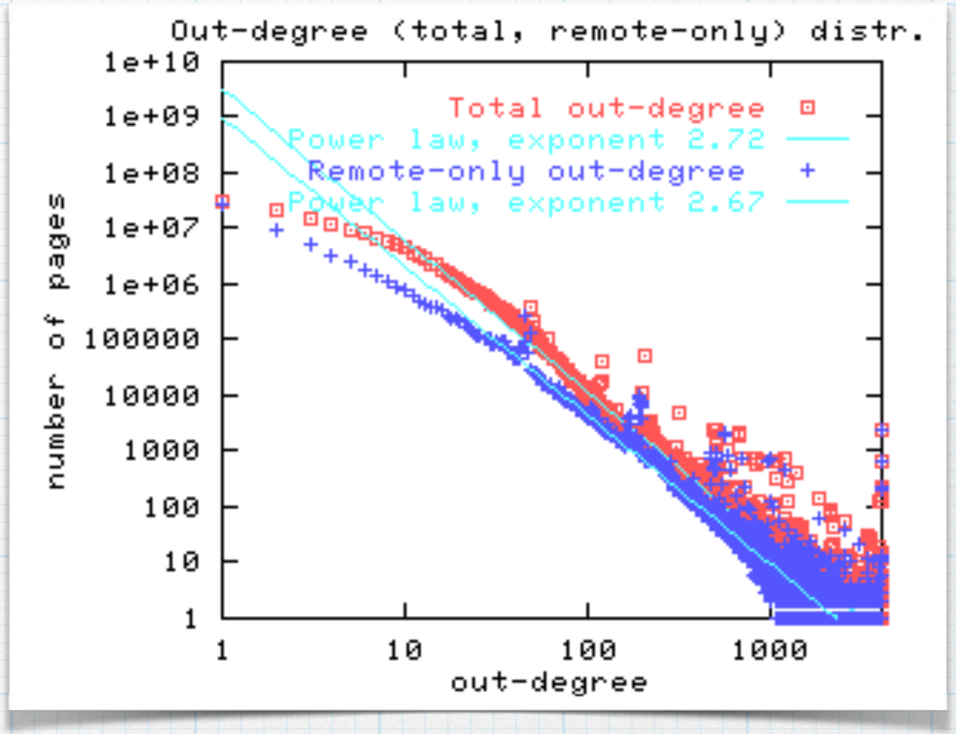
INDIANA UNIVERSITY BLOOMINGTON

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? 20billion?
* Question no longer makes sense
  * Search engines stopped reporting # pages indexed
  * Dynamic pages
  * Hidden Web

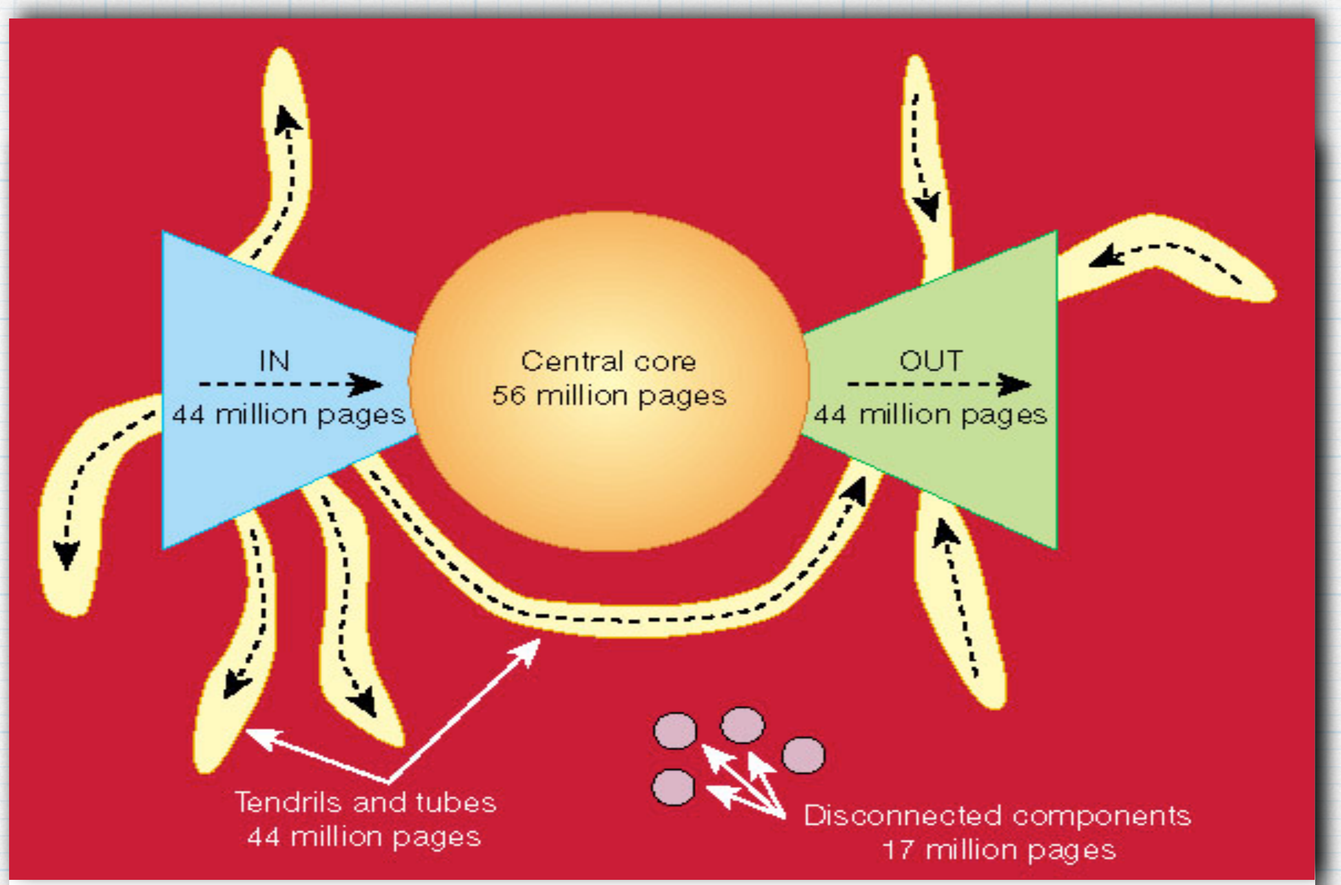
Scale free Web structure





* 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?

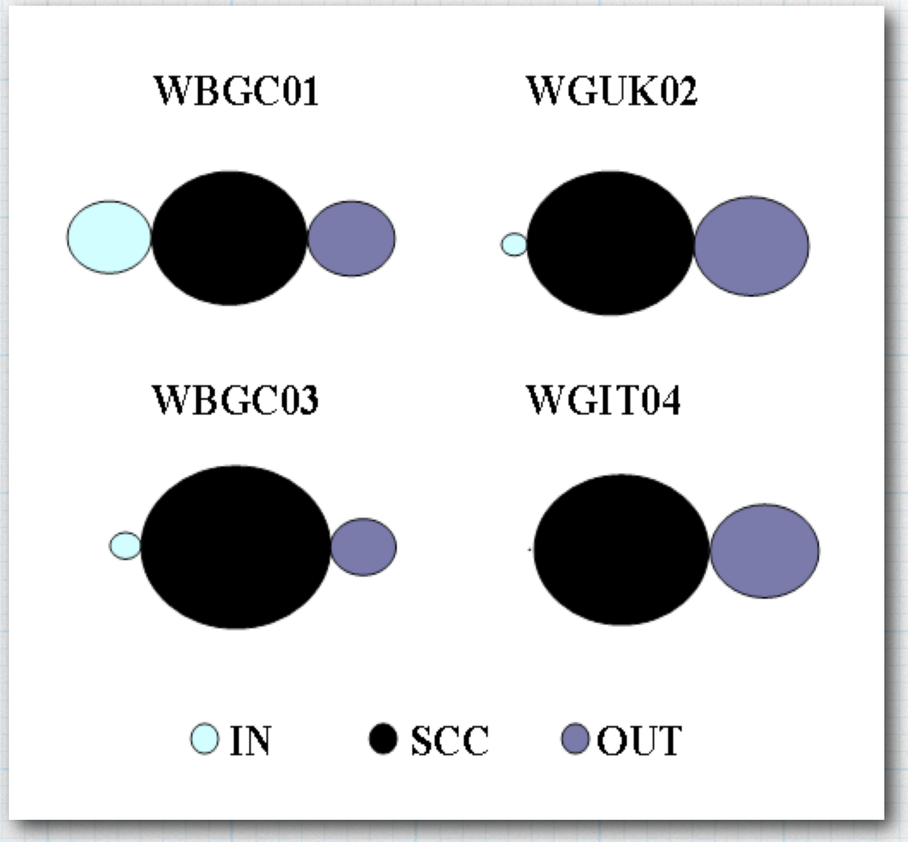
The Web looks like a "bow-tie" (why?)



Region:	SCC	IN	OUT	TENDRILS	Disconnected	TOTAL
Size:	56,463,993	43,343,168	43,166,185	43,797,944	16,777,756	203,549,046

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

Web structure from different crawls



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

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

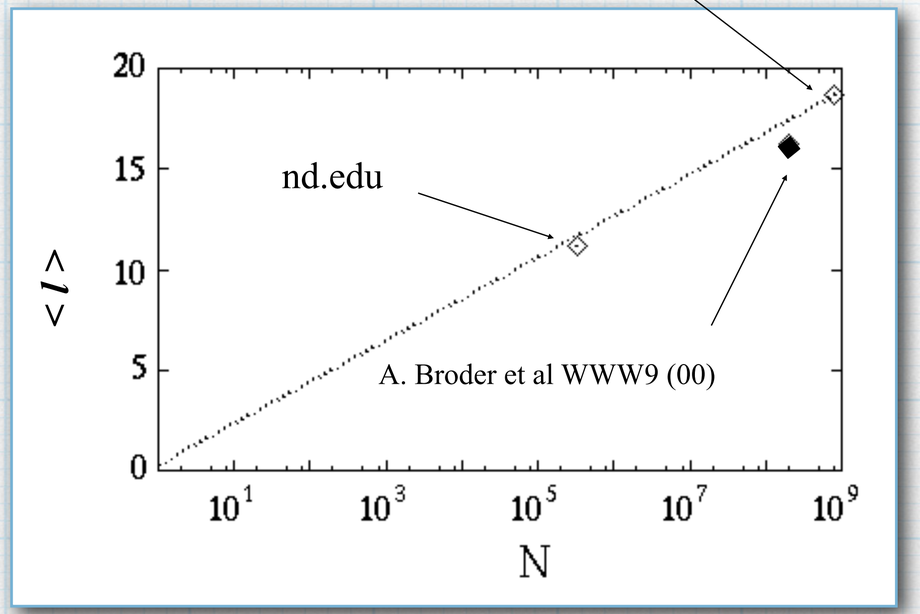
Data set	WBGC01	WGUK02	WBGC03	WGIT04
$\langle k_{in} \rangle$	9.3	15.8	24.1	27.5
k_{in}^{max}	788632	194942	378875	1326744
σ_{in}	200.2	143.3	421.6	881.4
κ_{in}	4298.6	1317.5	7414.9	28269.9
γ_{in}	1.9	1.7	2.2	1.6
	WBGC01	WGUK02	WBGC03	WGIT04
$\langle k_{out} \rangle$	9.3	15.8	24.1	27.5
k_{out}^{max}	552	2449	629	9964
σ_{out}	13.1	27.4	29.5	67.1
κ_{out}	27.7	63.4	60.3	191.0
γ_{out}	∞	∞	∞	∞

Serrano & al. 2007

Web average path length

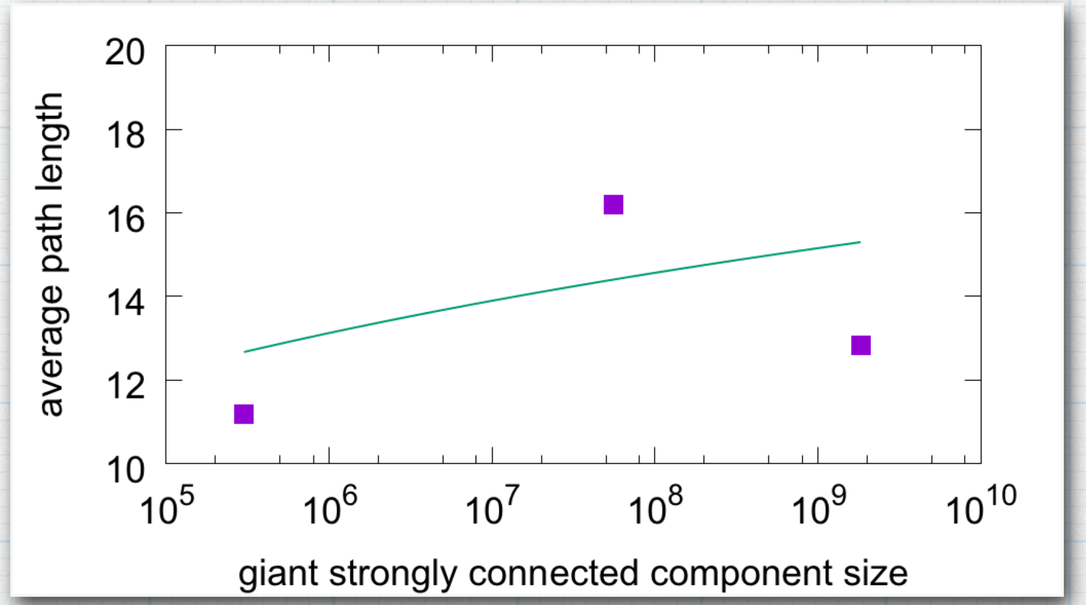
For the core (SCC)

19 degrees of separation
Albert et al 1999



$< l > \sim log(N)$

2012 commoncrawl.org
Meusel et al 2015



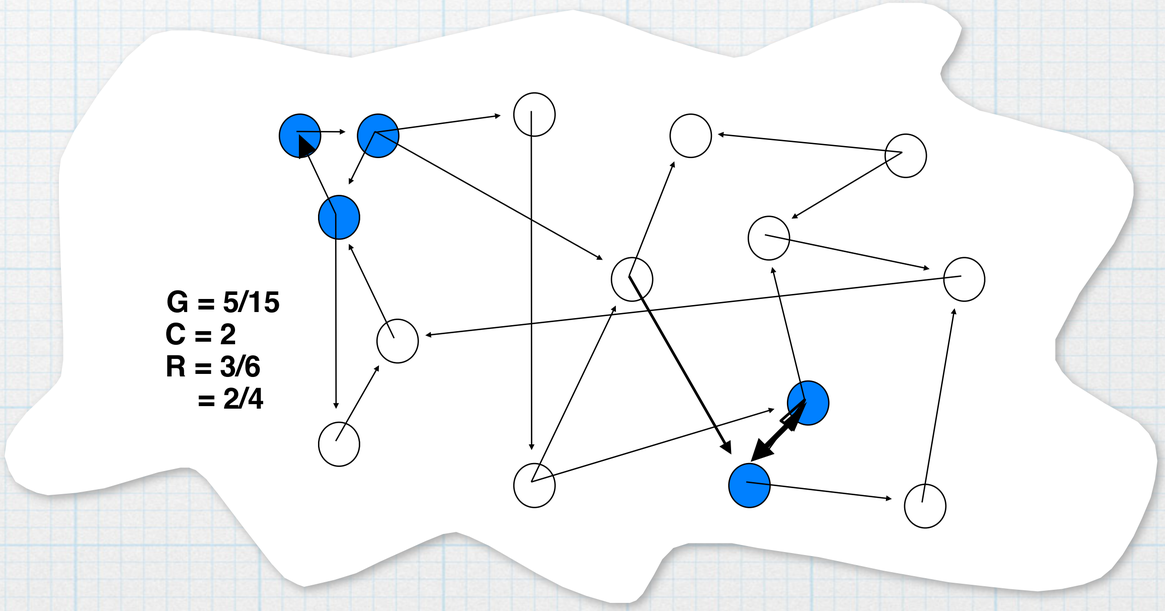
$< l > \sim log(log(N))$

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[rel(p)] : fraction of relevant/topical pages (topic generality)
  - R = Pr[rel(p) | rel(q) AND 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

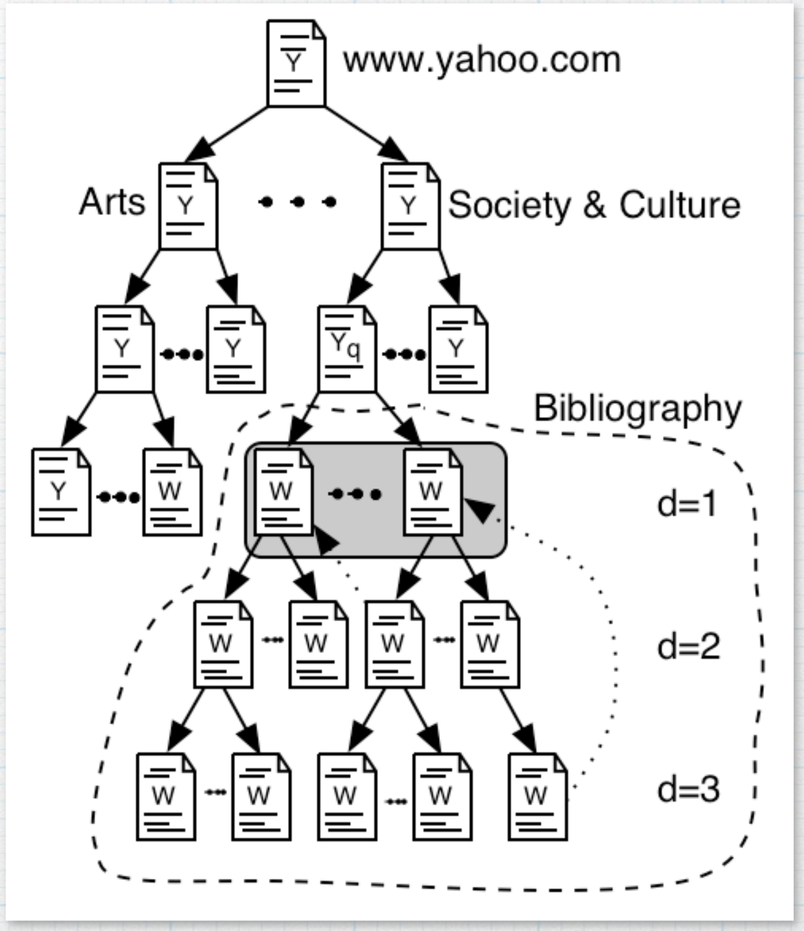


Topical locality

- 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)

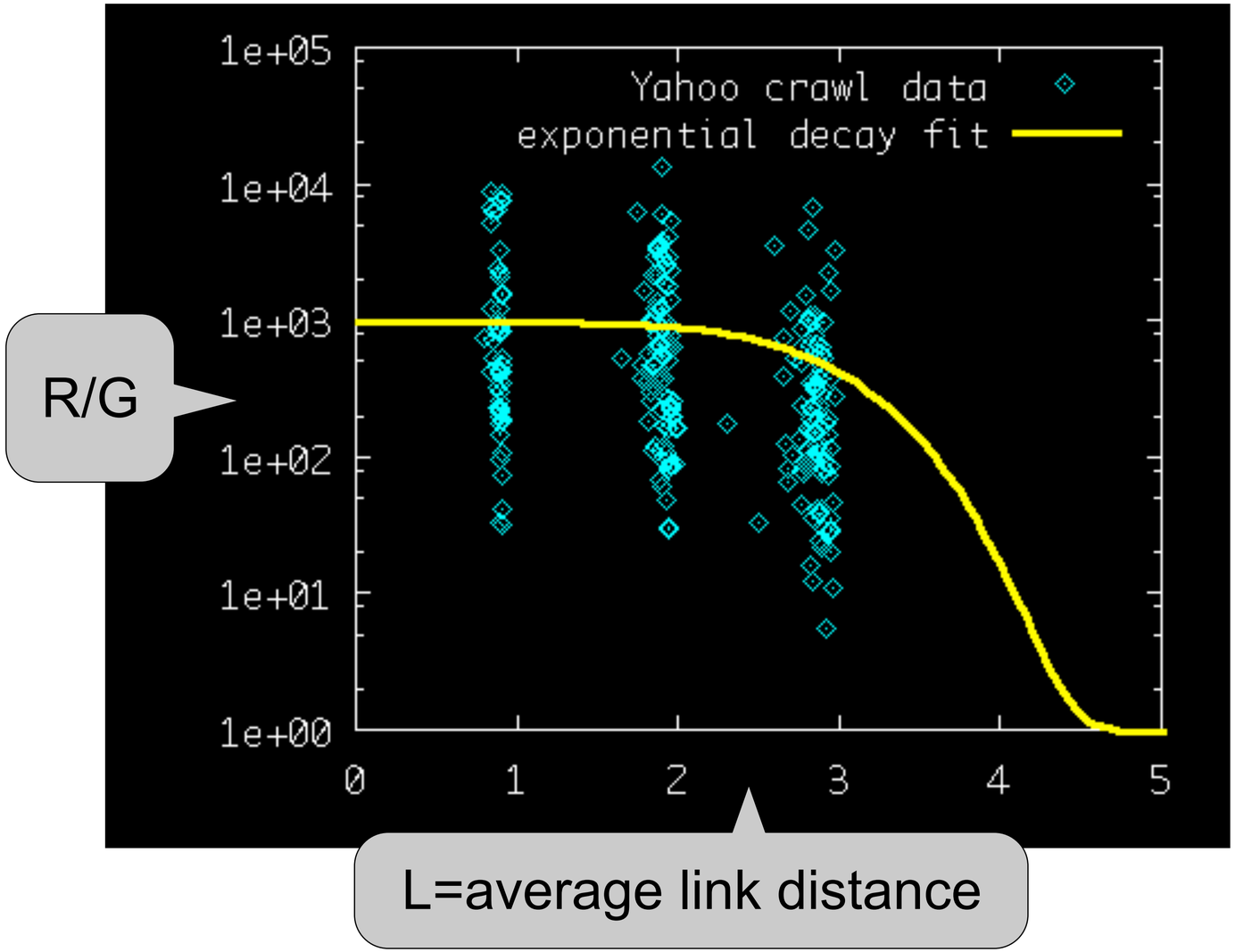
Quantifying topical locality

- 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



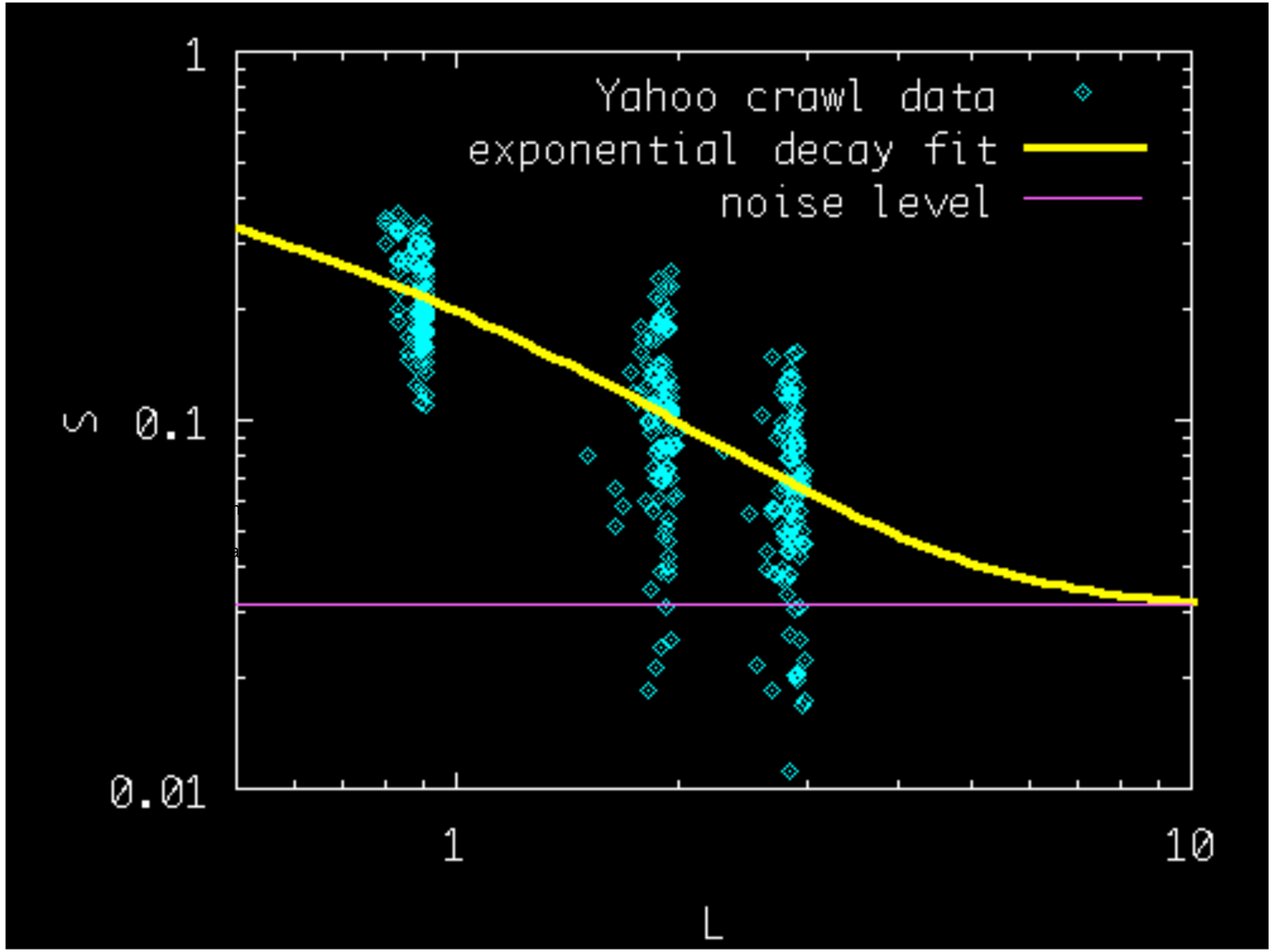
Semantic locality

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



Topic drift

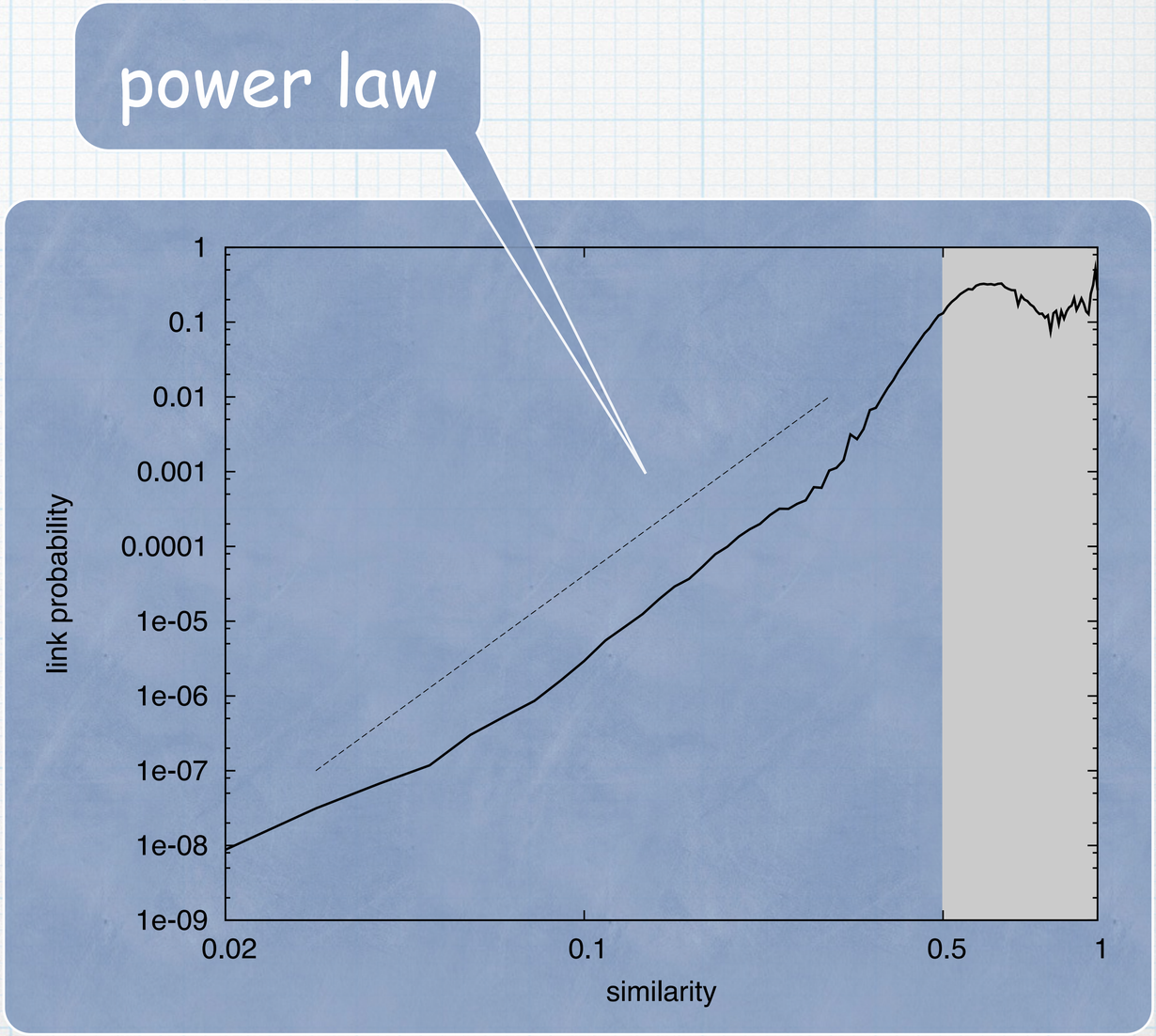
- Correlation of lexical (content) and link topology
- **L**: average link distance
- **S**: average content similarity to start (topic) page



Bing Liu: *Web Data Mining*. Springer, 2007
Ch. 8 *Web Crawling* by Filippo Menczer

Topical locality

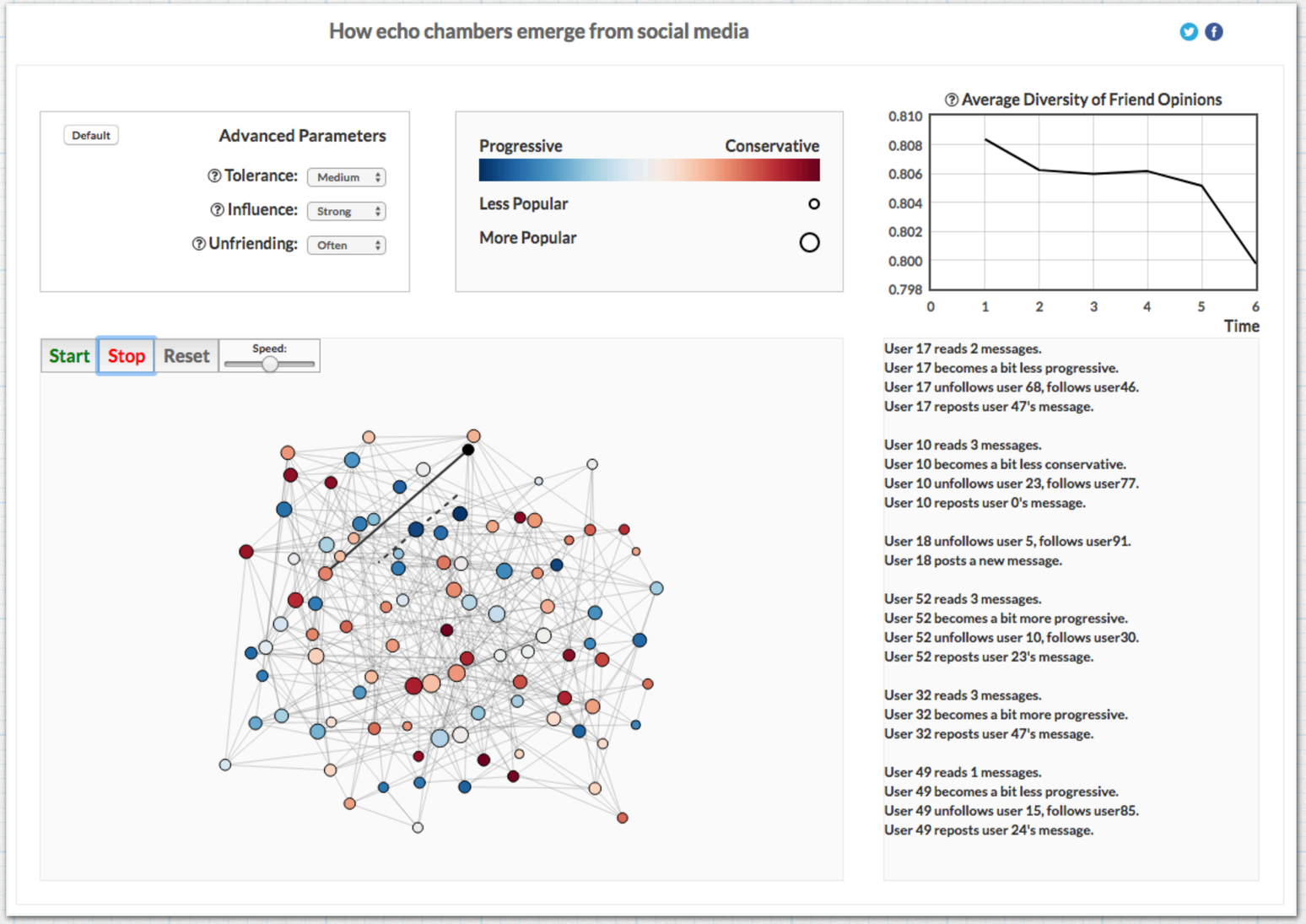
* What is the relationship between the similarity of two pages, and the probability they are linked?
* Are people more likely to create a link to a related page?



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/

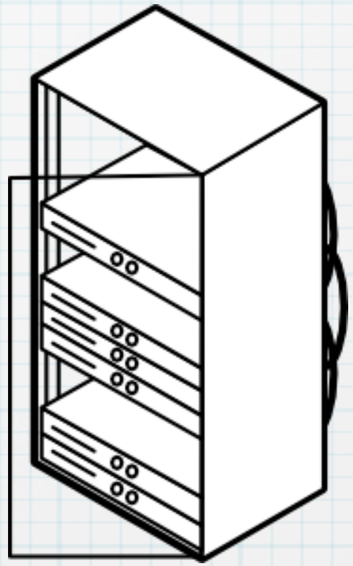
- Review
  - Web structure
    - Degree
    - Bow-tie
    - Paths
  - Homophily and topical locality

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?

Client

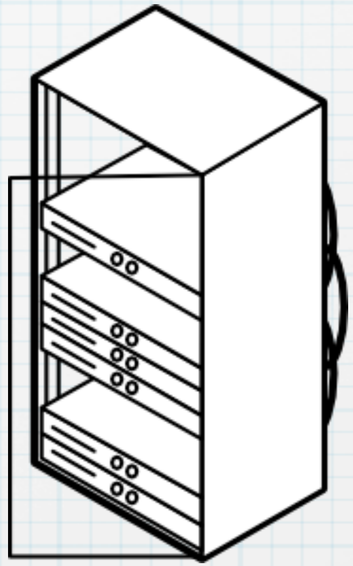


Server

How does the Web work?

Client

http://taylorswift.com/

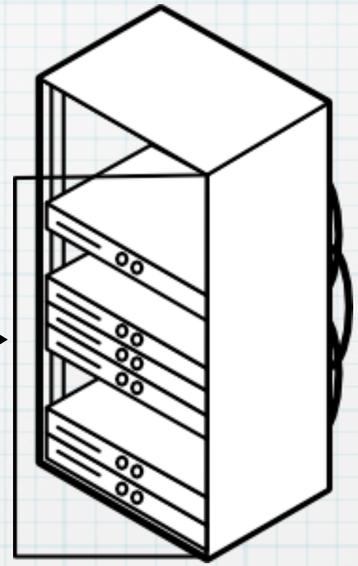


Server

How does the Web work?

Client

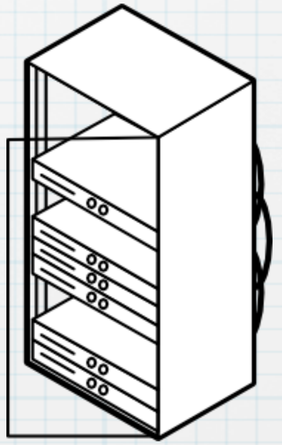
http://taylorswift.com/



Server

Let's try it the nerd way!

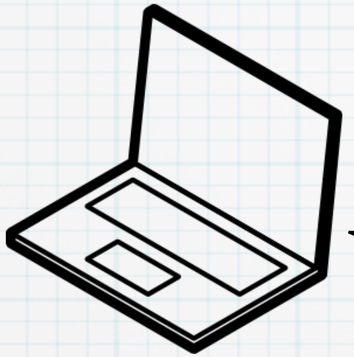
Client: "http://taylorswift.com/"



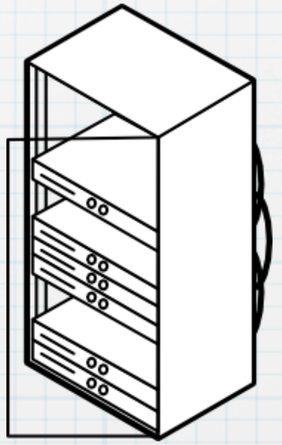
Server

Let's try it the nerd way!

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



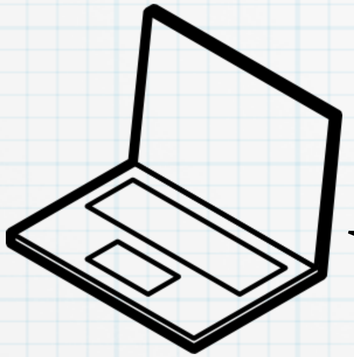
```
telnet taylorswift.com 80
Trying 2400:cb00:2048:1::6814:1a73...
Connected to taylorswift.com
Escape character is '^]'.
GET / HTTP/1.1
host: taylorswift.com
[return] (empty line means 'done')
```



Server

Let's try it the nerd way!

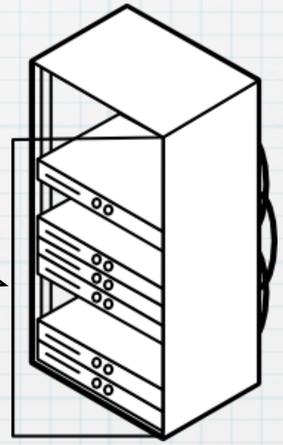
Client: "http://taylorswift.com/"



```
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

swift

Google

swift

Sign in

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


Objective-C


Python


C


Ruby


Java

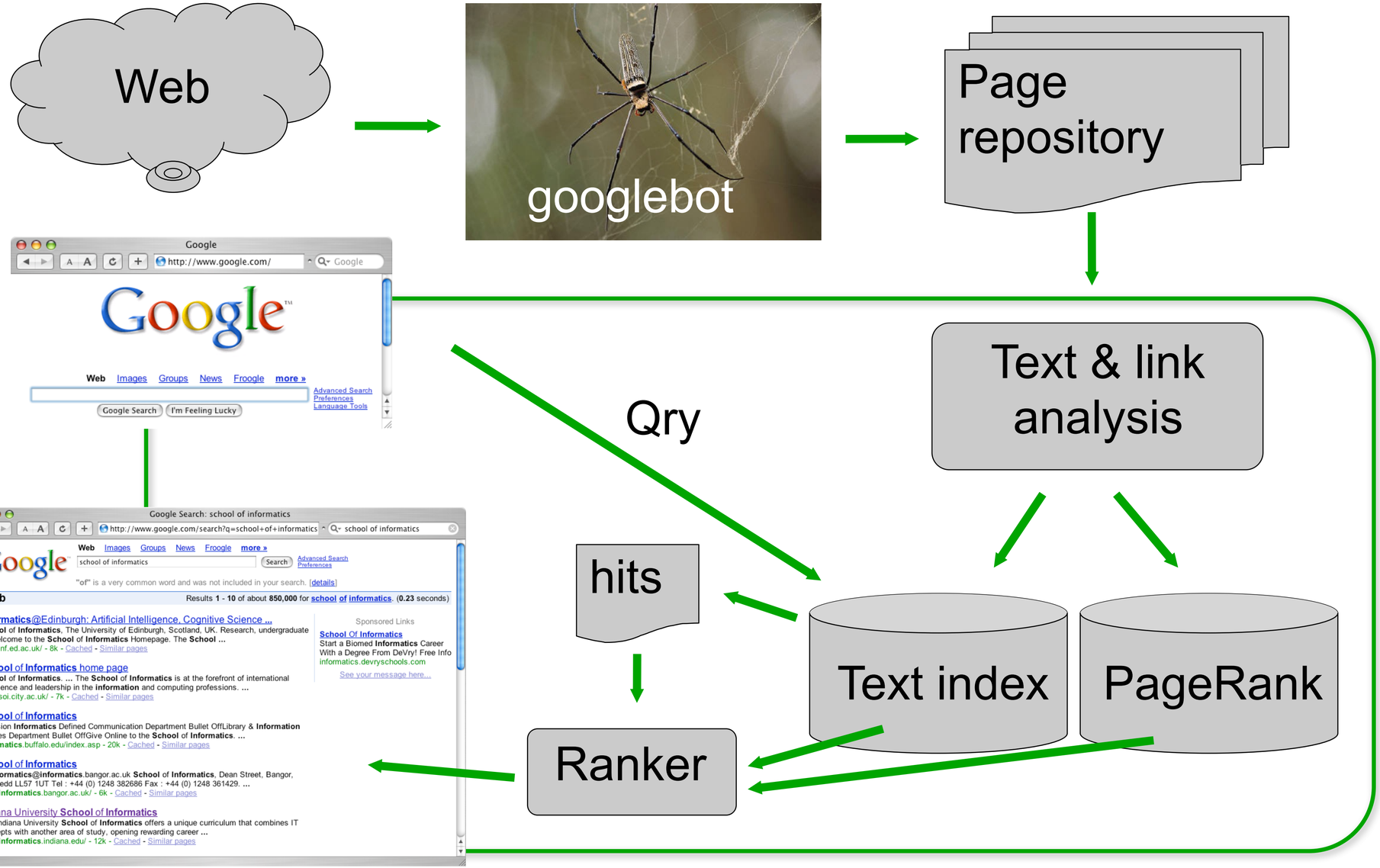
Feedback

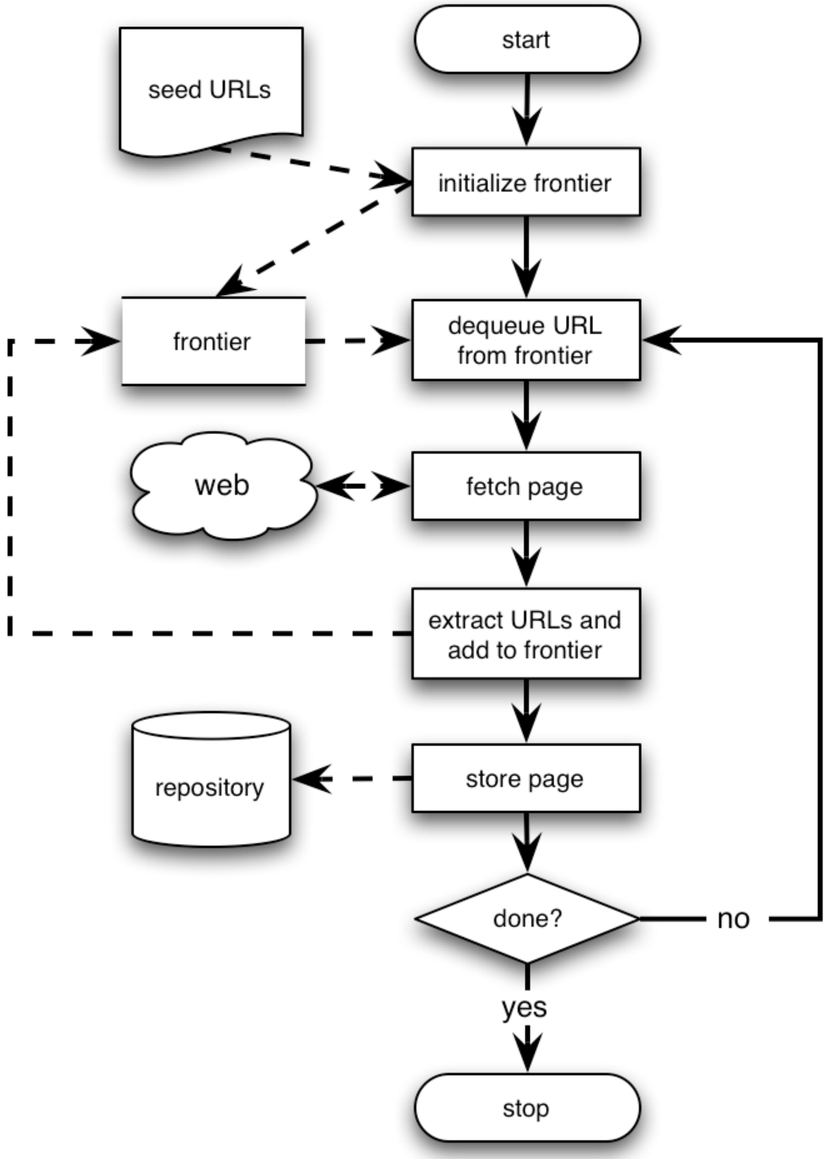
See results about

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



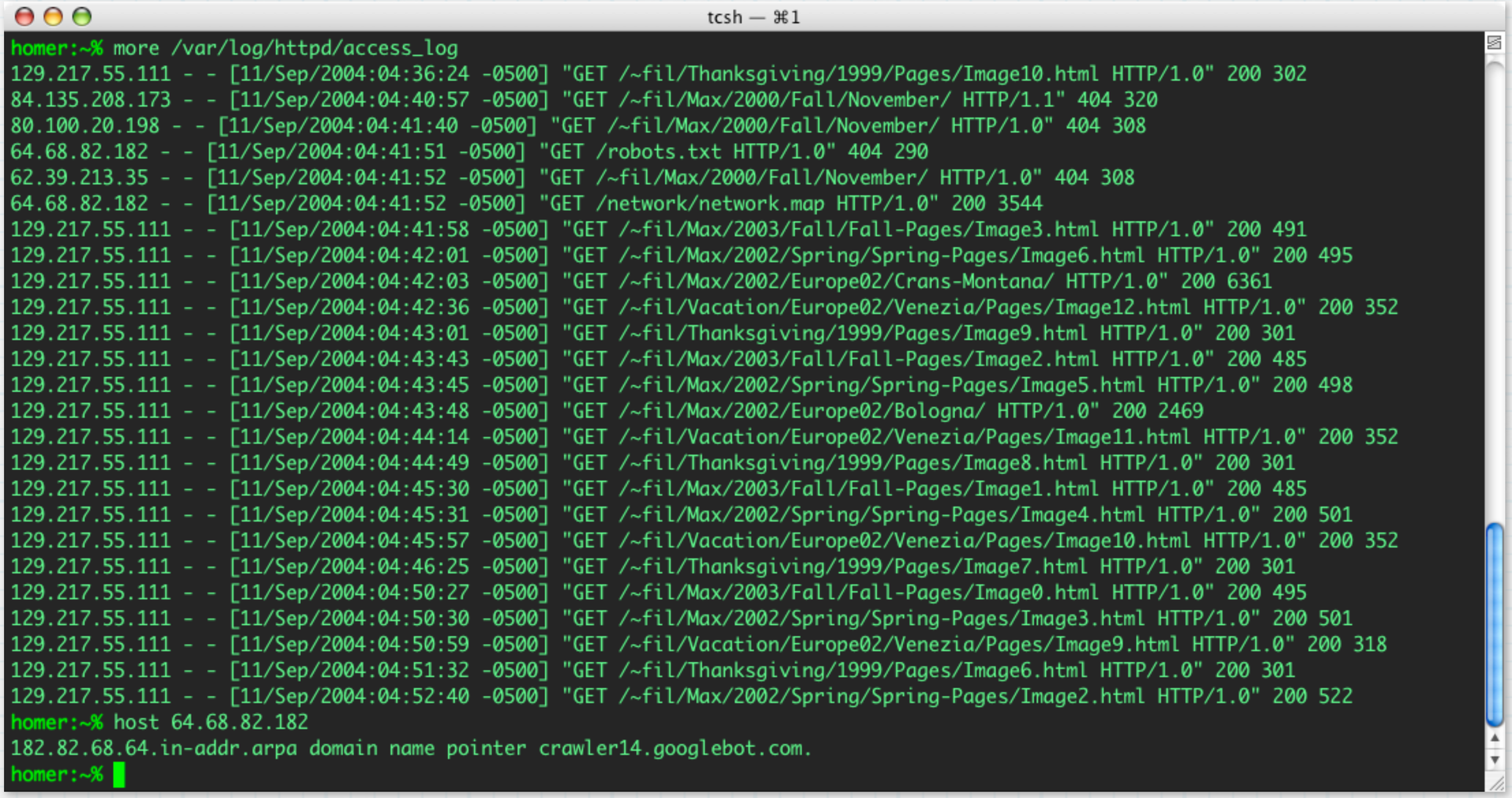


A basic crawler

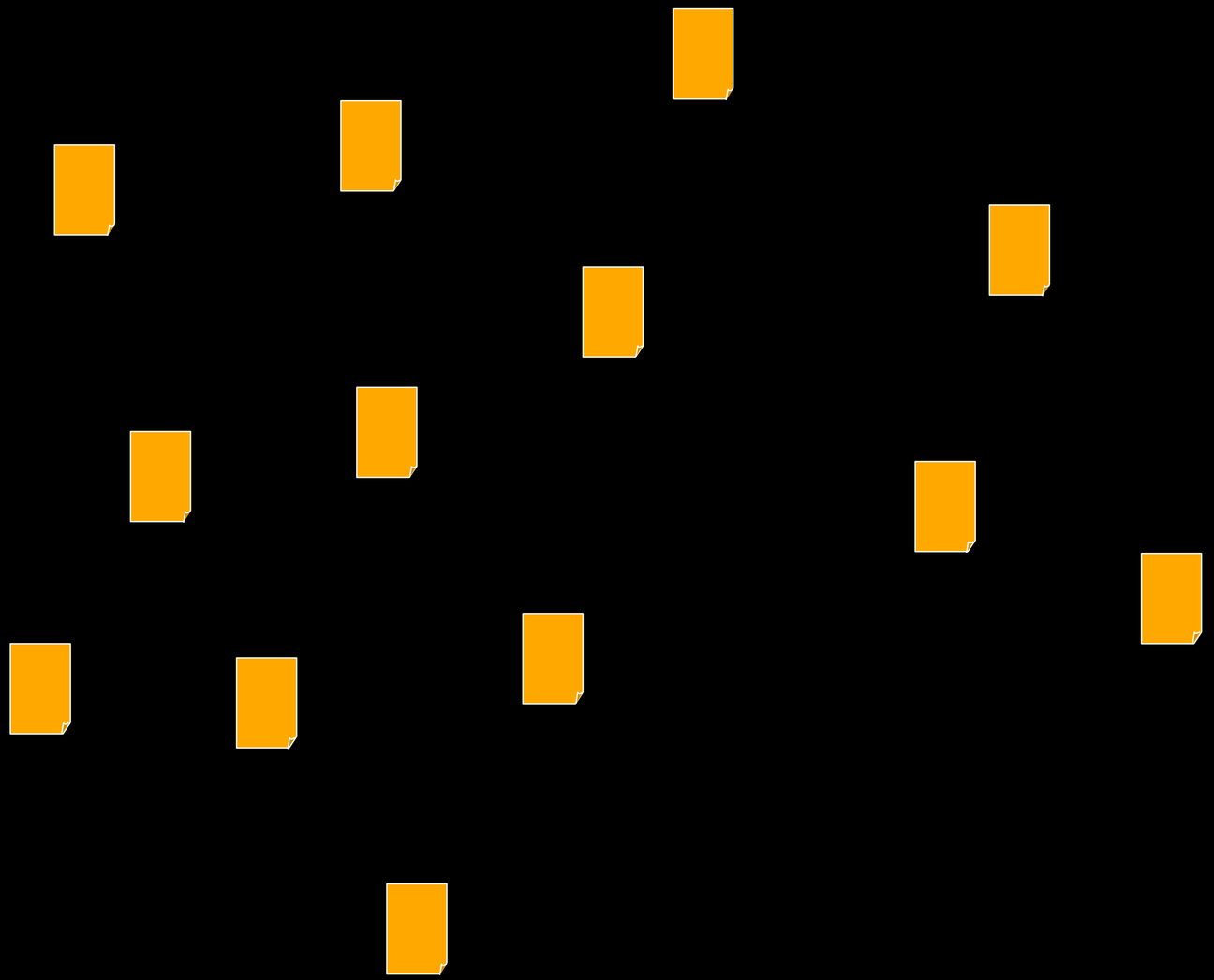
- This is a sequential crawler
- Seeds can be any list of starting URLs
- Order of page visits is determined by frontier data structure
- Stop criterion can be anything

Bing Liu: *Web Data Mining*. Springer, 2007
Ch. 8 *Web Crawling* by Filippo Menczer

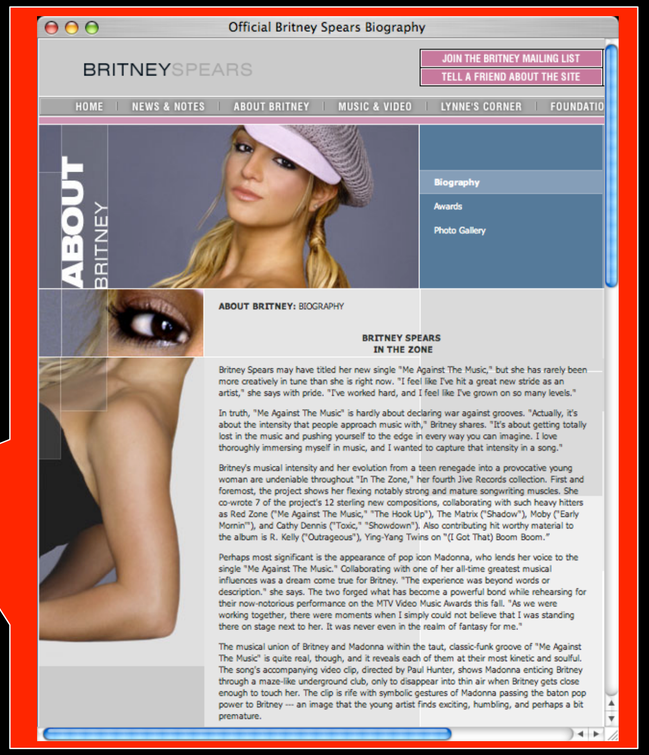
Googlebot & you



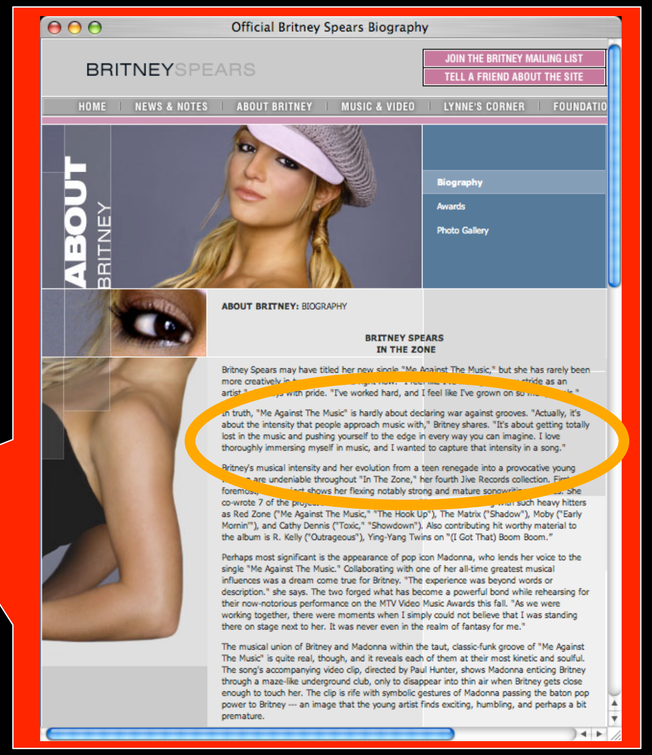
Index



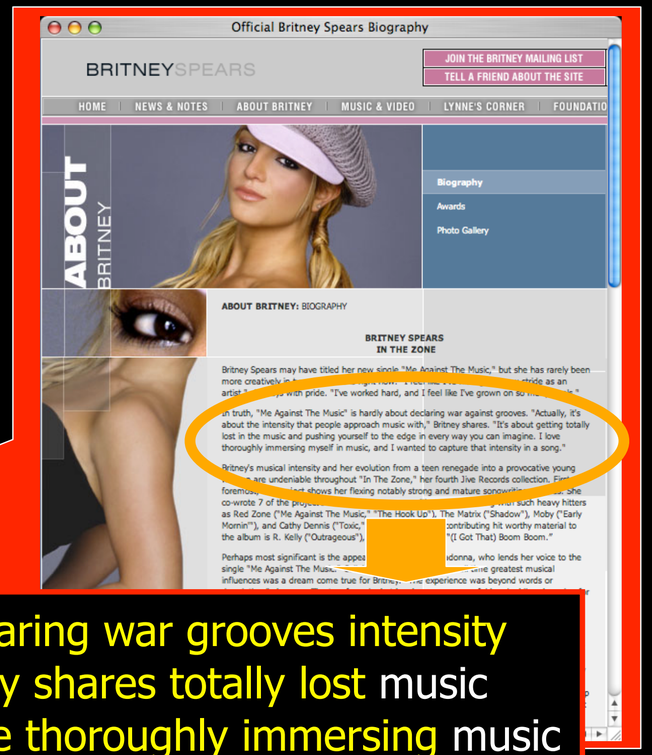
Index


Official Britney Spears Biography
BRITNEYSPEARS
JOIN THE BRITNEY MAILING LIST
TELL A FRIEND ABOUT THE SITE
HOME | NEWS & NOTES | ABOUT BRITNEY | MUSIC & VIDEO | LYNNE'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 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

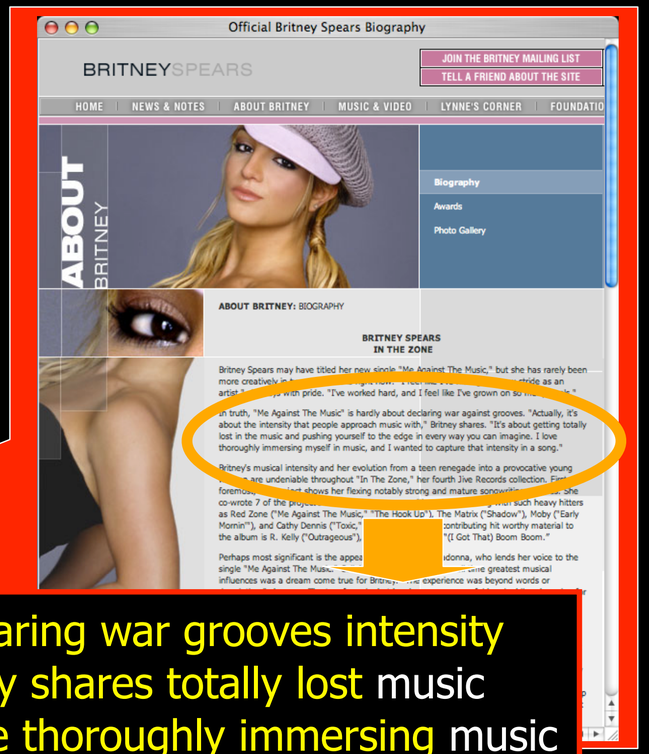

Official Britney Spears Biography
BRITNEYSPEARS
JOIN THE BRITNEY MAILING LIST
TELL A FRIEND ABOUT THE SITE
HOME | NEWS & NOTES | ABOUT BRITNEY | MUSIC & VIDEO | LYNNE'S CORNER | FOUNDATIO
ABOUT BRITNEY
Biography
Awards
Photo Gallery
ABOUT BRITNEY: BIOGRAPHY
BRITNEY SPEARS
IN THE ZONE
Britney Spears may have titled her new single "Me Against The Music," but she has rarely been
more creatively
with pride. "I've worked hard, and I feel like I've grown on so
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
are undeniable throughout "In The Zone," her fourth Jive Records collection.
foremost
shows her flexing notably strong and mature songwriting
co-wrote 7 of the project
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



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

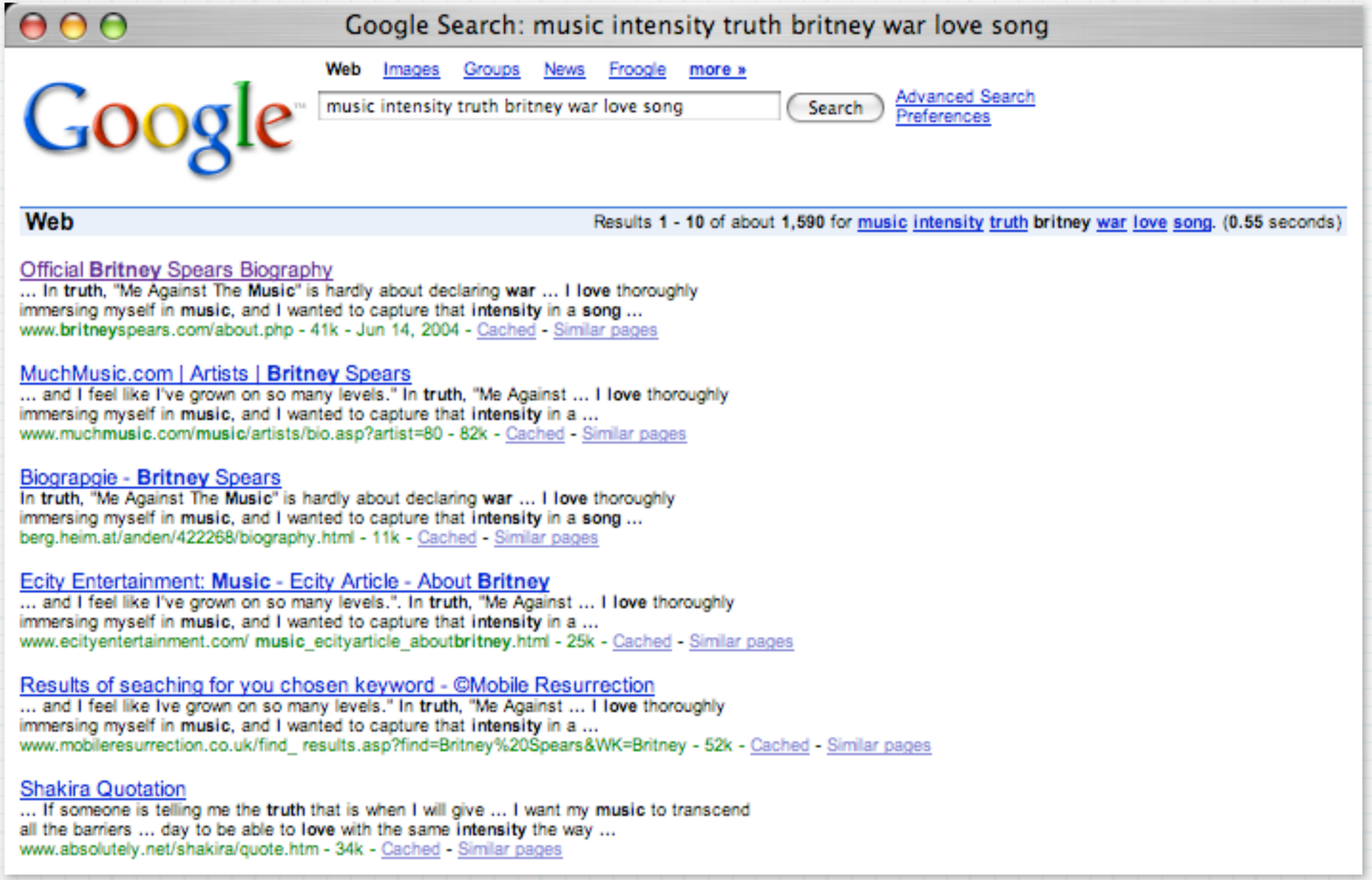
Index



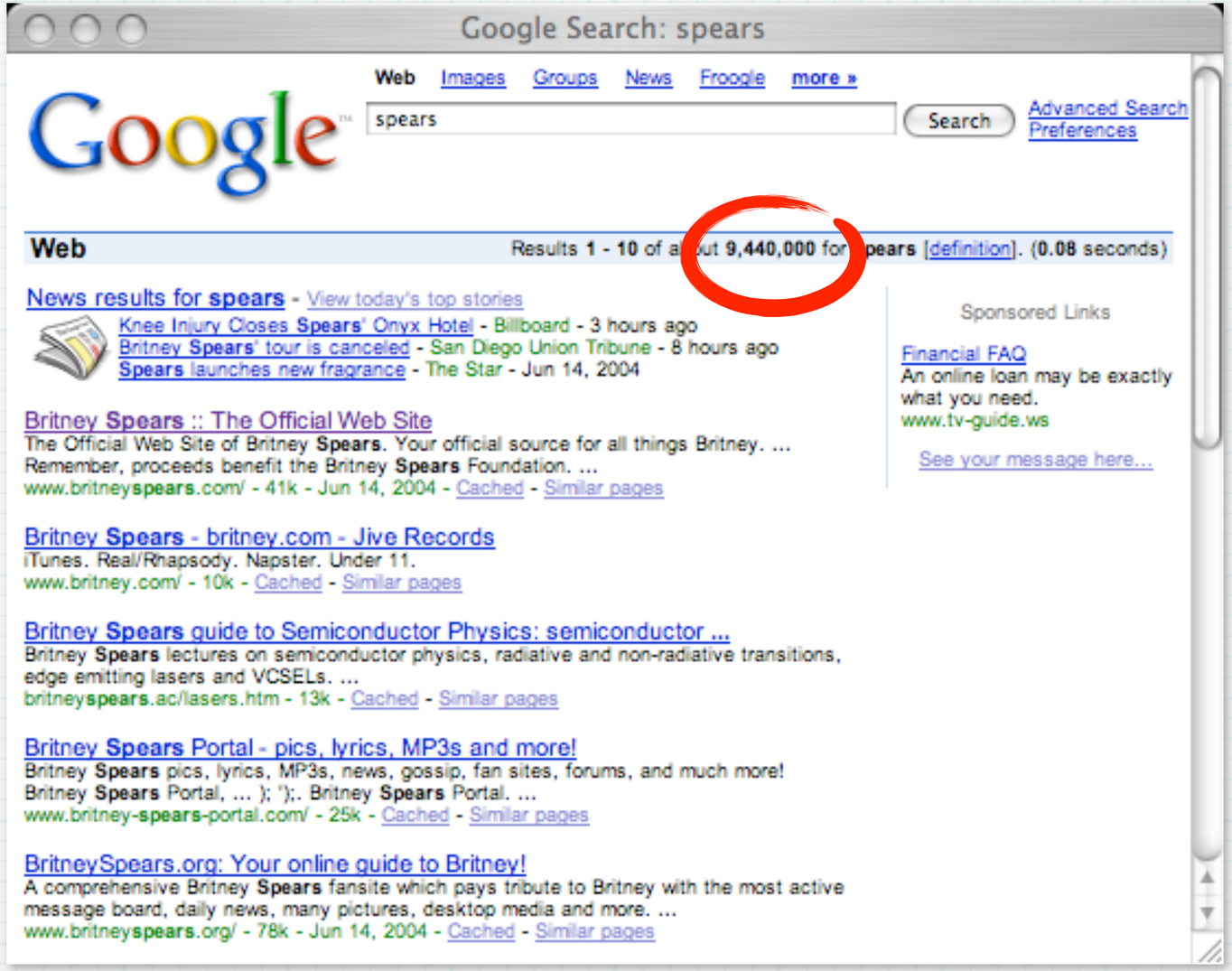
truth music hardly declaring war grooves intensity approach music britney shares totally lost music pushing edge imagine love thoroughly immersing music wanted capture intensity song

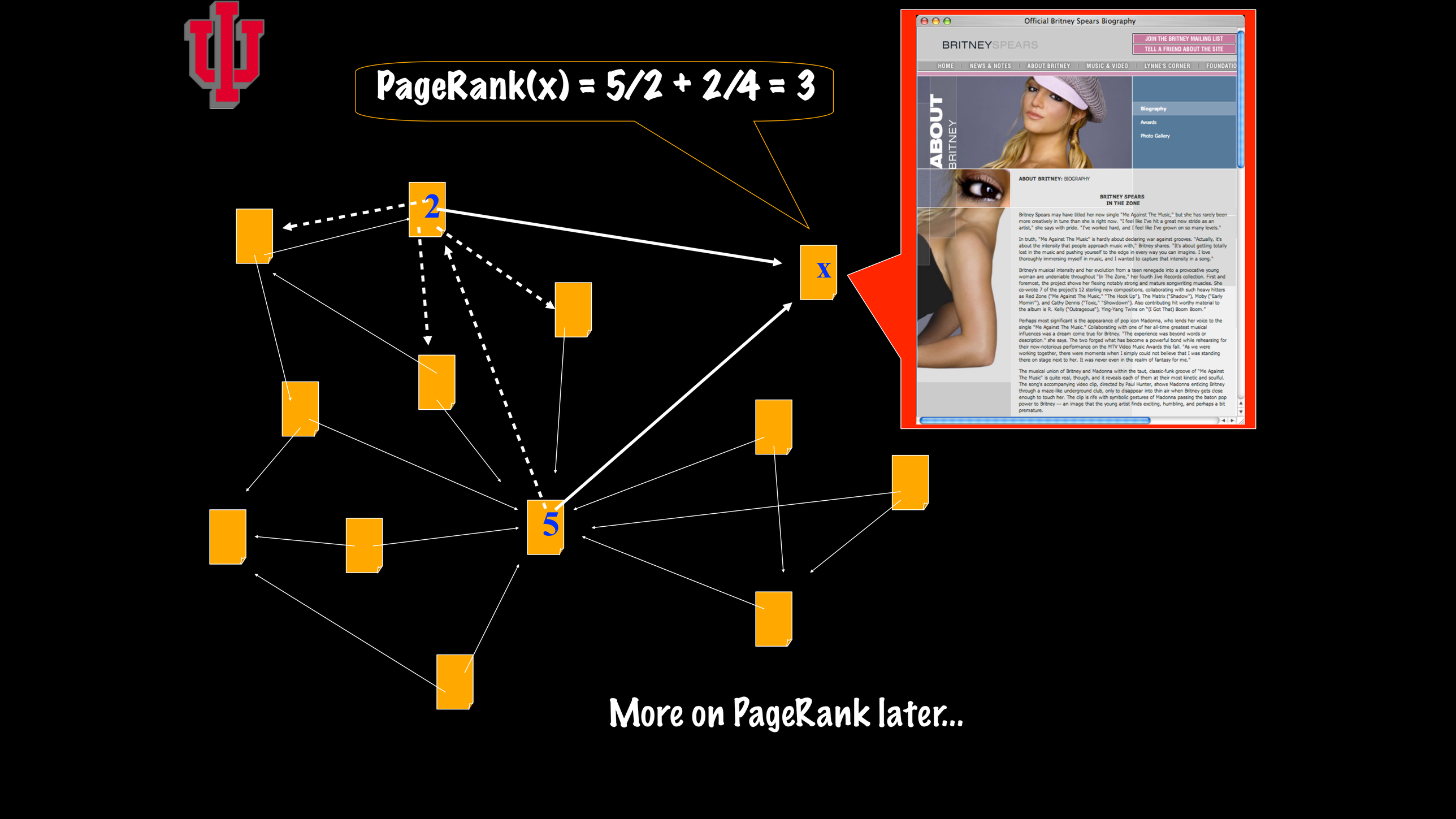
britney	1	http://www.britneyspears.com/about-biography.php
edge	1	http://www.britneyspears.com/about-biography.php
grooves	1	http://www.britneyspears.com/about-biography.php
intensity	2	http://www.britneyspears.com/about-biography.php
love	1	http://www.britneyspears.com/about-biography.php
music	4	http://www.britneyspears.com/about-biography.php
song	1	http://www.britneyspears.com/about-biography.php
truth	1	http://www.britneyspears.com/about-biography.php
war	1	http://www.britneyspears.com/about-biography.php
…		

Query & Retrieval



Is text enough to rank hits?




PageRank(x) = 5/2 + 2/4 = 3
2
x
5
Official Britney Spears Biography
BRITNEYSPEARS
JOIN THE BRITNEY MAILING LIST
TELL A FRIEND ABOUT THE SITE
HOME | NEWS & NOTES | ABOUT BRITNEY | MUSIC & VIDEO | LYNNE'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 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.
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

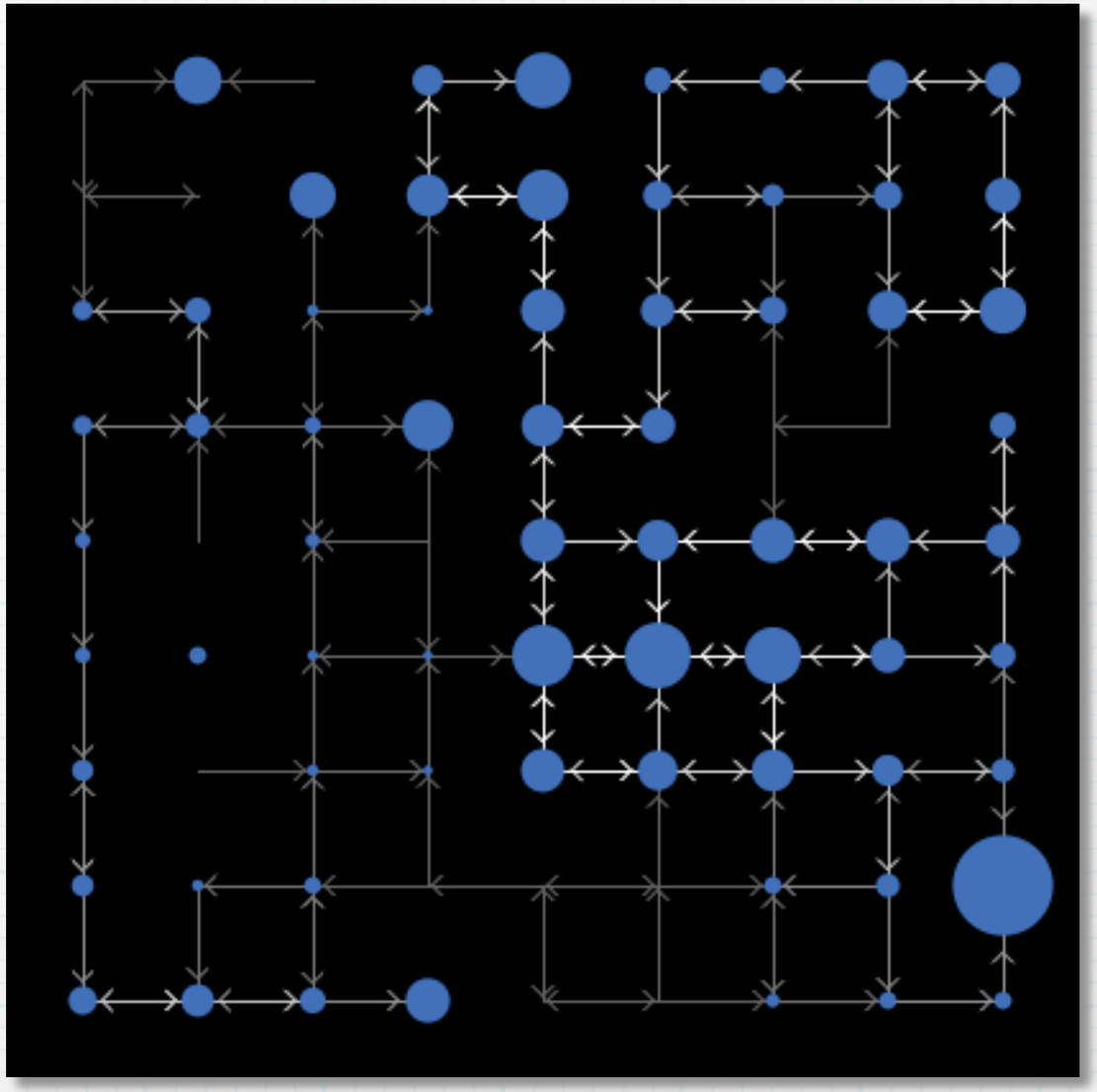
* Brin and Page introduced a measure of "page importance" or "prestige" called PageRank in their seminal Google paper (1998)
  * Invented by Seeley in 1949 (few good ideas are original)
  * Related to HITS authority (Kleinberg)
  * Computed from hyperlink structure
  * Today, PageRank is just one of 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?

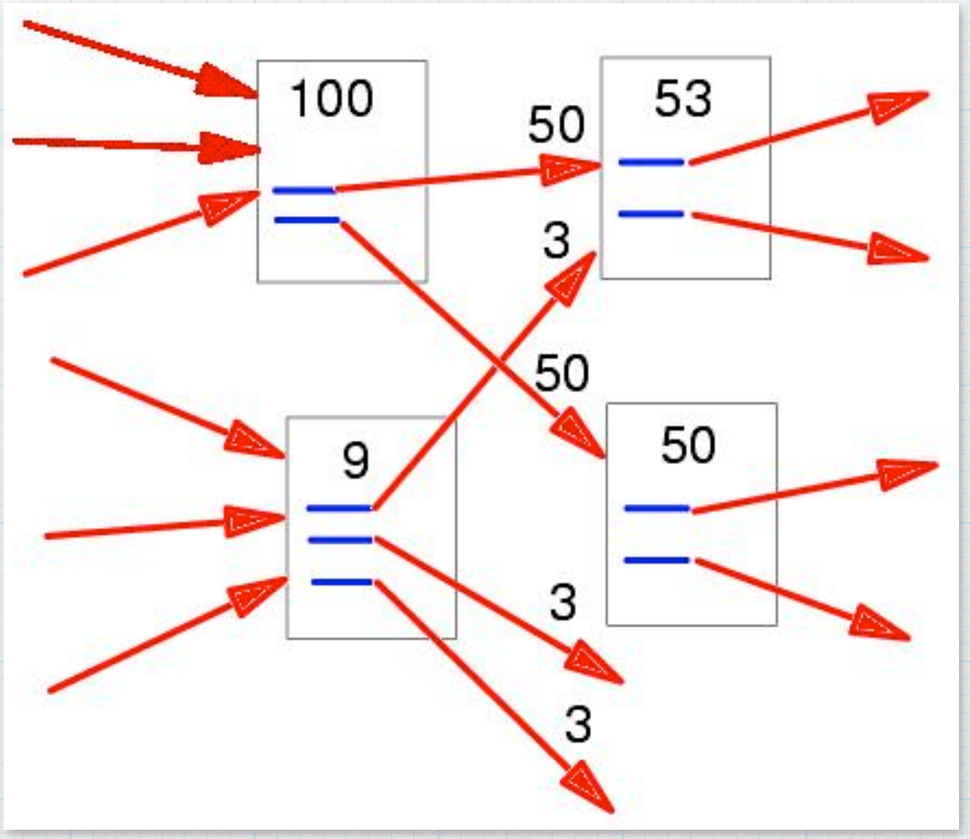


Simplified PageRank definition

- PageRank is received via inliks and distributed via outlinks
- PageRank is "conserved": divided among outlinks

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

- Note recursive definition!



Power (iterative) computation of PR

- Initialize: $R_0(i) = \frac{1}{N}$ (unimportant)
- Repeat foreach time step t, until convergence: $R_{t+1}(i) = \sum_{j:j\to i} \frac{R_t(j)}{k_{out}(j)}$

importance of page i

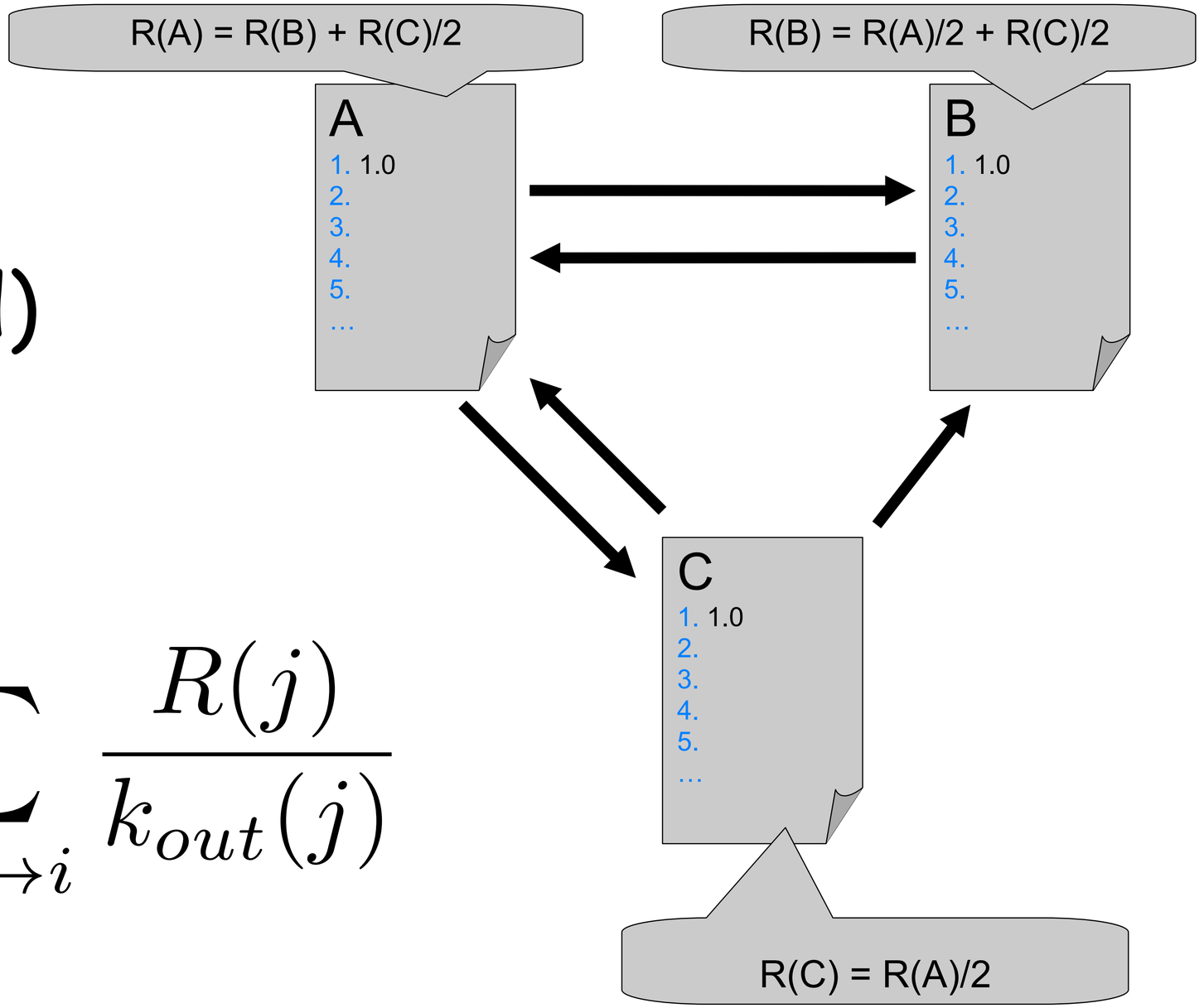
pages j that link to page i

out-degree of page j

importance of page j

(Simplified) PageRank exercise

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



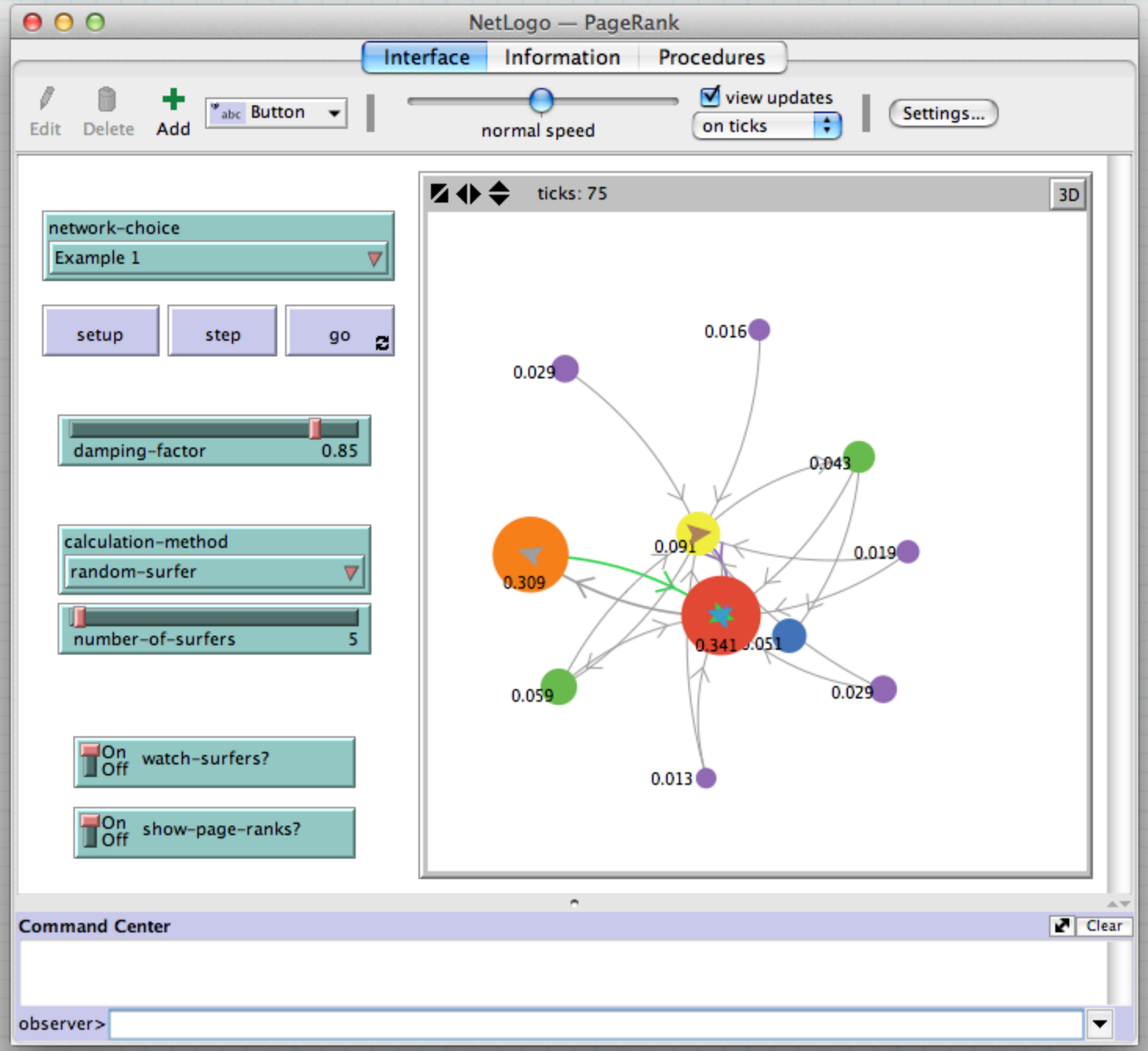
Actual PageRank

- 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 α =0.15 (jump or teleportation factor)
  - Damping factor (1-α), typically 0.85
- Another issue:
  - Cliques of links to artificially inflate PR
  - Search engine use various measures to combat "spamdexing"

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

NetLogo

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

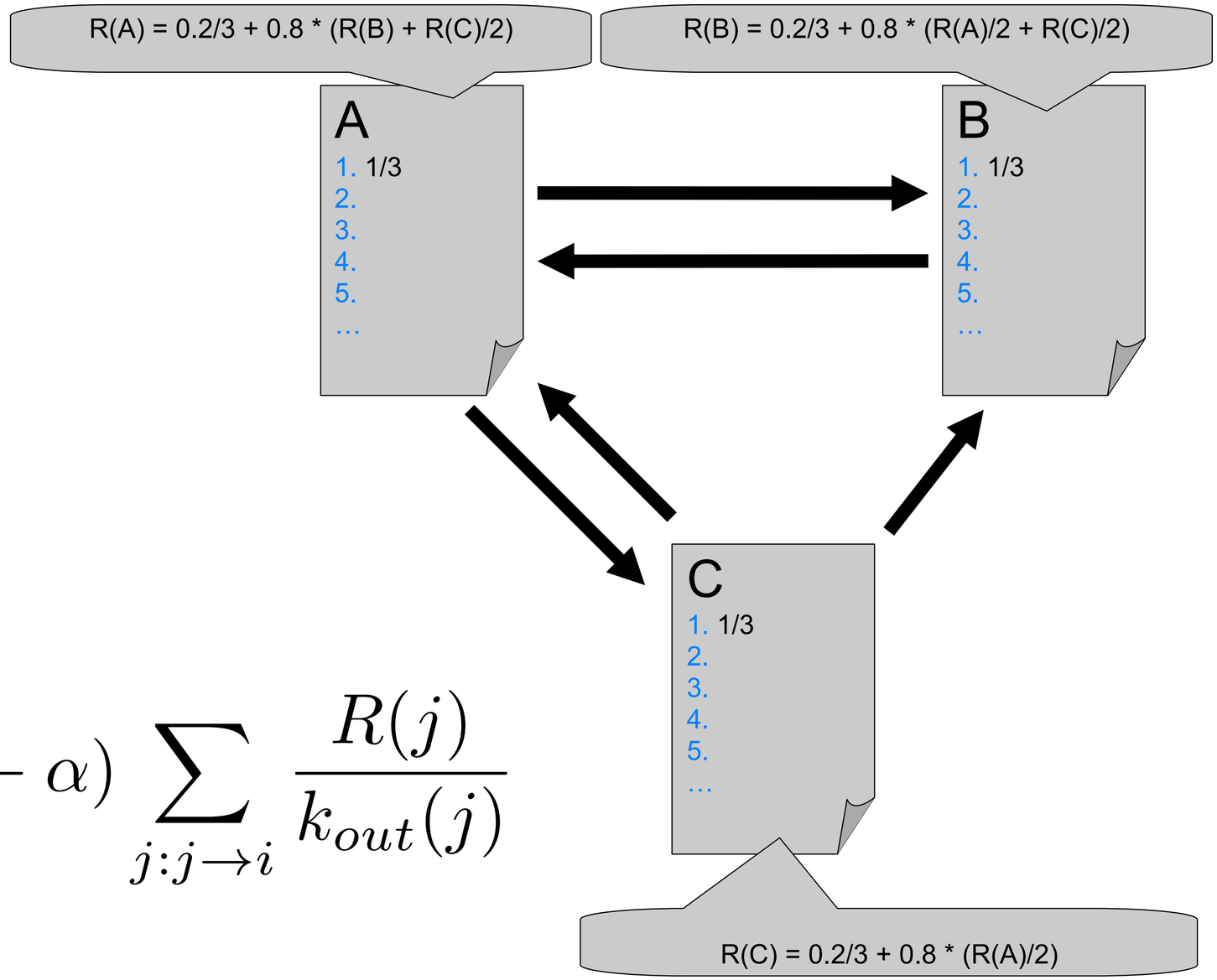


INDIANA UNIVERSITY BLOOMINGTON

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

carl.cs.indiana.edu/fil/cgi-bin/pagerank/

go.iu.edu/pagerank



(Actual) PageRank exercise

α =0.2

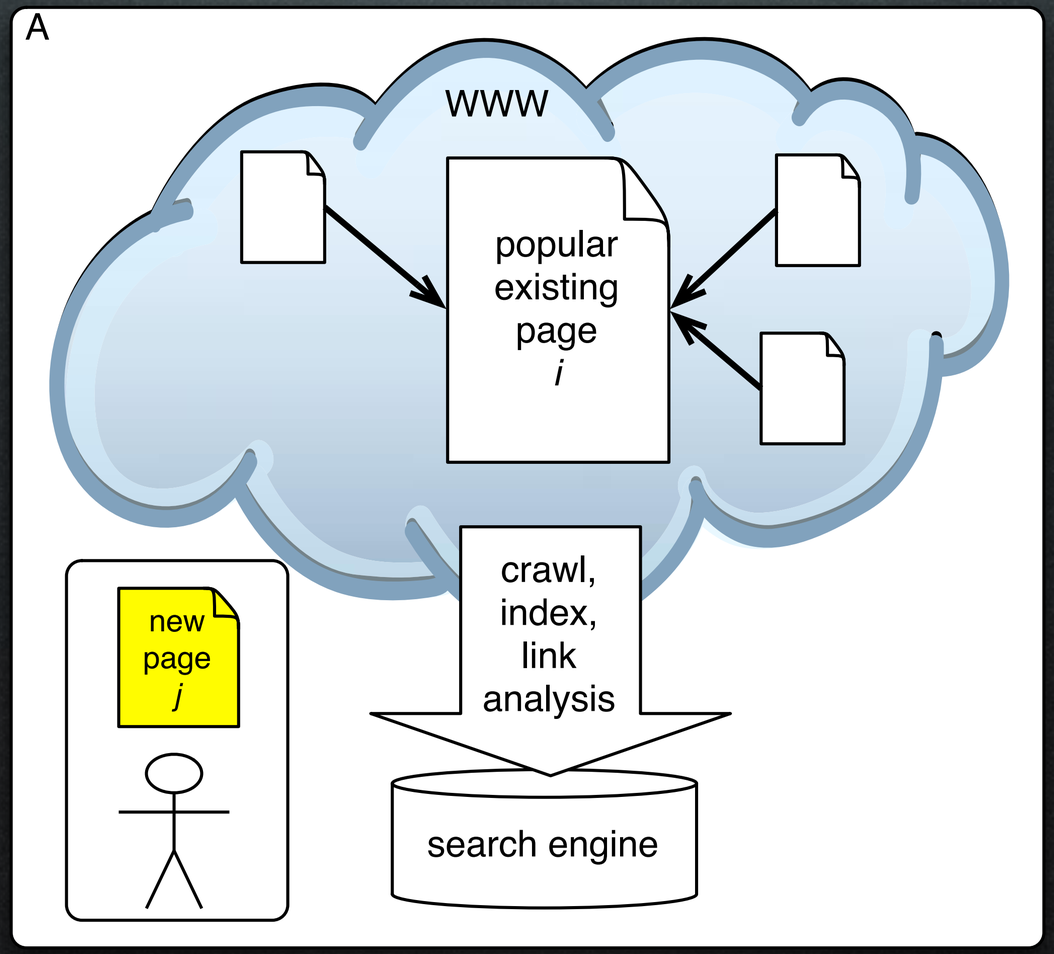
$$R(i) = \frac{\alpha}{N} + (1-\alpha) \sum_{j:j\to i} \frac{R(j)}{k_{out}(j)}$$

Also see the PR calculator at http://www.webworkshop.net/pagerank_calculator.php

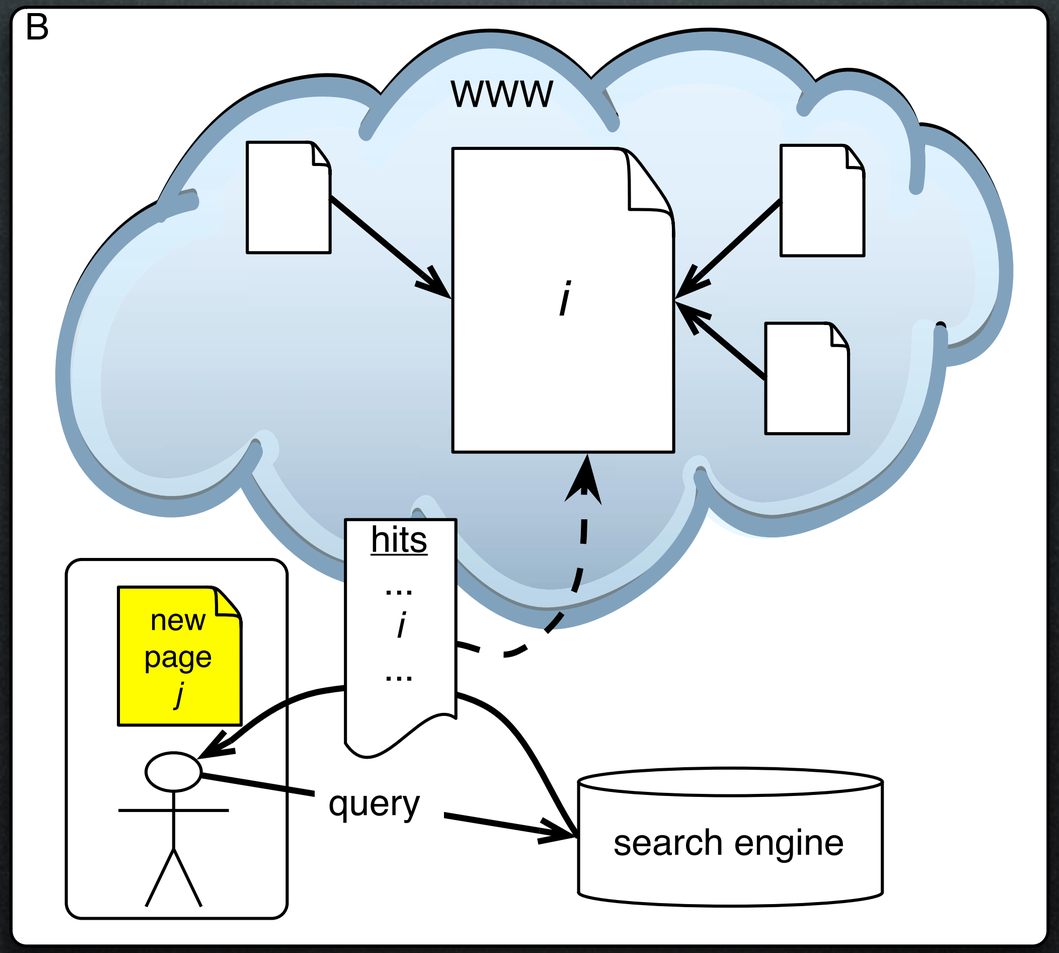
PageRank competition

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

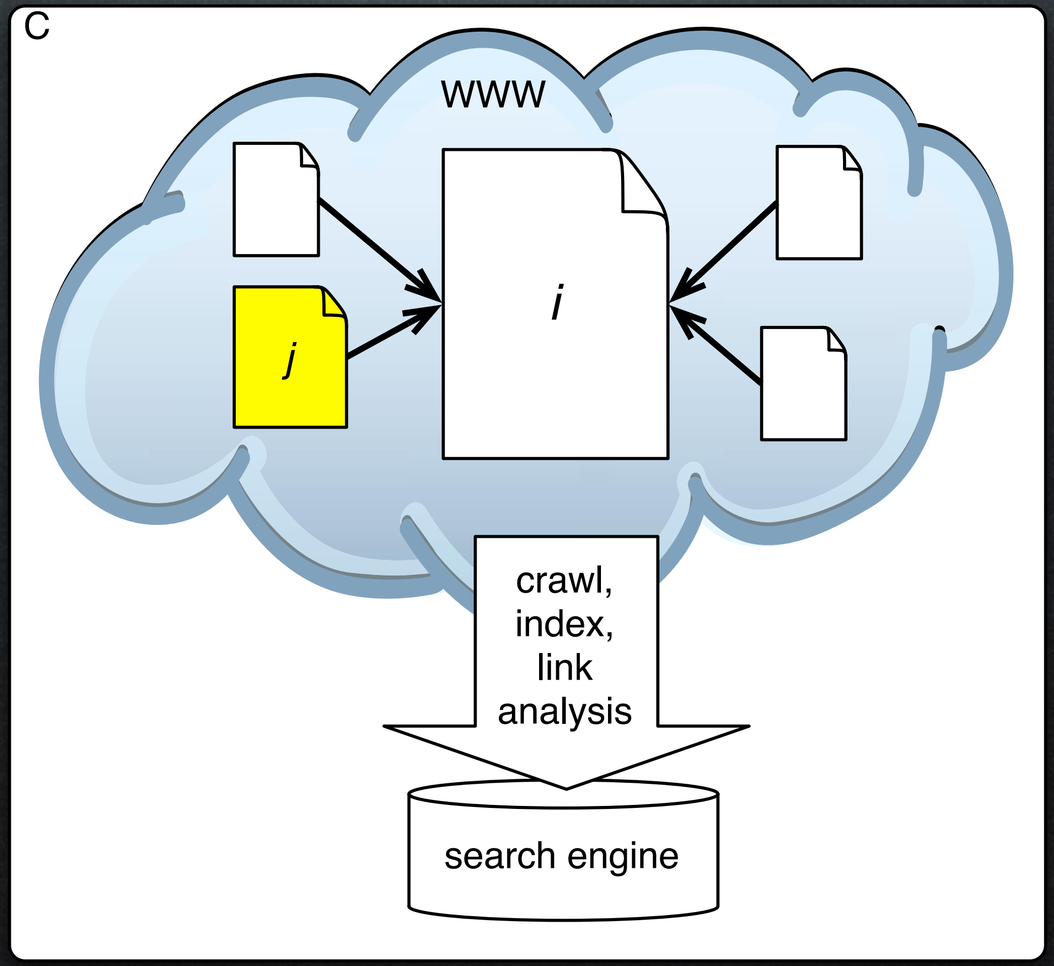
Is there a popularity bias?
("Entrenchment", "Googlearchy")

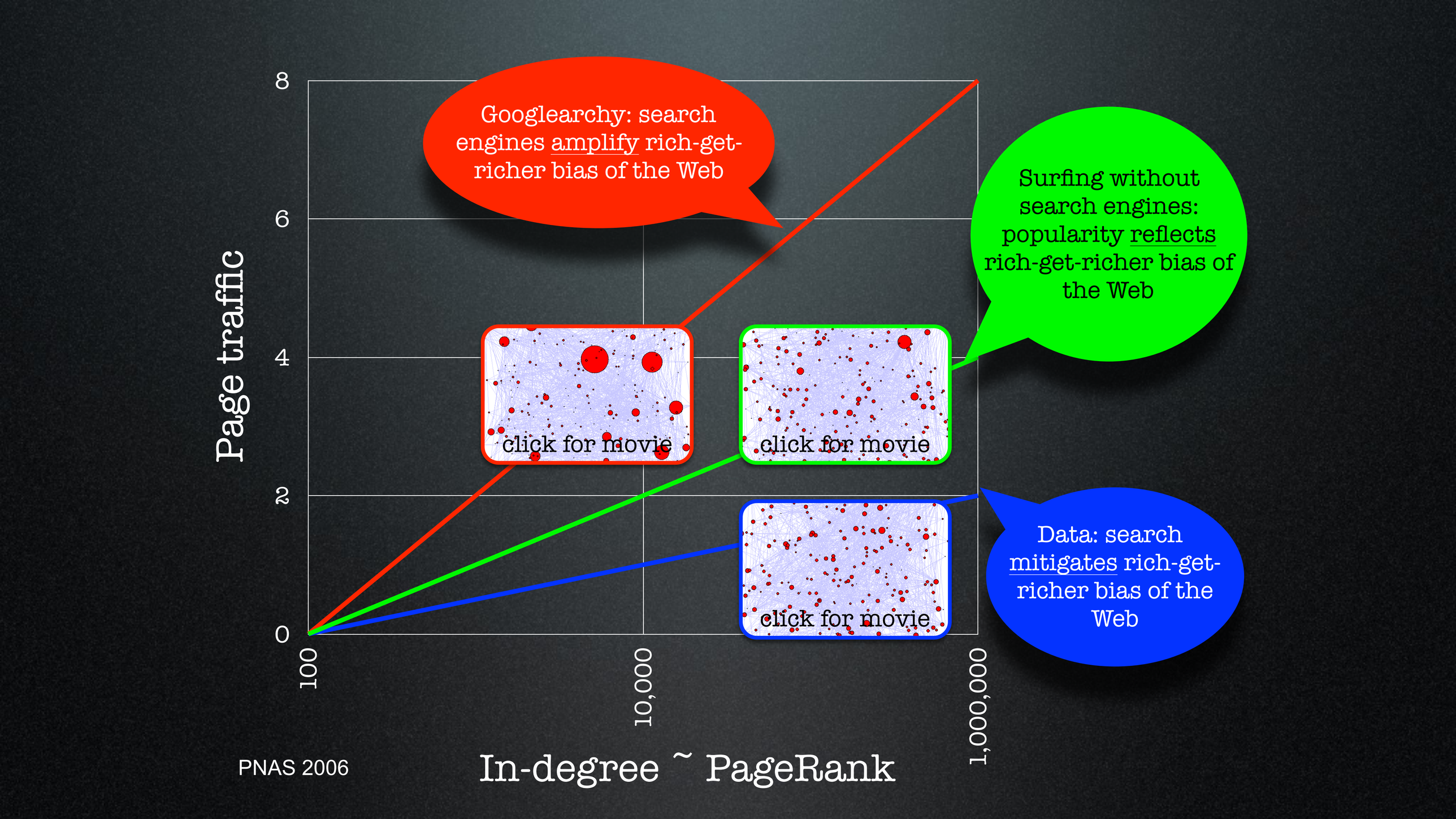


Is there a popularity bias? ("Entrenchment", "Googlearchy")



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



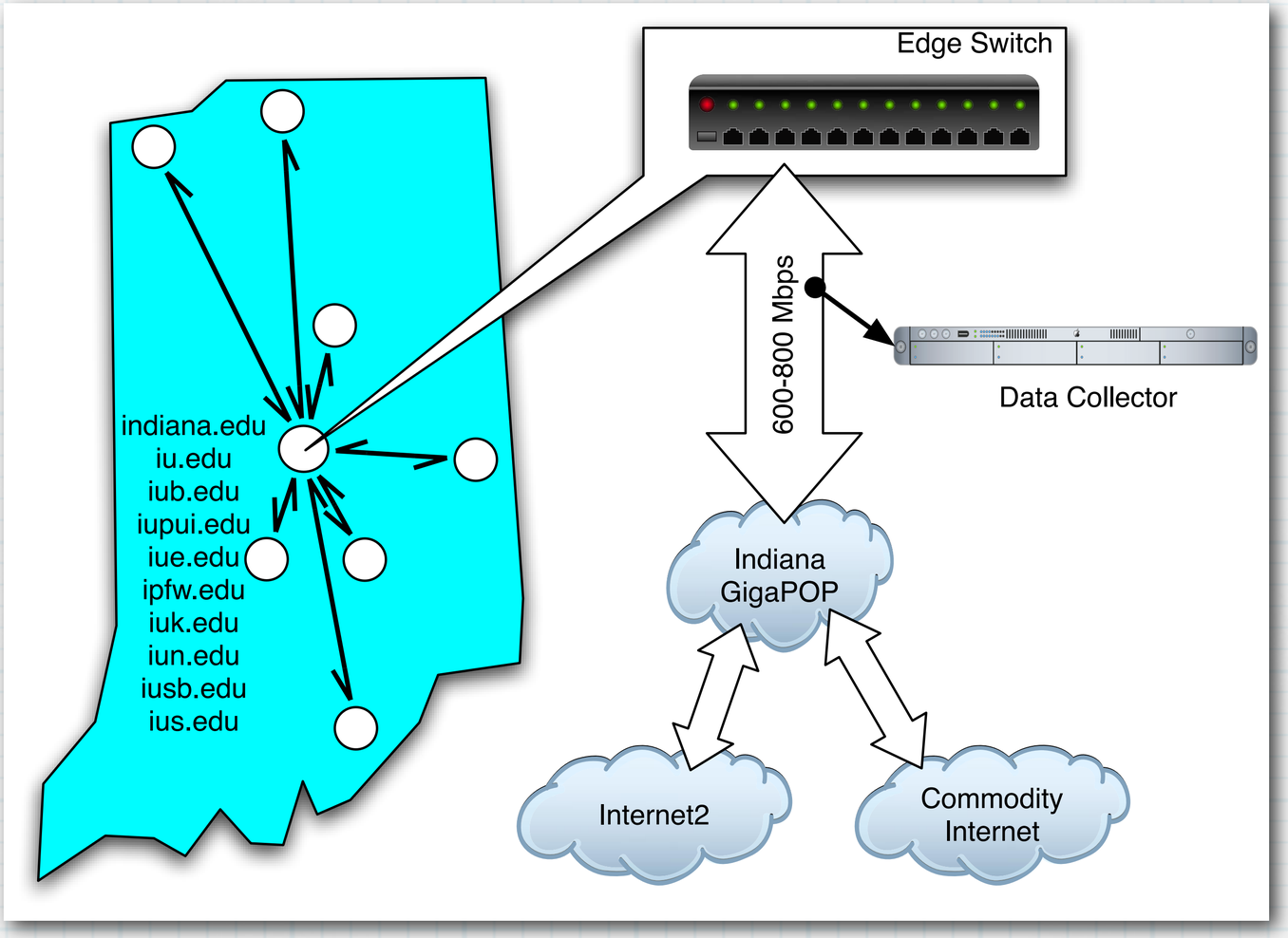

Googlearchy: search engines amplify rich-get-richer bias of the Web
Surfing without search engines: popularity reflects rich-get-richer bias of the Web
Data: search mitigates rich-get-richer bias of the Web
click for movie
click for movie
click for movie
Page traffic
0
2
4
6
8
100
10,000
1,000,000
In-degree ~ PageRank
PNAS 2006

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

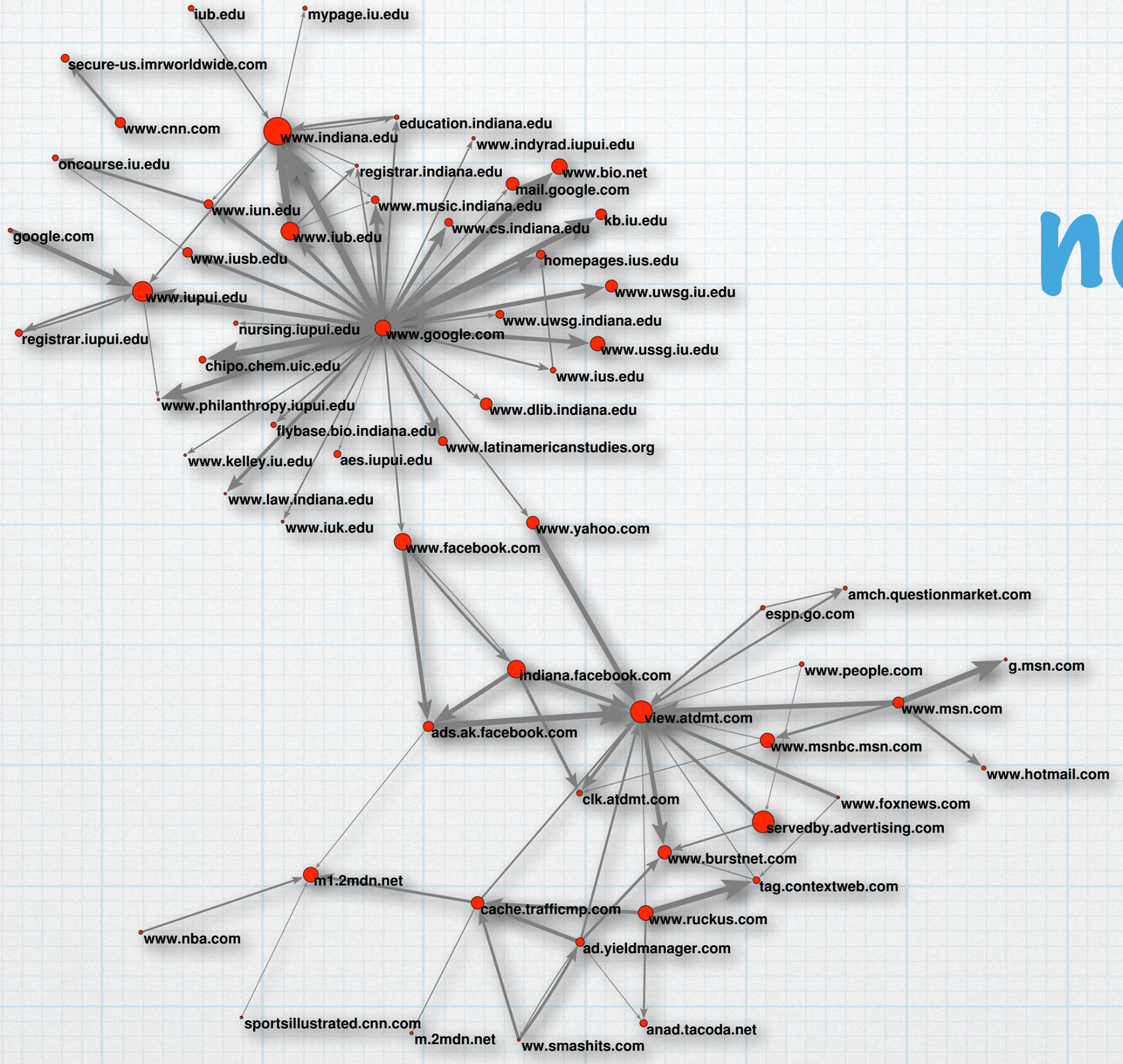
- PageRank competition!
- Review
  - Web protocol
  - Search engines components
  - Crawlers
  - Essential elements of search ranking
  - PageRank centrality

IU Web traffic



Traffic networks

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



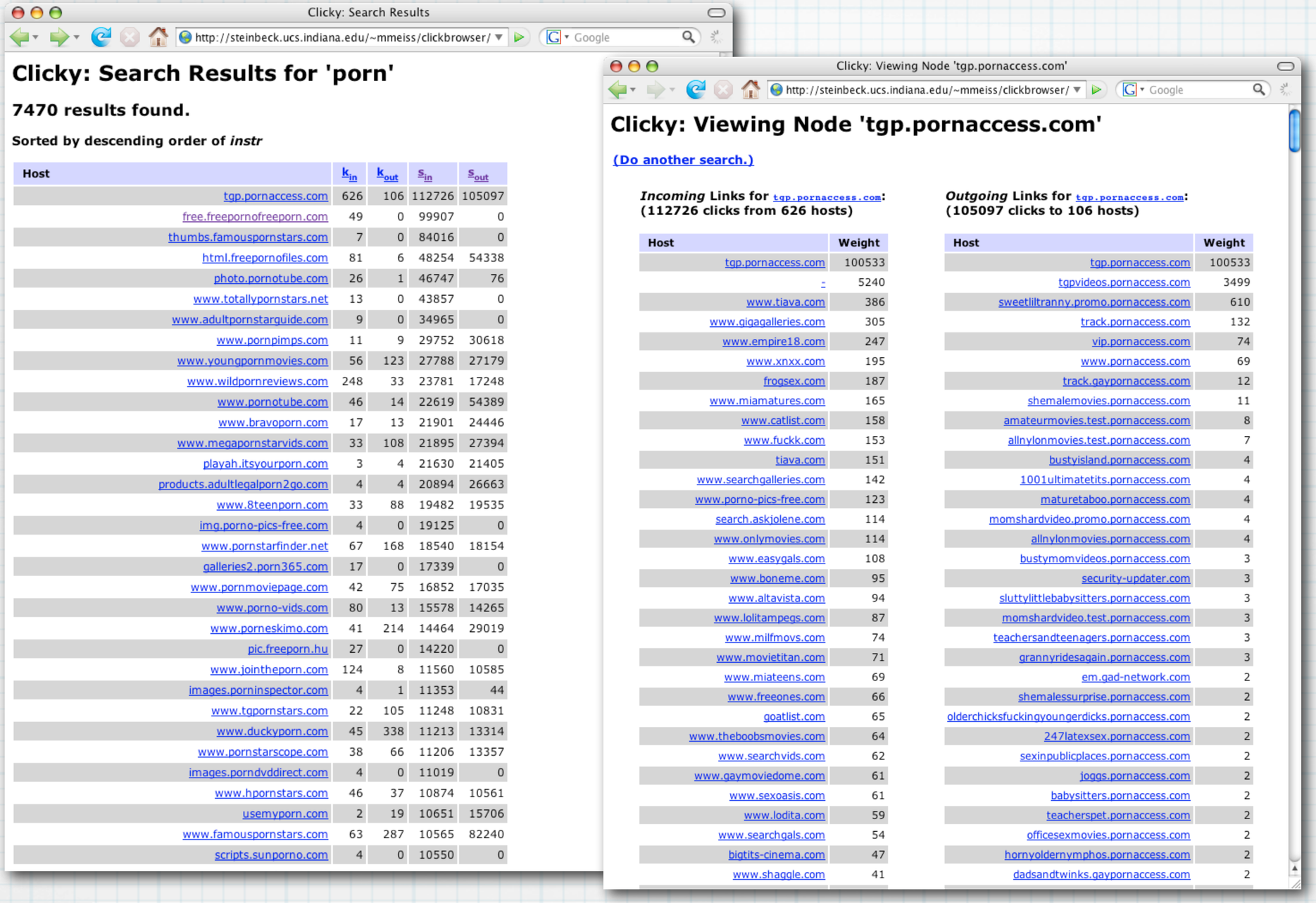
* Weighted
* Directed

Clicky: Search Results for 'porn'

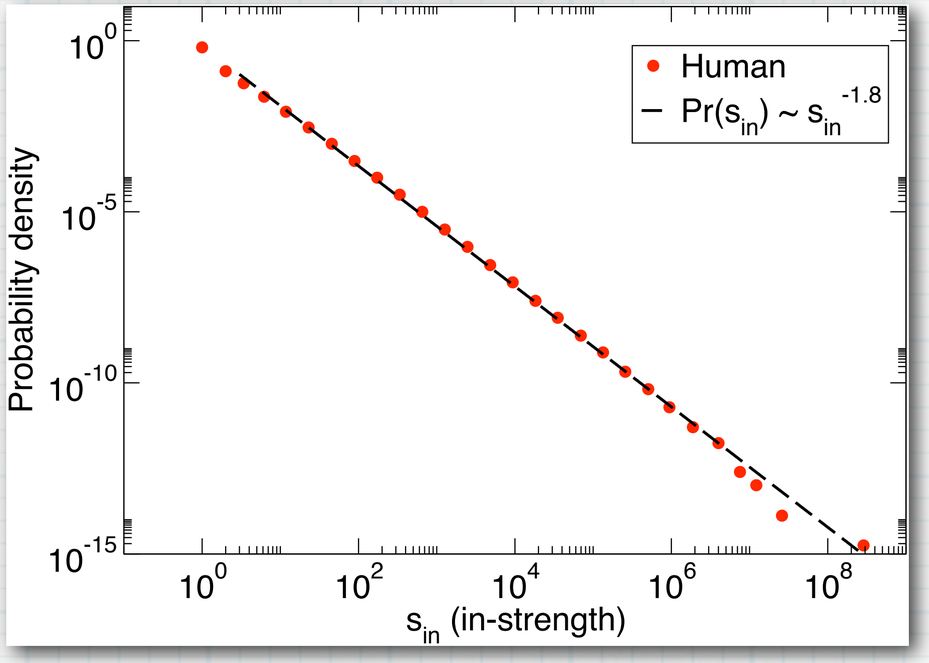
7470 results found.

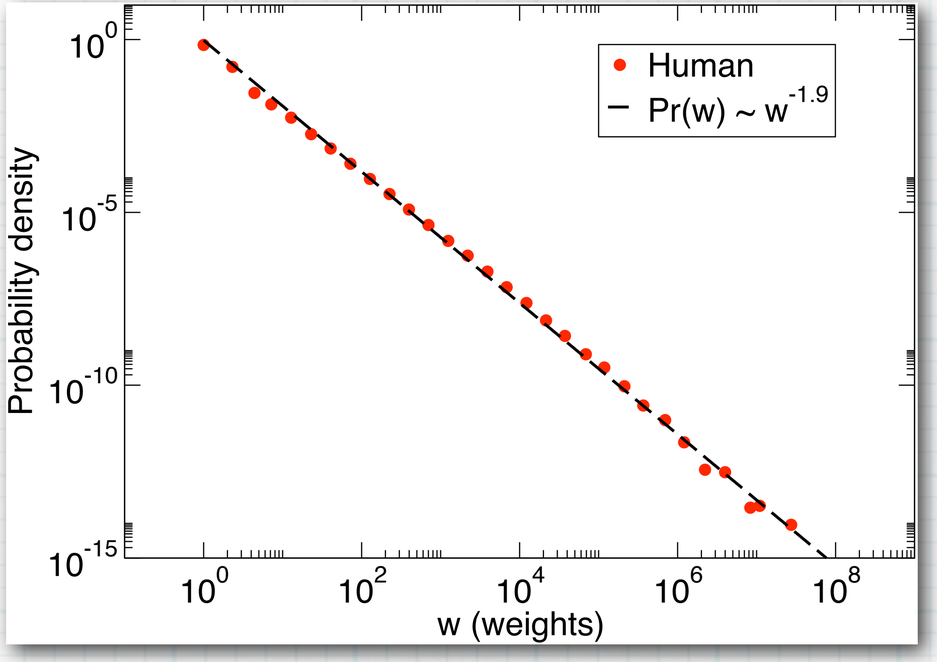
Sorted by descending order of *instr*

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



Web traffic properties

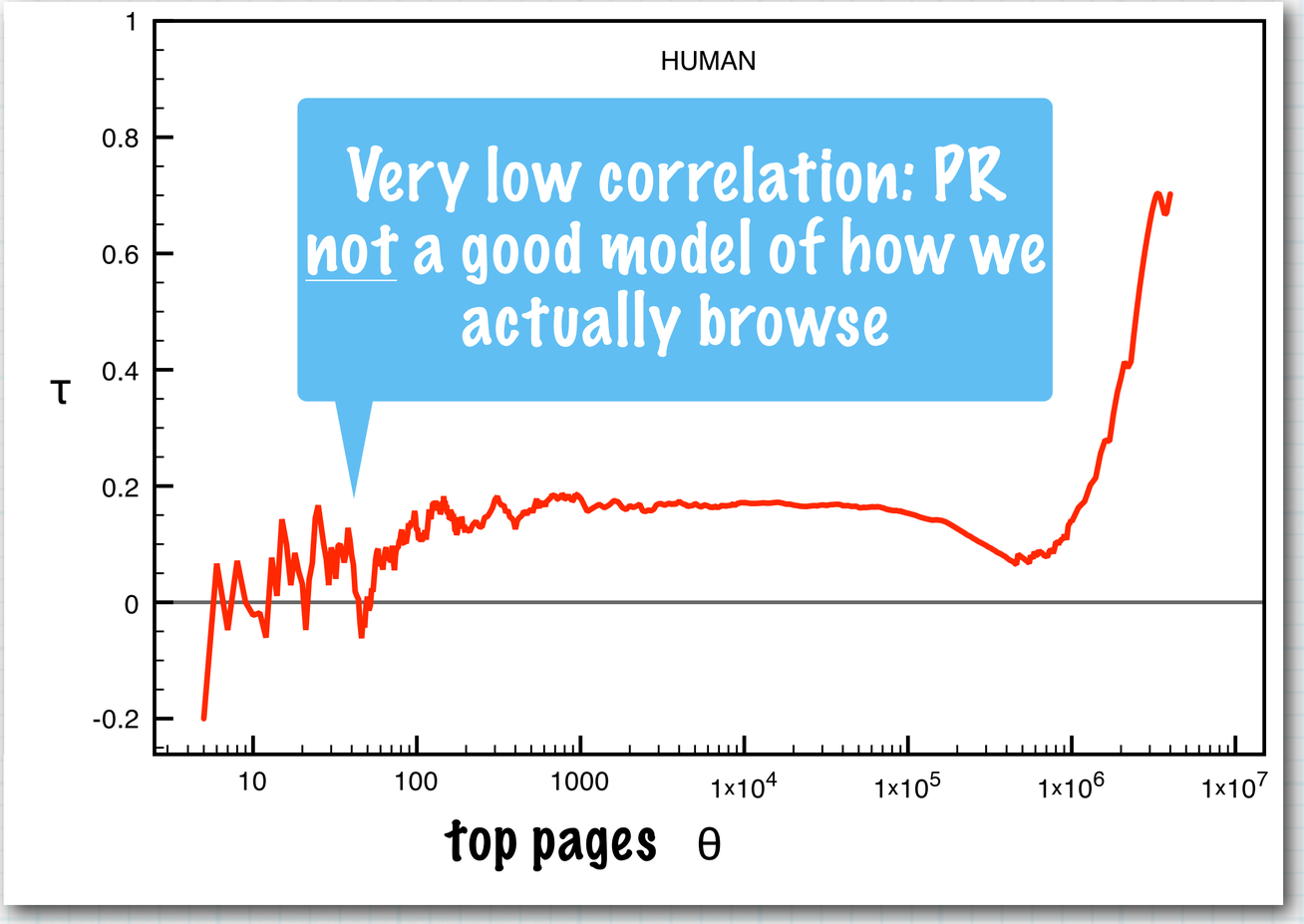




PageRank revisited

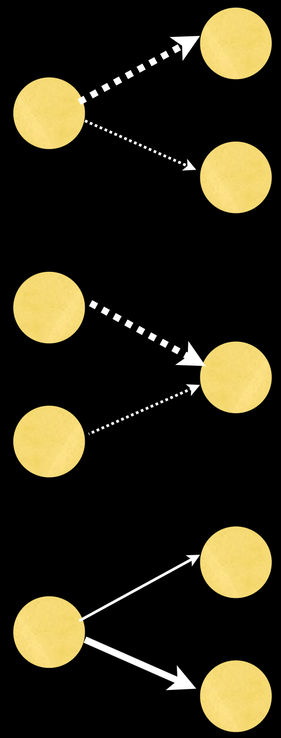
* PR as a model of Web navigation: stationary distribution of visit frequency by a modified random walk (with jumps) on the Web graph
* Is this a good model of Web traffic?
  * From an application perspective, we care about the ranking of sites
  * Compare with actual site traffic (in-strength) by measuring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

Kendall's rank correlation



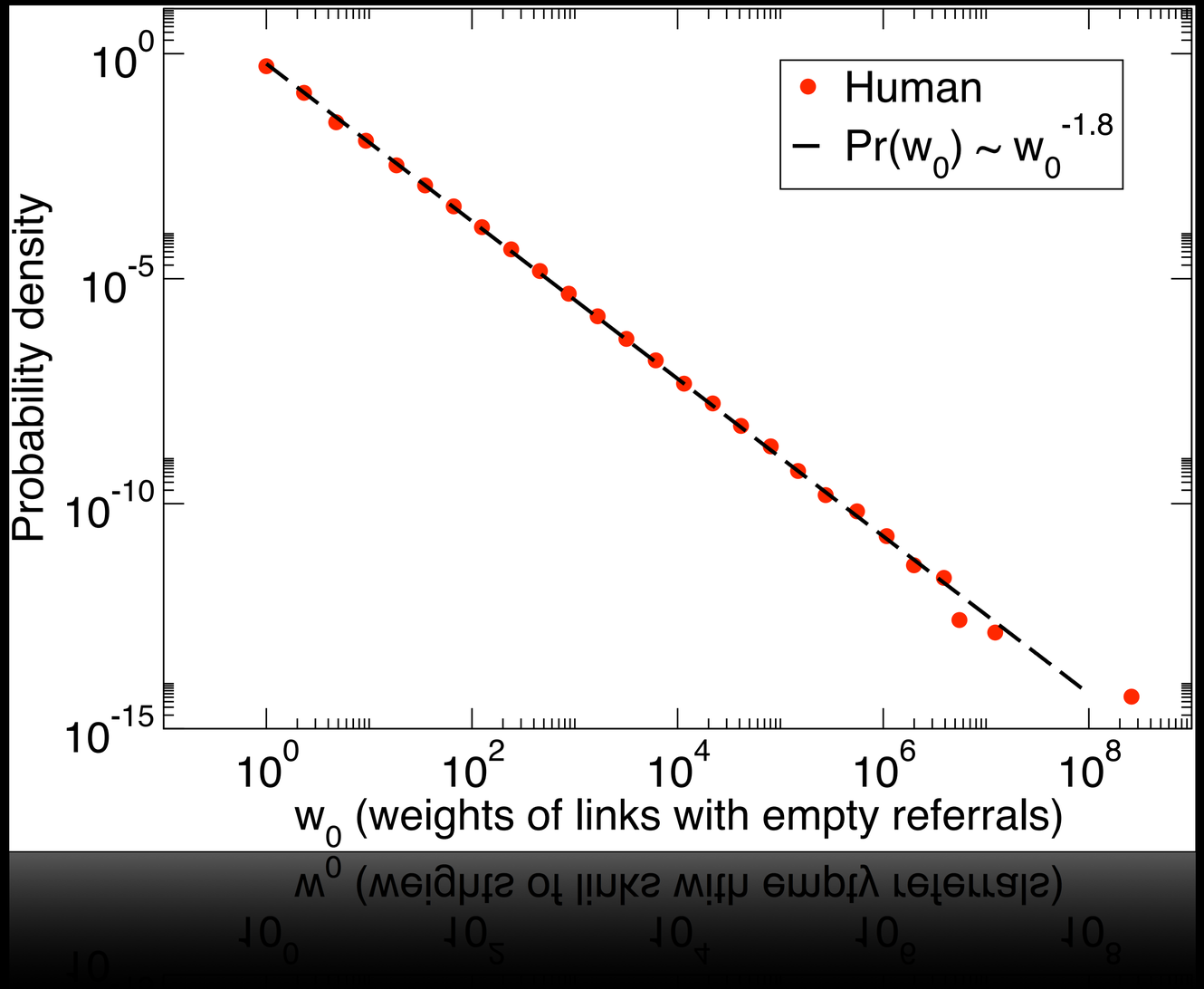
PageRank assumptions

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

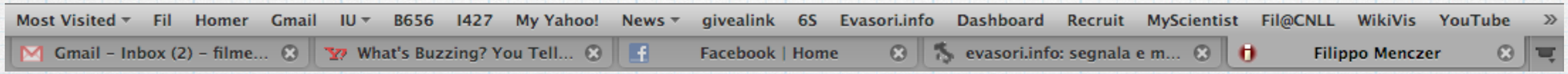


$$PRW(j) = \frac{\alpha}{N} + (1-\alpha) \sum_{i:w_{ij}\neq 0} \frac{w_{ij}}{s_{out}(i)} PRW(i)$$

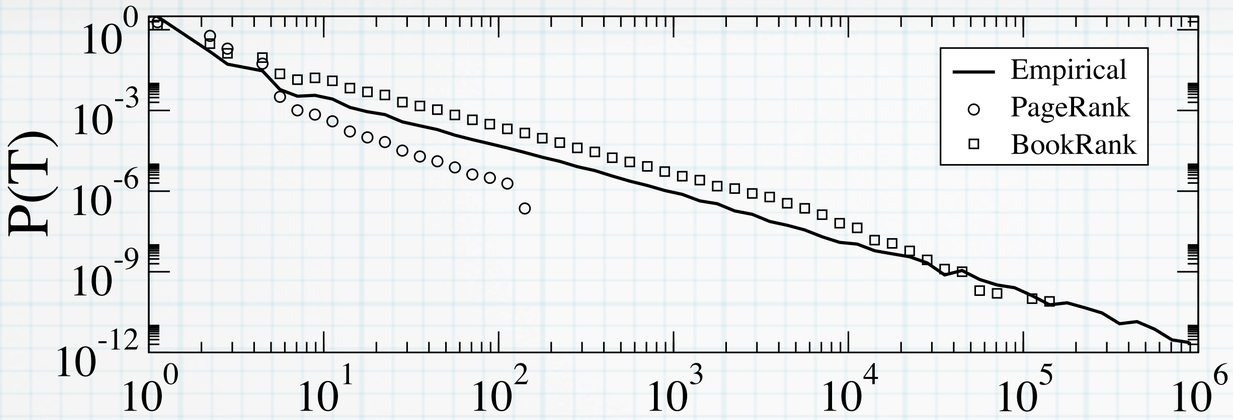
Teleportation target heterogeneity



Ingredients for an agent-based model

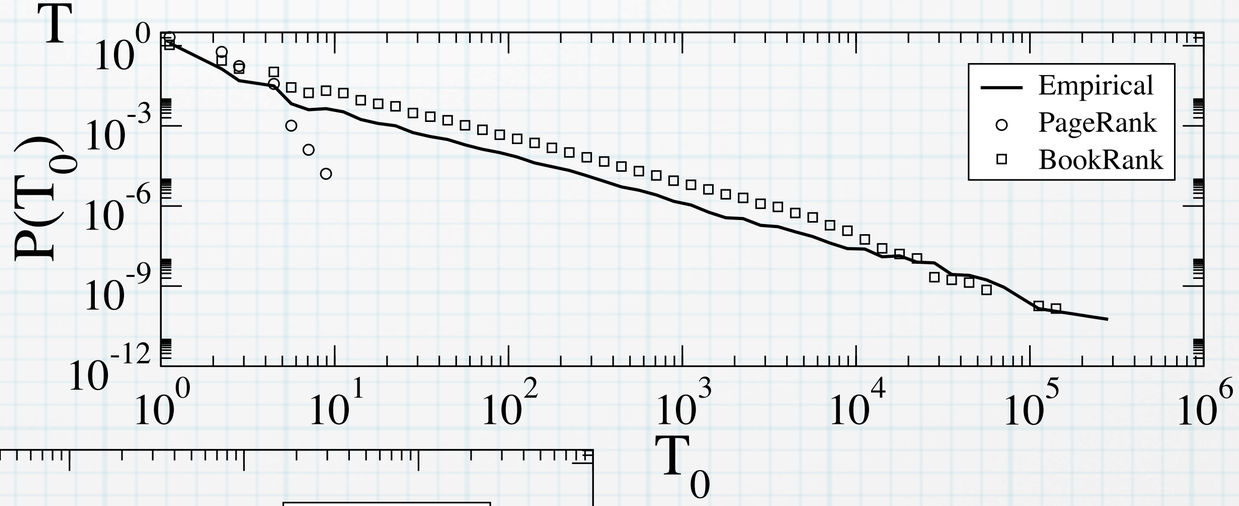


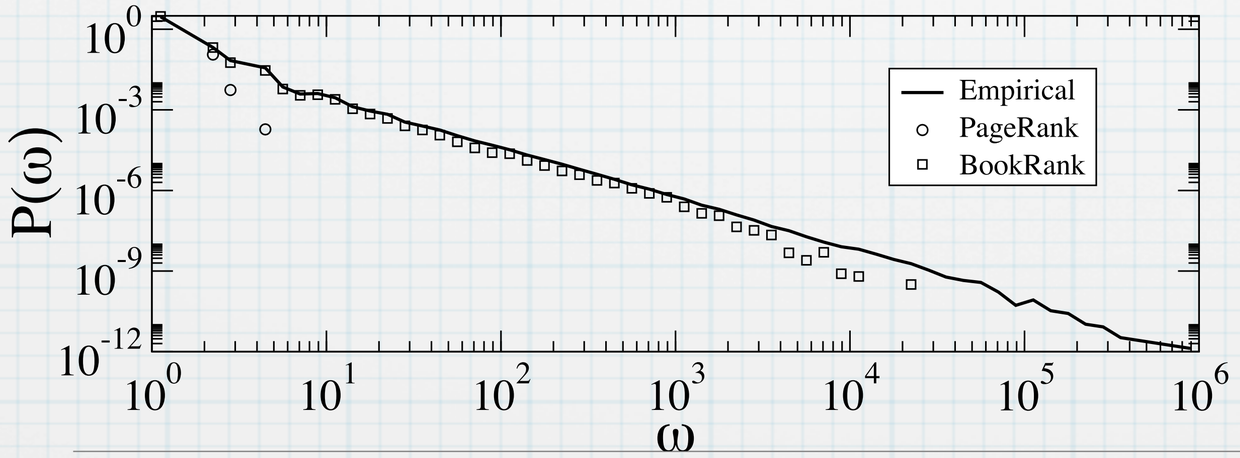
1. Bookmarks (memory): we remember the pages we like and are more likely to start browsing from those



Page Traffic

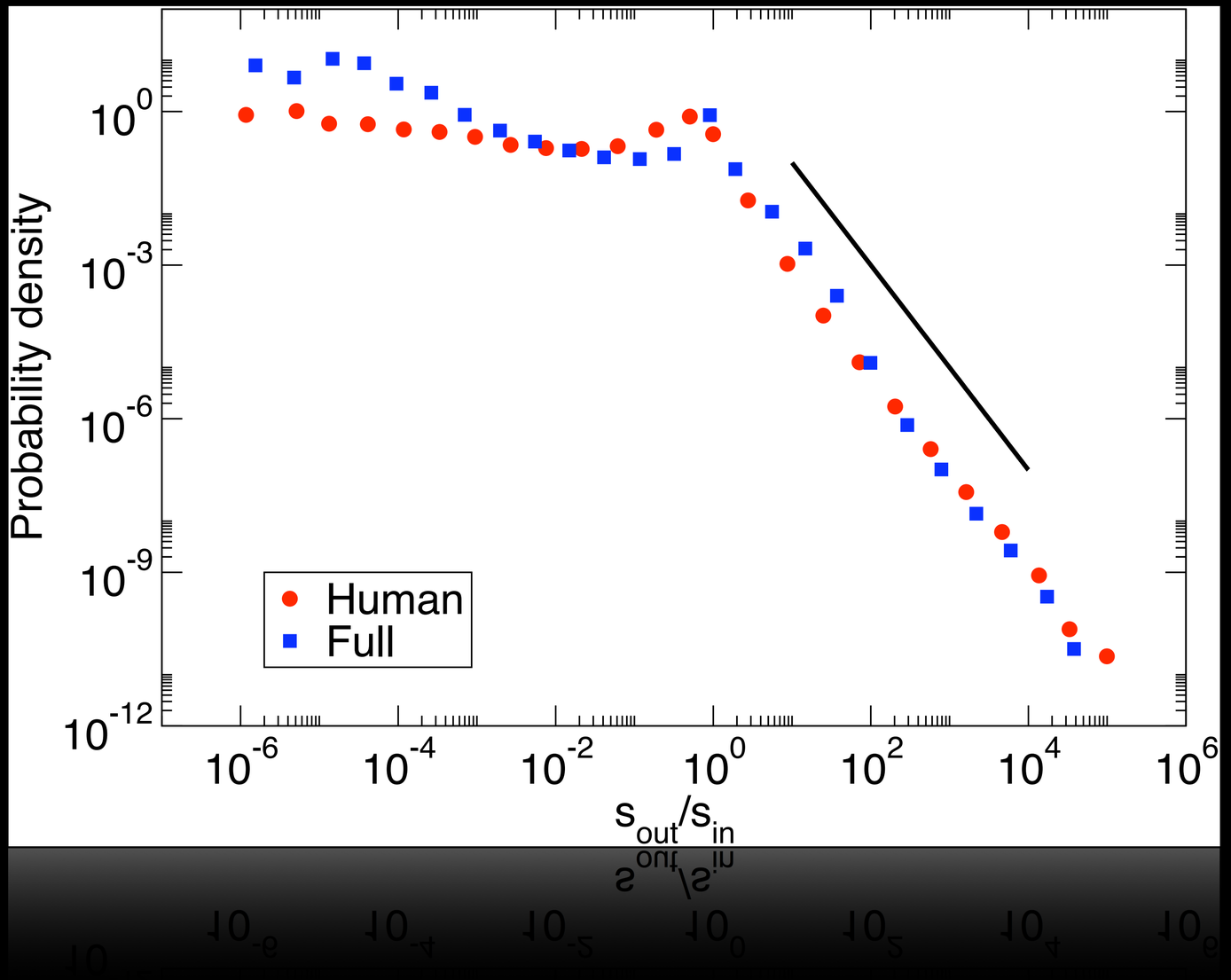
Jump Traffic



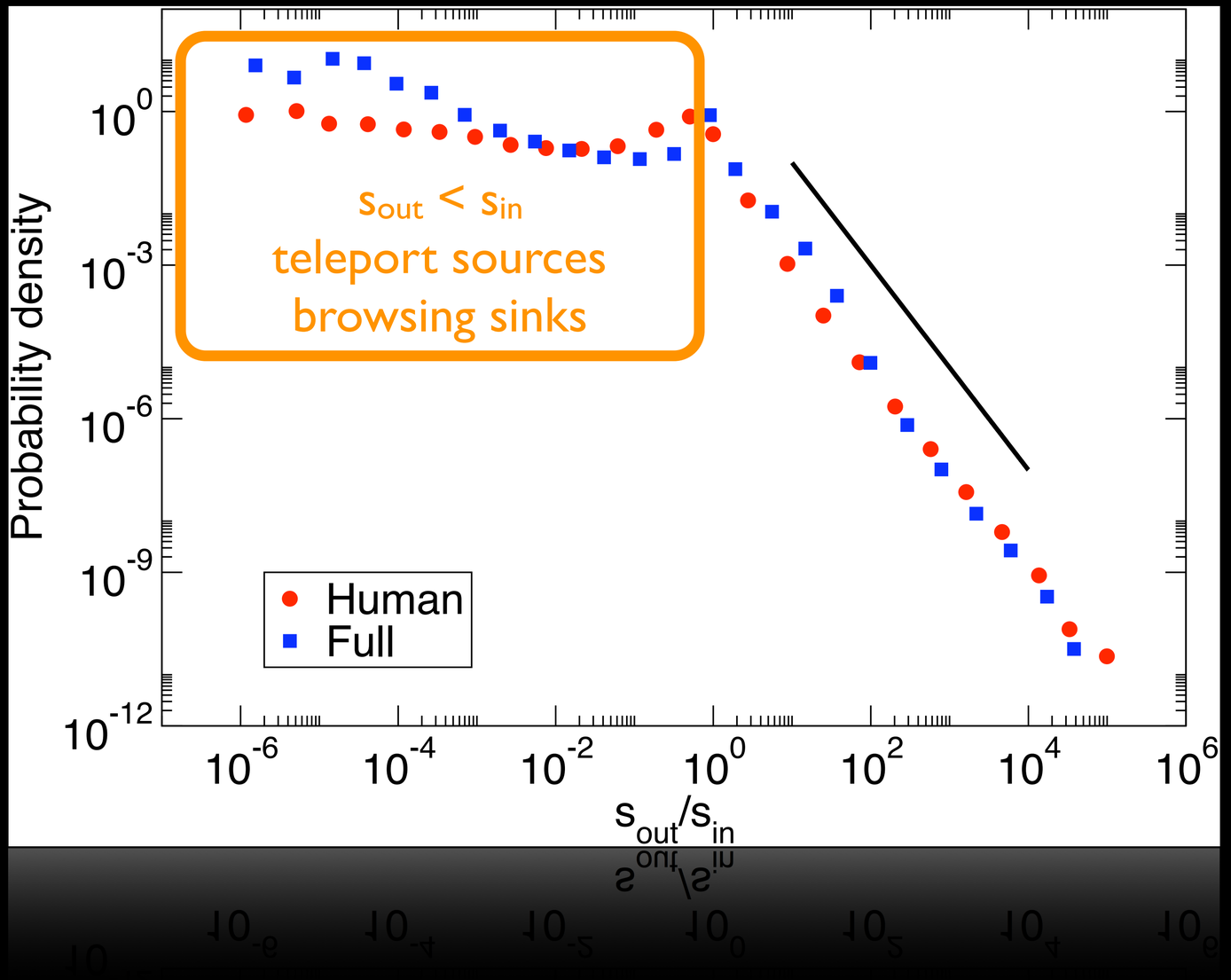


Link Traffic

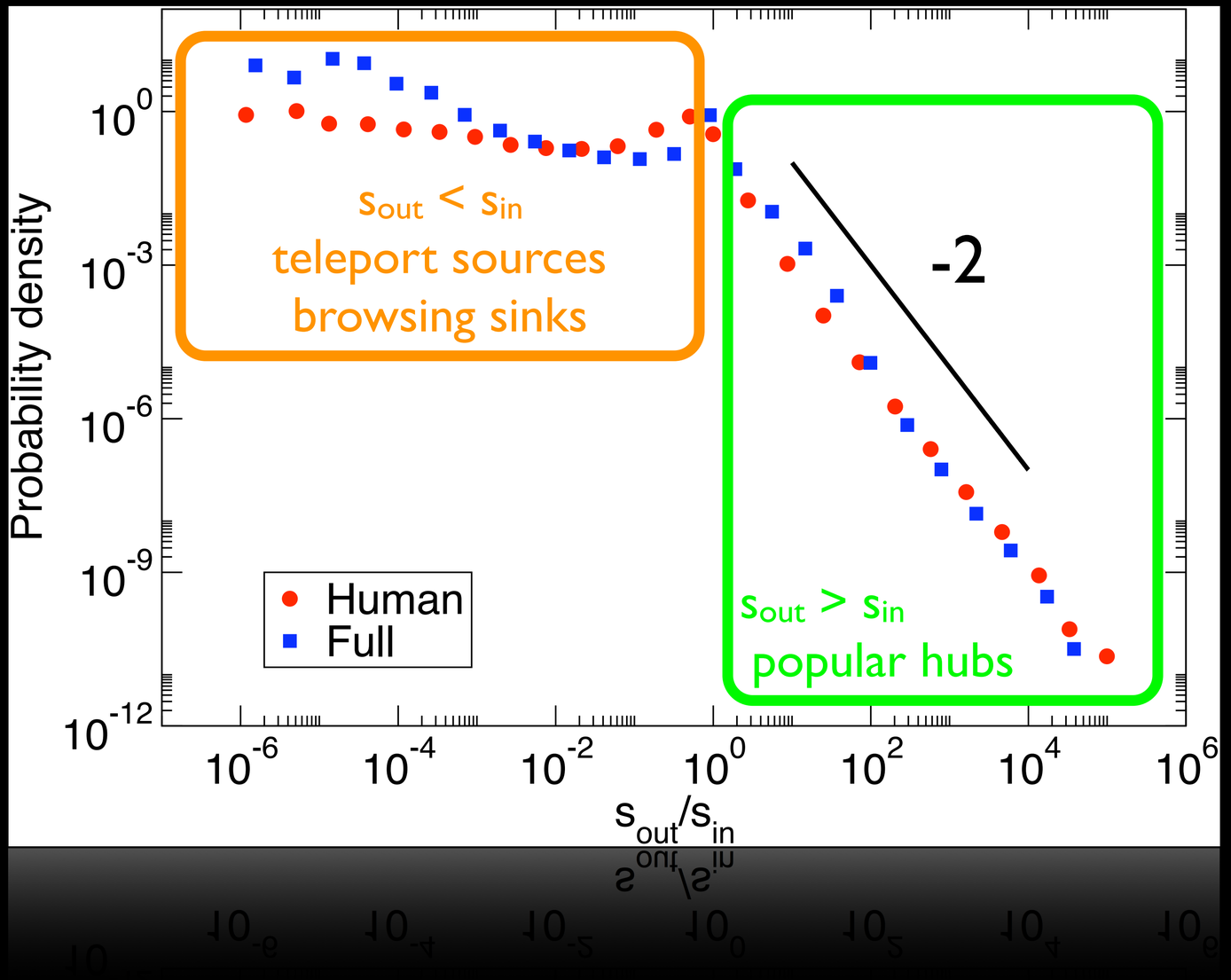
Teleportation source heterogeneity: “hubness”



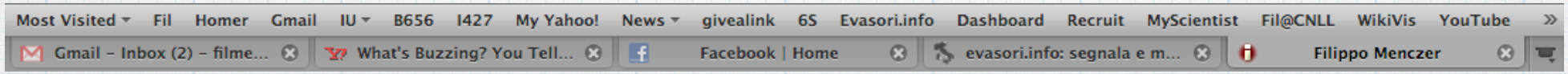
Teleportation source heterogeneity: “hubness”



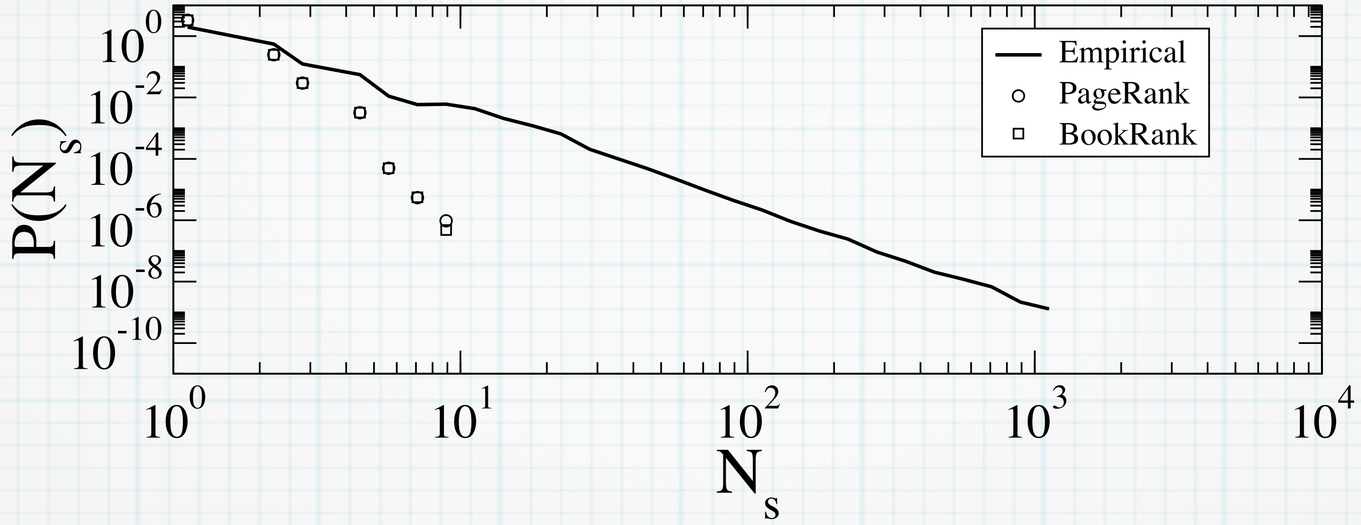
Teleportation source heterogeneity: “hubness”

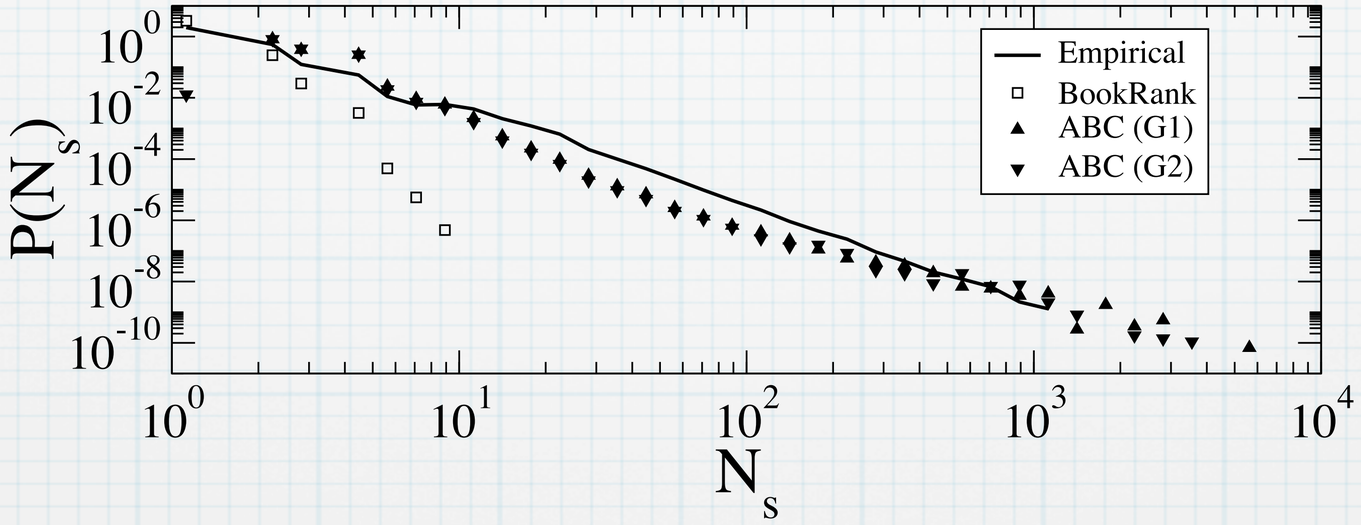


Ingredients for an agent-based model



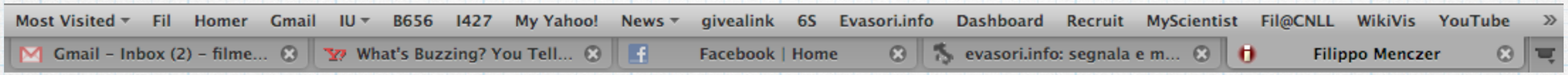
1. Bookmarks (memory)
2. Backtracking (or tabs)
3. Topicality (sessions)





Session length (& depth)

Summary

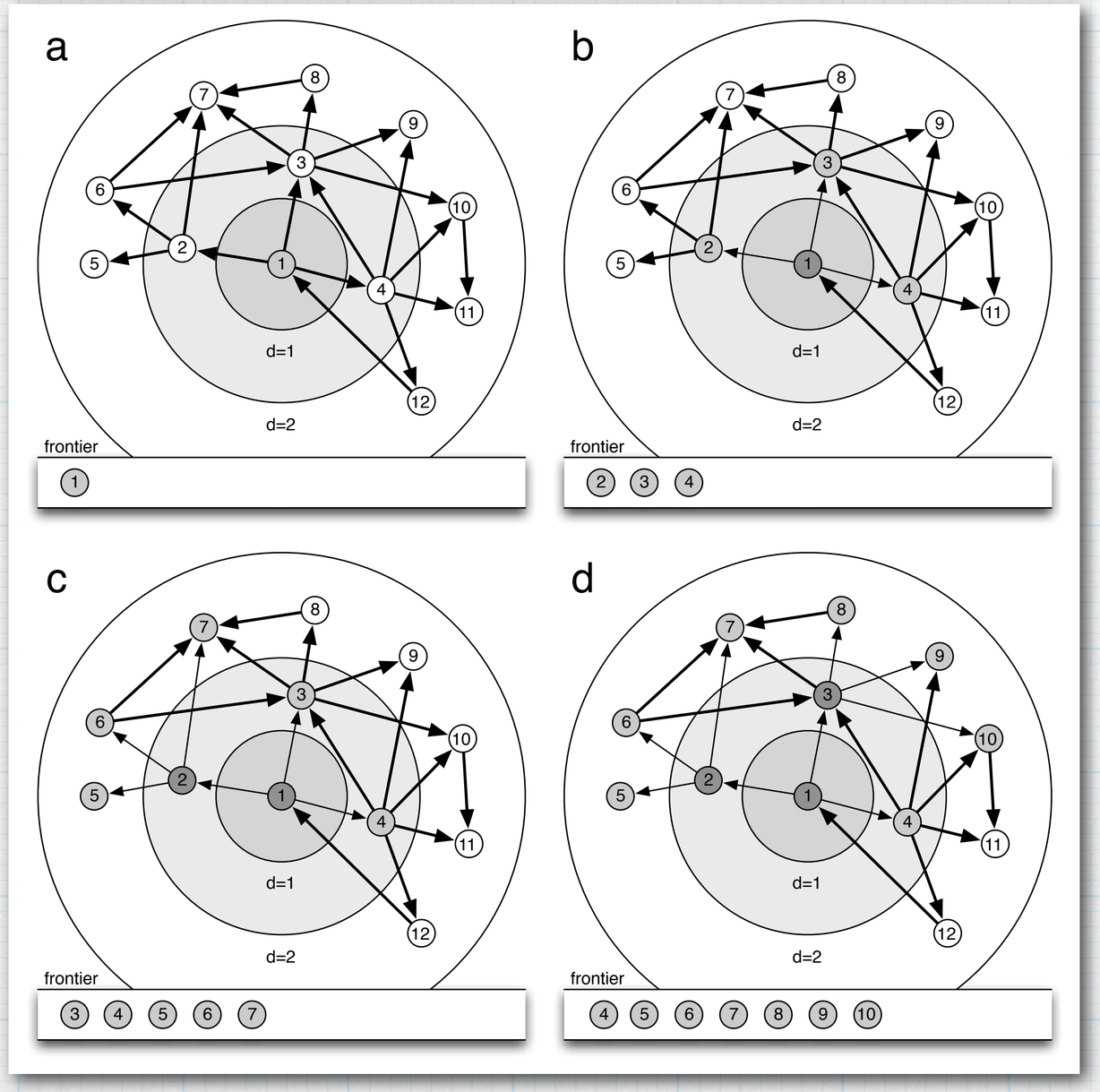


* Memory/bookmarks allow to reproduce page, jump, and link traffic
* To capture full user browsing behavior, including session length, the model needs to also account for backtracking and topicality

- 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 Erdös 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?

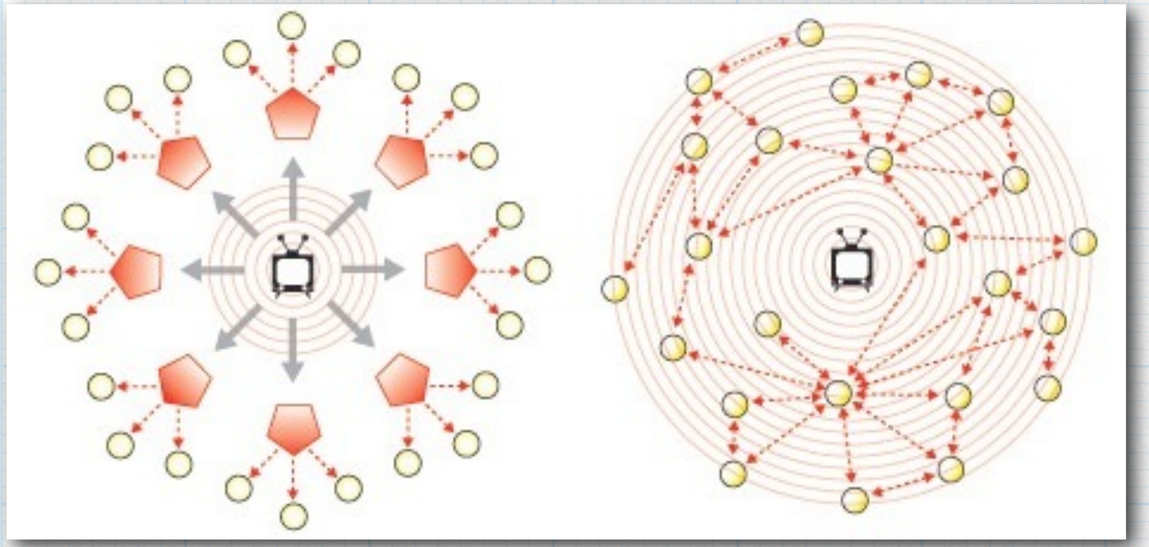


Remember BFS...

* 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.

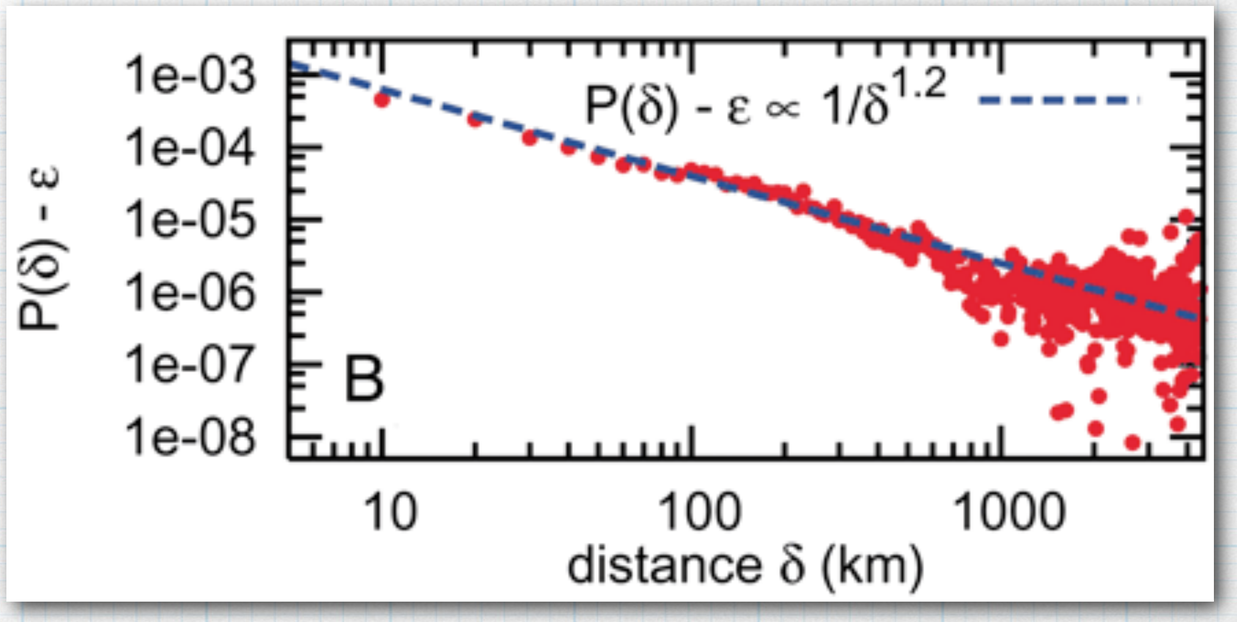
Practice: hard to find them!

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

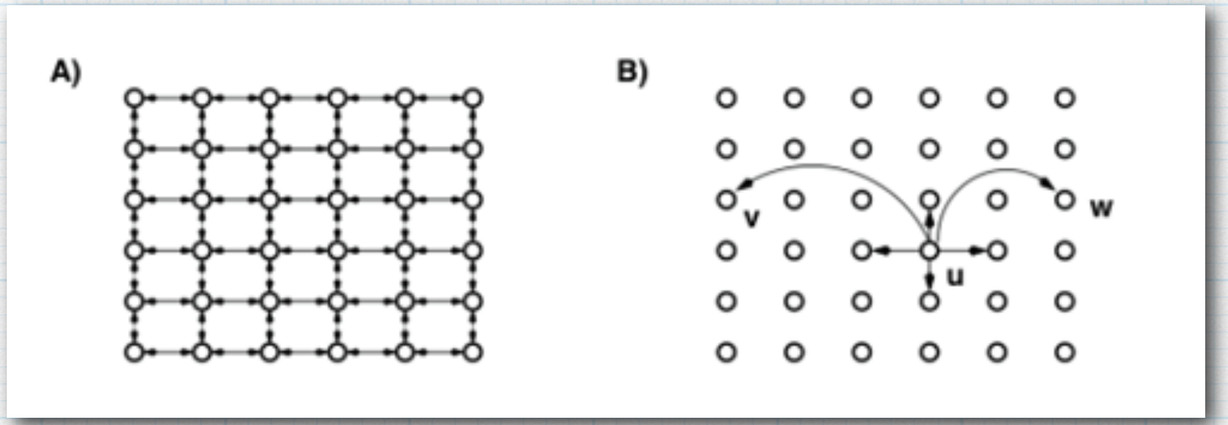


Liben-Nowell et al. 2005 10.1073/pnas.0503018102

Locality and geography

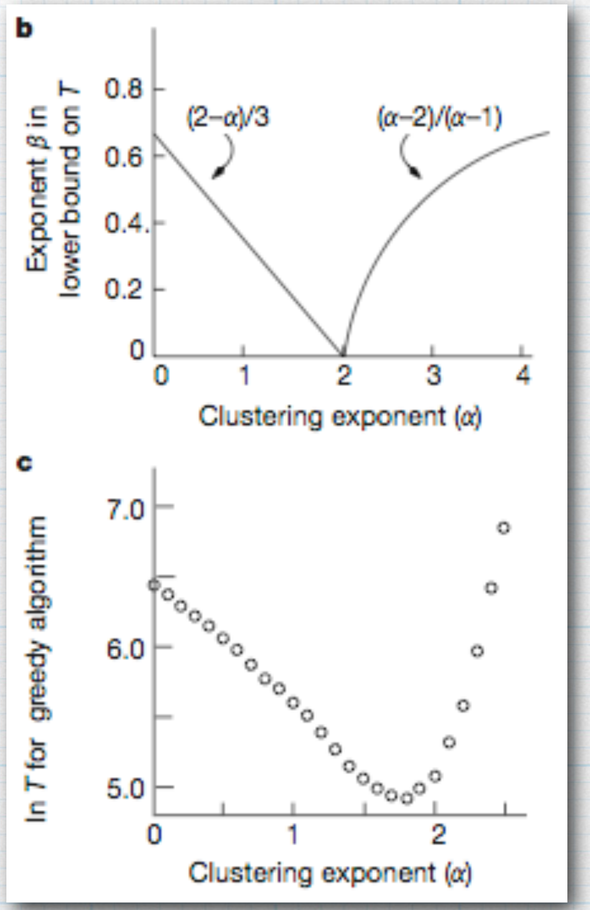
* Jon Kleinberg (2000): Assume we have information about position
  * Local links to all lattice neighbors
  * Long-range link probability distribution:

    power law $Pr(r) \sim r^{-\alpha}$
    * r: lattice (Manhattan) distance
    * α: clustering exponent



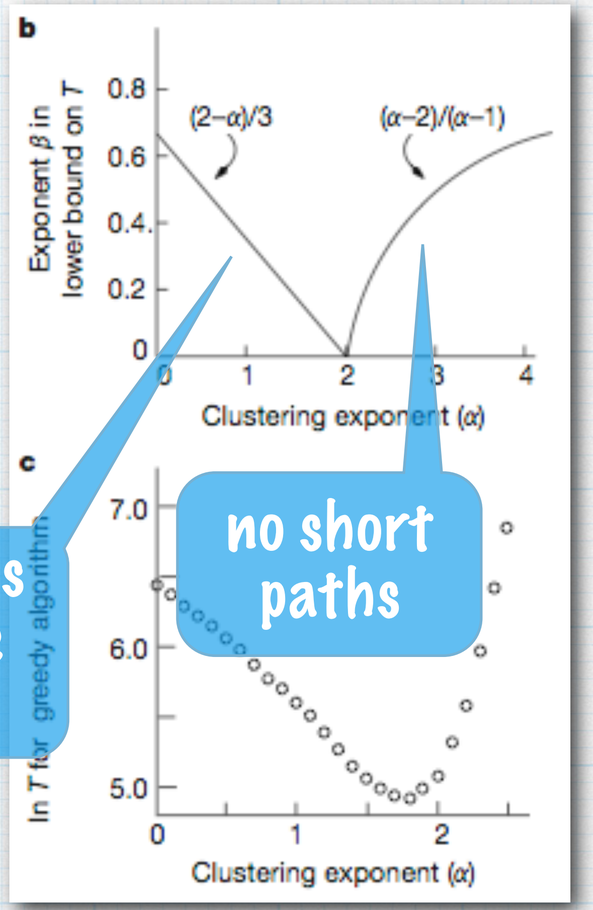
Locality and geography

* 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 $L \sim N^{\beta}$



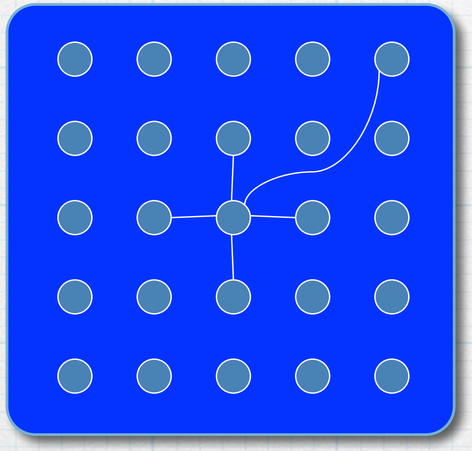
Locality and geography

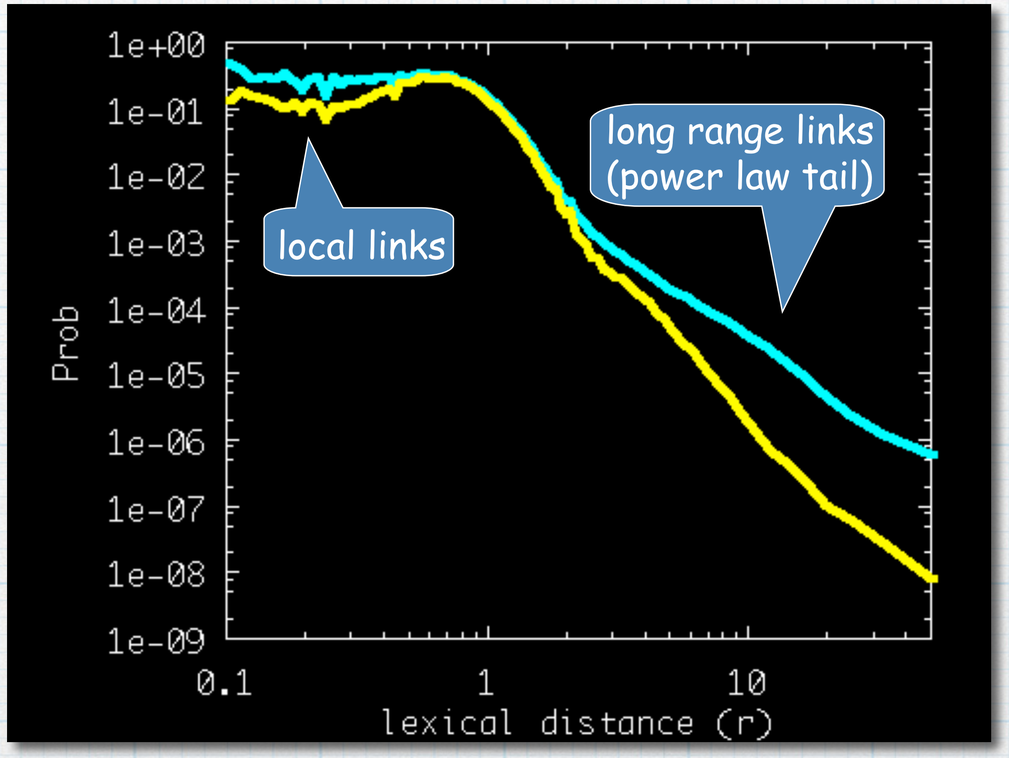
* Jon Kleinberg (2000): Assume we have information about position/location/ distance
* Greedy algorithms: pick neighbor node closest to target
  * $L \sim \log^2 N$ if $\alpha=2$ (critical exponent = dimension of lattice)
  * Else $L \sim N^{\beta}$



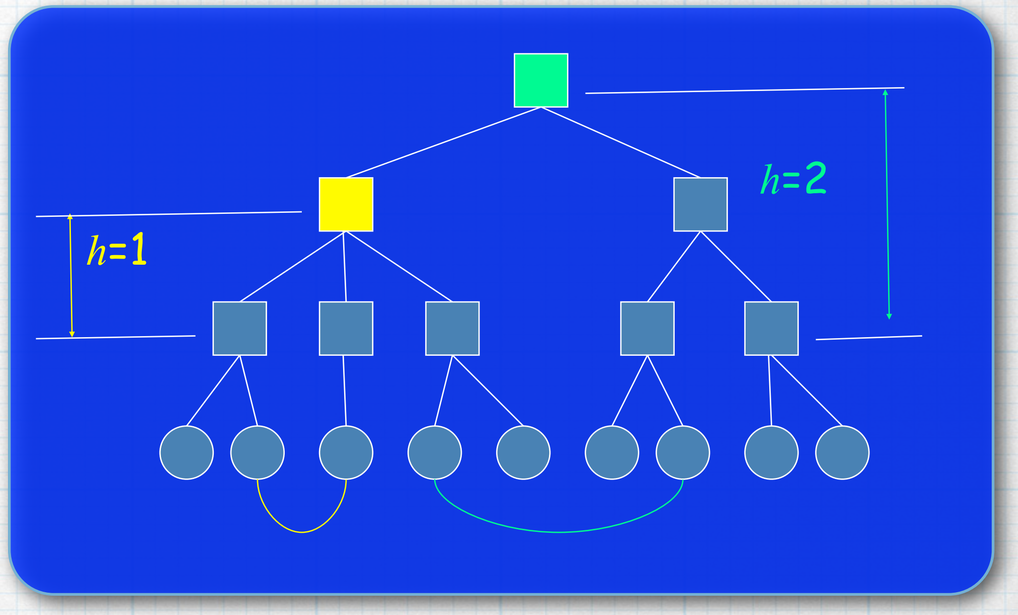
Topical Locality & Searchability

* 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

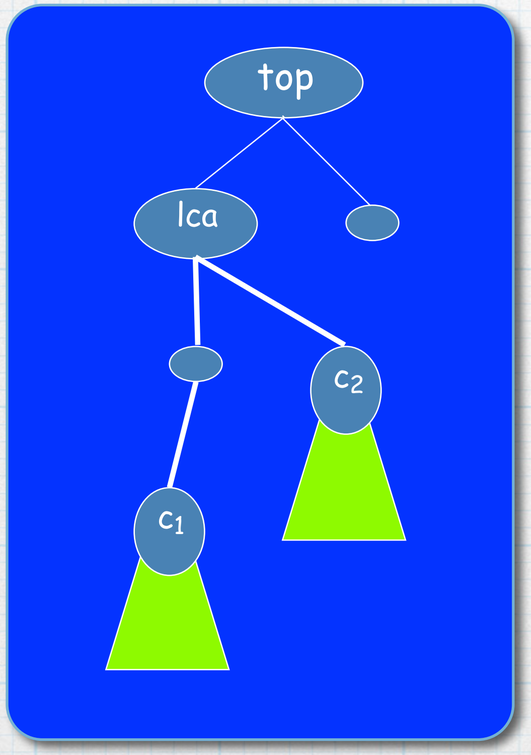


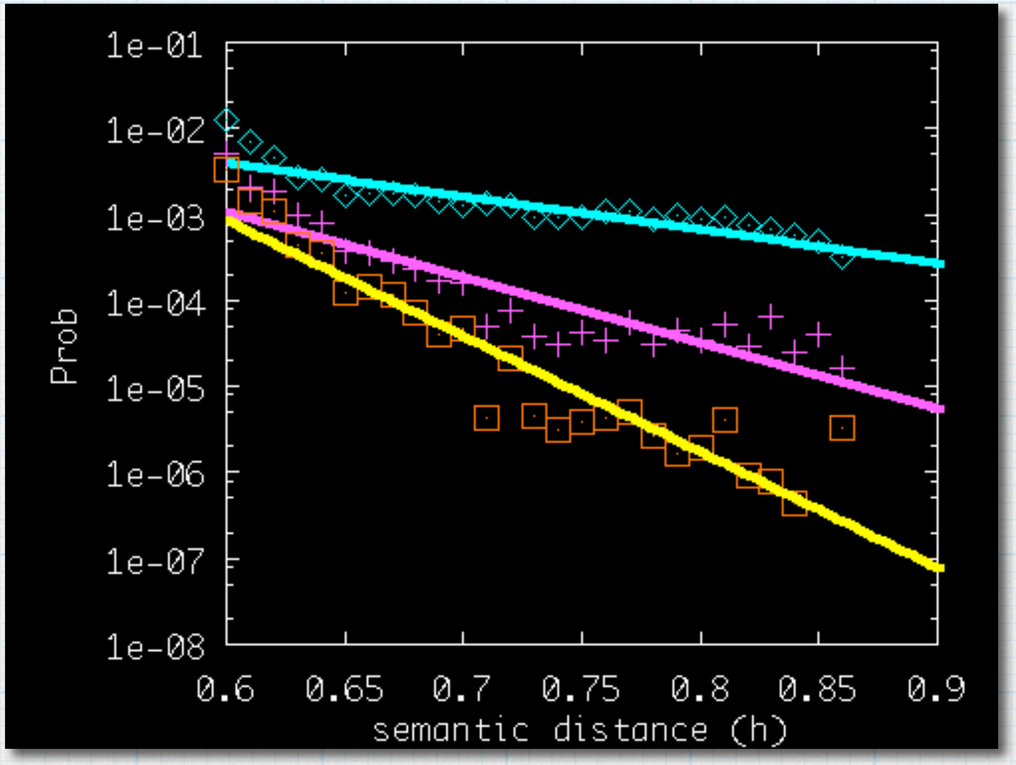
- Kleinberg (2002), Watts & al. (2002)
- Nodes are classified at the leaves of tree
  - Eg, by topical interests, profession, etc.
  - Link probability distribution: exponential $Pr(h) \sim e^{-h}$
    - h: tree distance (height of lowest common ancestor)

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

Topical Hierarchy & Searchability

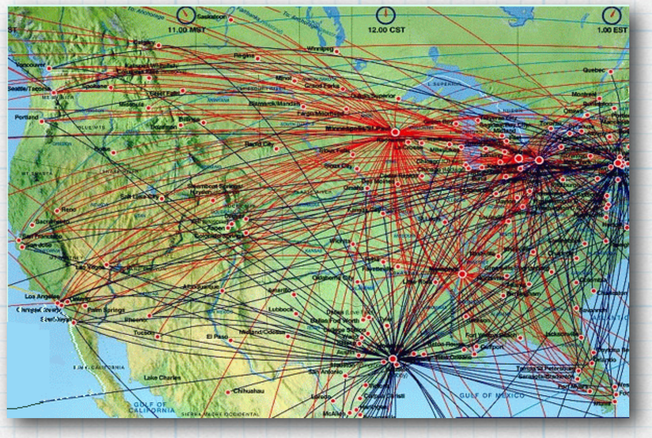
* It turns out that the Web is a special case of "searchable" (hierarchical) network





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



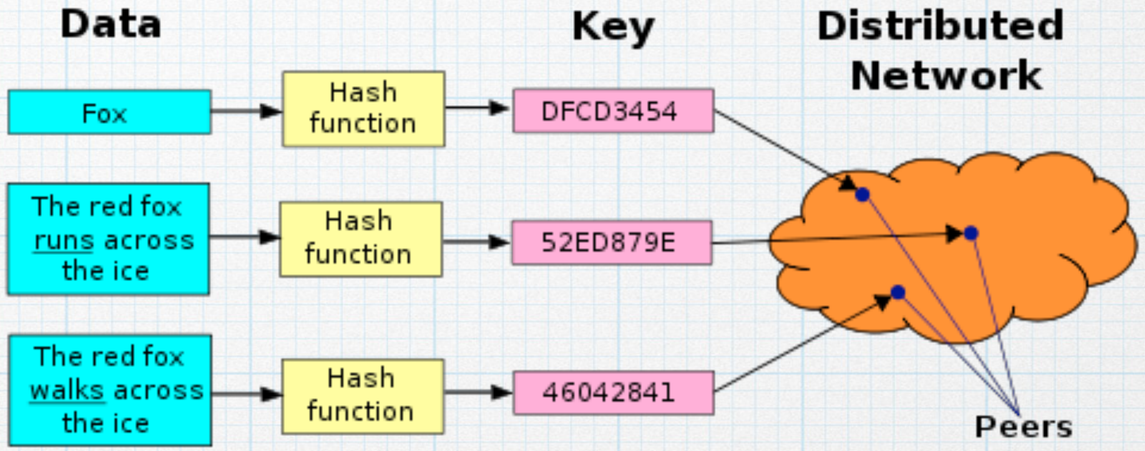
* Not good to search the Web at query time: too many links to explore! But crawlers implicitly leverage hubs, then information reused many times by search engines

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

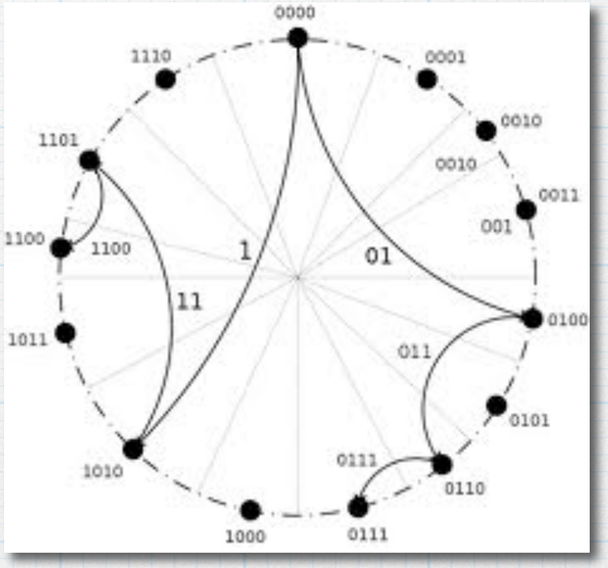
* Structured networks, distributed hash tables (DHT)



* 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