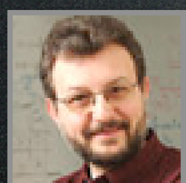
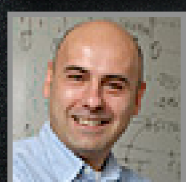
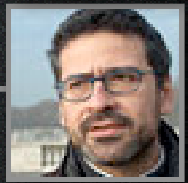
Googlearchy or Googlocracy?

How search affects Web traffic and growth


Filippo Menczer

Santo Fortunato

Sandro Flammini

Alex Vespignani

Indiana University School of informatics

Web [Images](#) [Groups](#) [News](#) [Froogle](#) [Local](#)New! [Scholar](#) **[more »](#)**

global warming | Search | [Advanced Search](#) [Preferences](#)

Web Results **1** - **100** of about **282,000** for **global warming**

[The **Global Warming** hoax by James K. Glassman -- Capitalism Magazine](#)
The delegation met Wednesday with counterparts from Europe, and Inhofe and many of his colleagues were shocked at the Europeans' refusal even to consider ...
capmag.com/article.asp?ID=3400 - 19k - Cached - Similar pages

[Capitalism Magazine: The **Global Warming** hoax by James K. Glassman](#)
The delegation met Wednesday with counterparts from Europe, and Inhofe and many of his colleagues were shocked at the Europeans' refusal even to consider ...
capmag.com/articlePrint.asp?ID=3400 - 13k - Cached - Similar pages
[More results from capmag.com]

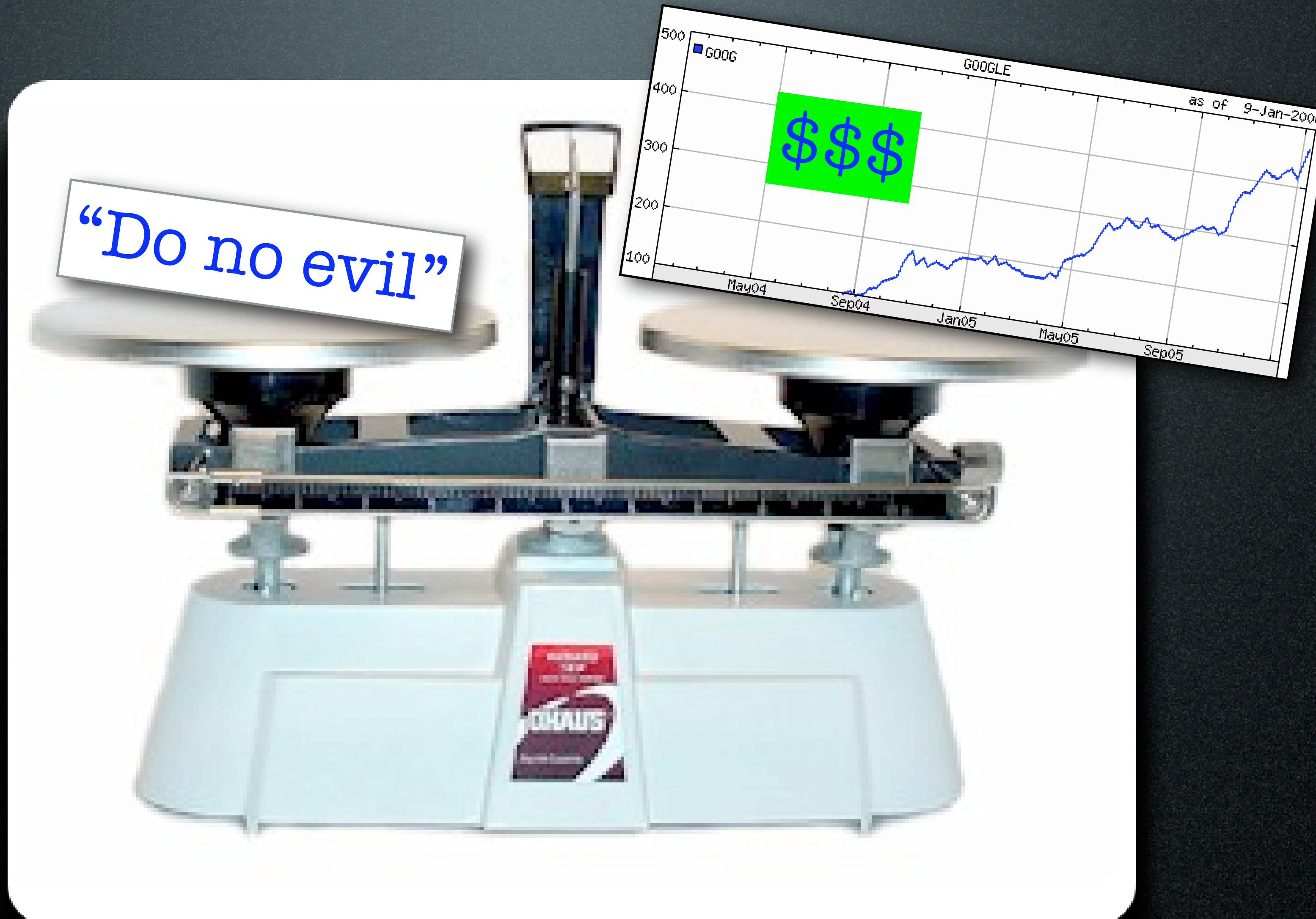
[No Scientific Consensus On **Global Warming**](#)
The United Nations Intergovernmental Panel on Climate Change (IPCC) changed or deleted more than 15 sections in Chapter 8 of the report --sections setting ...
www.zianet.com/wblase/endtimes/gwarm.htm - 15k - Cached - Similar pages

[ESR | June 9, 2003 | Revisiting the **global warming** hoax](#)
The entire **global warming** hoax is based on computer models and they are designed to produce ... The **global warming** hoax is not about the Earth's climate. ...
www.enterstageright.com/ archive/articles/0603/0603**warming**.htm - 11k - Cached - Similar pages

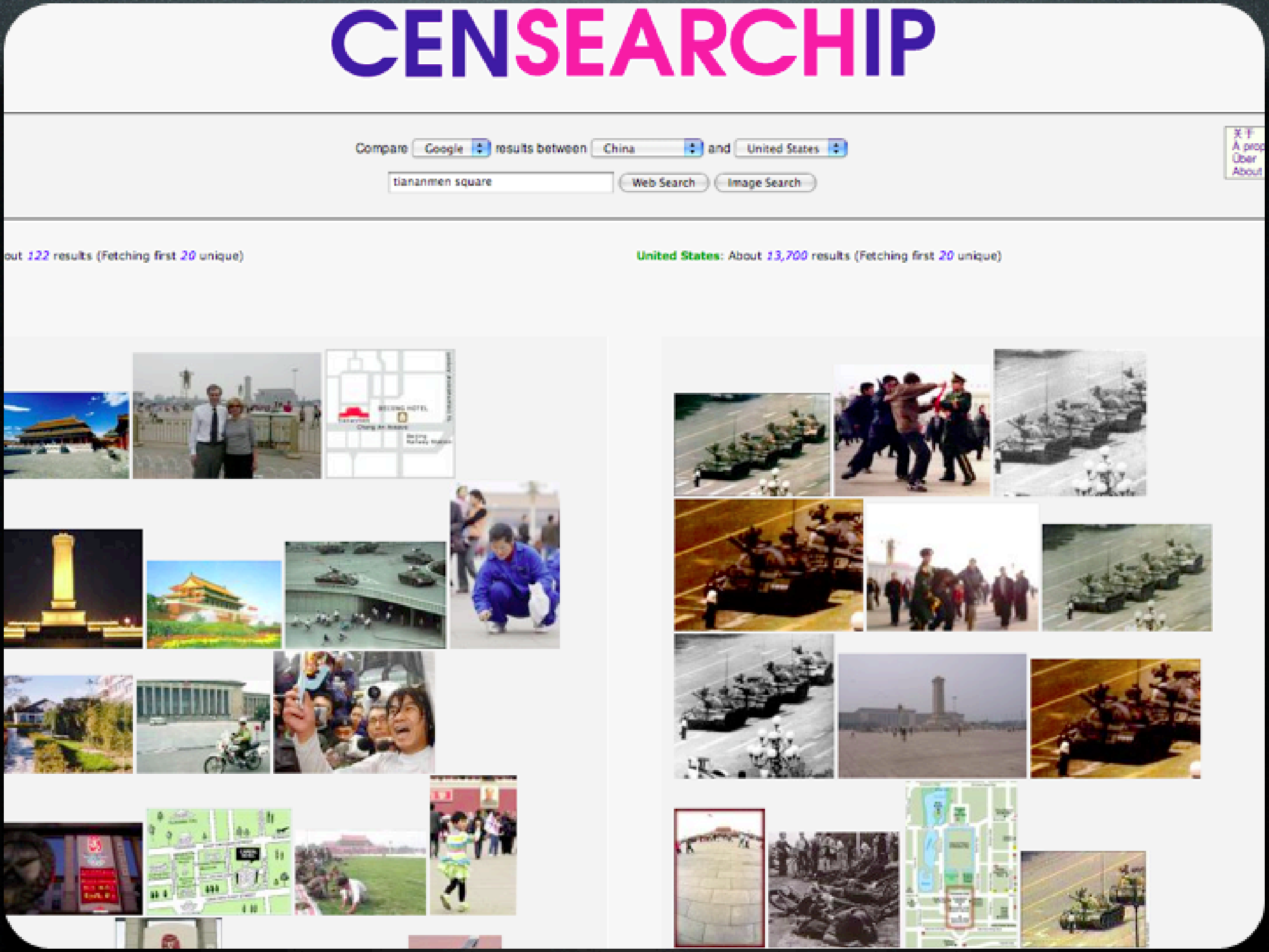
[Archive | November 13, 2000 | Desperate times call for desperate acts](#)
The Greens **global warming** hoax, the cause that followed the Ice Age fiasco, ... Vice President Gore's star witness for the **global warming** hoax, ...
www.enterstageright.com/archive/ articles/1100environmentalism.htm - 11k - Cached - Similar pages
[More results from www.enterstageright.com]

["Man-Made **Global Warming** hoax" by Tom Gremillion](#)
Global warming is a hoax, invented in 1988, that combines old myths including limits to growth, sustainability, the population growth time bomb, ...
www.chronwatch.com/content/contentDisplay.asp?aid=12594 - 36k - Cached - Similar pages

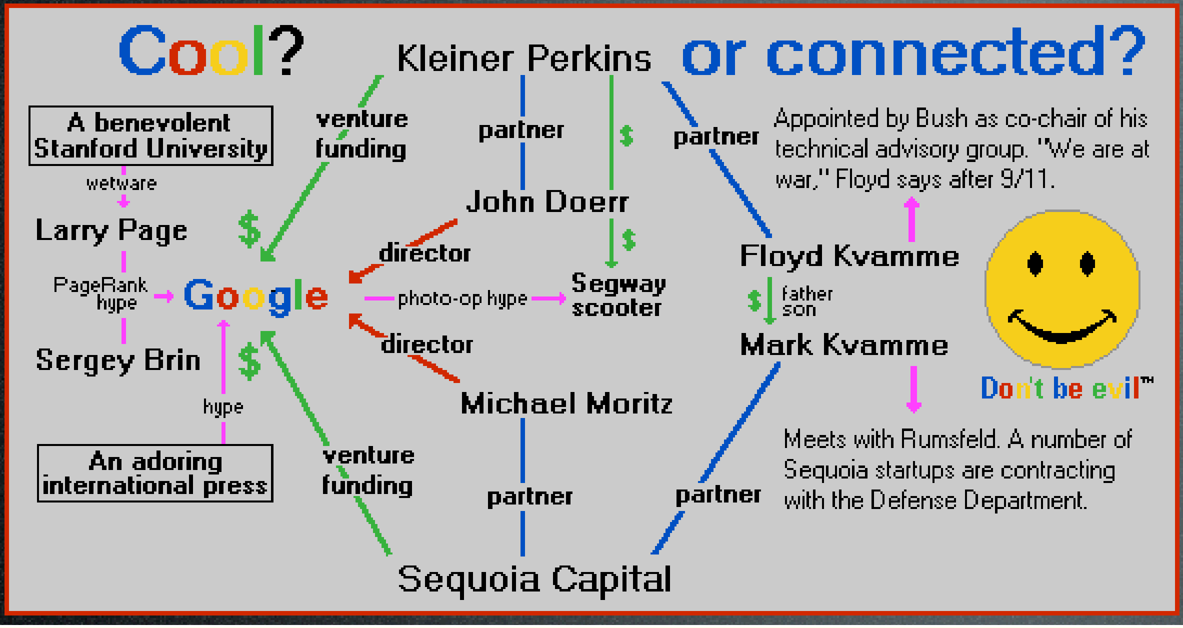
[**Global Warming** Is Greatest hoax Ever -- America's Future -- Week ...](#)
DeWeese calls **global warming** "the greatest hoax ever perpetrated on the people of the world, bar none. Those who have been fighting against the green agenda ...
www.americasfuture.net/1997/nov97/97-1123a.html - 9k - Cached - Similar pages

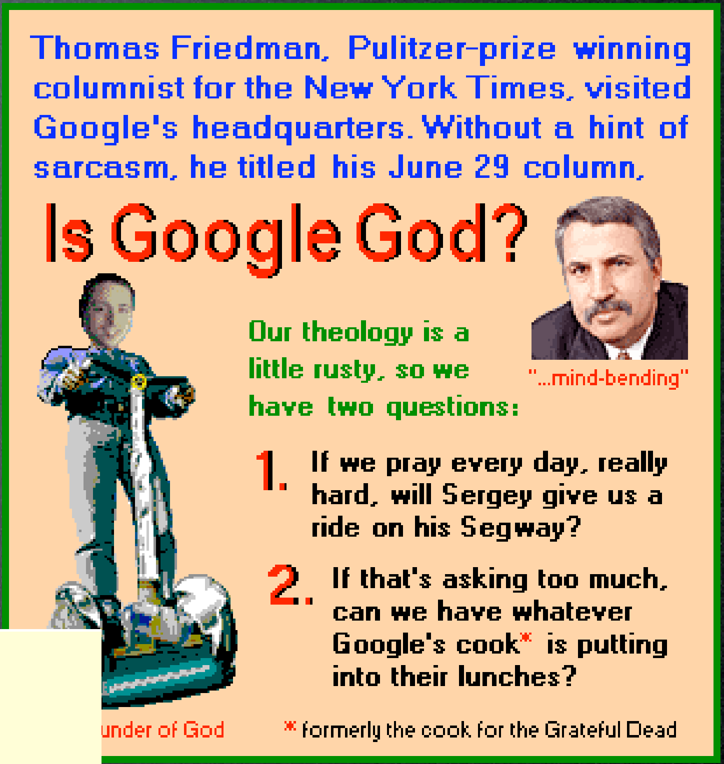


Corporate bias?



homer.informatics.indiana.edu/censearchip





GoogleWatch

Searching on the phrase "troubled teen"

Total number of results	448,000
Number of listings per page, total	20
Sponsored listings per page	10
Non-sponsored listings per page	10
Percent of first page, non-sponsored results that are commercial	80%
Percent of the top 100 results that are noncommercial	15%
Number of valuable news articles in the first 100 results	13
Number of commercial sites with multiple listings in top 100	16

COVER FEATURE

Definin

The Po

Helen Nissenbaum
Princeton University

at the ex
mated th
engines
total ind
search e

first monday

PEER-REVIEWED JOURNAL ON THE INTERNET

by Susan L. Gerhart

Do Web search engines suppress controversy?

Available online at www.sciencedirect.com

SCIENCE DIRECT

ELSEVIER

Information Processing and Management 41 (2005) 1193–1205

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Measuring search engine bias

Abbe Mowshowitz *, Akira Kawaguchi [1]

Department of Computer Science, The City College of New York, Convent Avenue at 138th Street, New York, NY 10031, USA

S ON PRODUCT INFORMATION

BIAS ON THE WEB

When it comes to measuring bias on the Web, there is clearly strength in numbers (of search engines, that is).

among those retrieved.
an be illustrated by examining the
d names in the URL strings. Con-

"Googlearchy": How a Few Heavily-Linked Sites Dominate Politics on the Web*

Matthew Hindman[†] Kostas Tsioutsiouliklis[‡] Judy A. Johnson[§]

March 31, 2003

Photo by Hector Garcia-Molina

Junghoo "John" Cho
Assistant Professor

Department of Computer Science
University of California, Los Angeles

Office: Boelter Hall 3532E
Email: cho@cs.ucla.edu
Phone: (208) 979-5012
Fax: (208) 979-5012

Search-Engine Bias Project

Search engines are a big part of our everyday life. Most of us rely on search engines to discover and access contents from the Web. Does this mean that now we can be biased by what search engines process and present to us? What kind of and how much bias can search engines potentially introduce? The primary goal of this research project is to investigate the potential bias of search engines problem and come up with technical solutions to this problem.

Publications

1. Junghoo Cho, Sourashis Roy "Impact of Web Search Engines on Page Popularity." *In Proceedings of the World-Wide Web Conference (WWW)*, May 2004.

2. Feng Qiu, Zhenyu Liu, Junghoo Cho "Analysis of User Web Traffic with a Focus on Search Activities." *In Proceedings of the International Workshop on the Web and Databases (WebDB)*, June 2005.

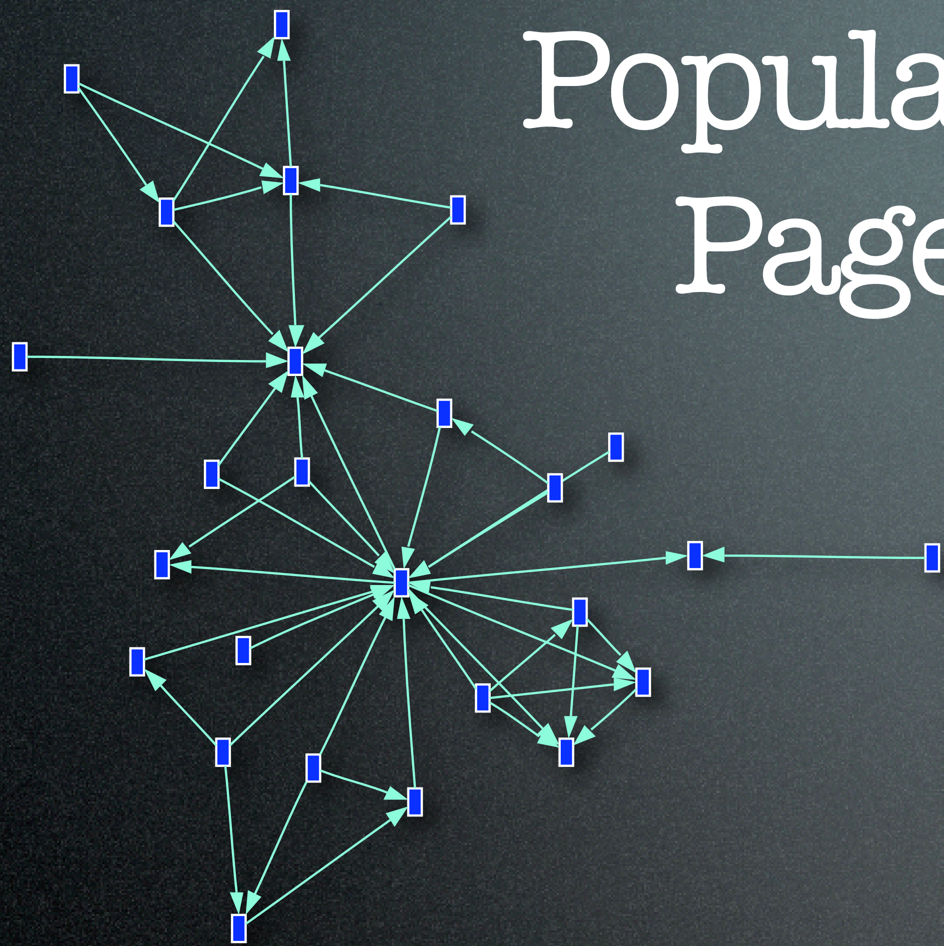
3. Junghoo Cho, Sourashis Roy, Robert E. Adams "Page Quality: In Search of an Unbiased Web Ranking." *In Proceedings of 2005 ACM International Conference on Management of Data (SIGMOD)*, May 2005.

4. Sandeep Pandey, Sourashis Roy, Christopher Olston, Junghoo Cho, Soumen Chakrabarti "Shuffling a Stacked Deck: The Case for Partially Randomized Ranking of Search Engine Results" *In Proceedings of 31st International Conference on Very Large Databases (VLDB)*, September 2005.

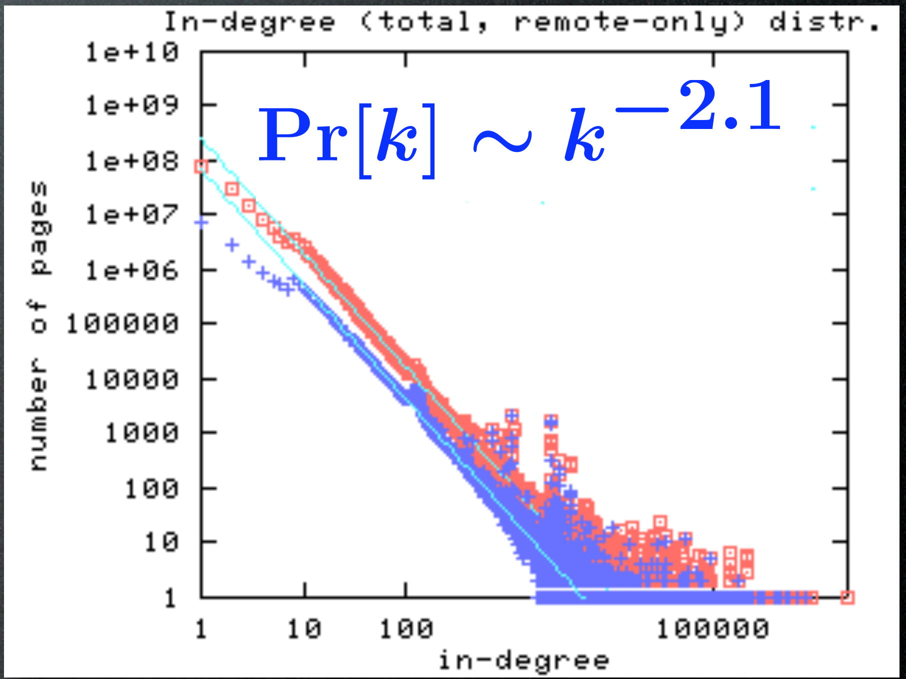
Questions

- Are we witnessing a monopolization of the Web by an oligarchy of sites?
- Can we quantify popularity bias from empirical evidence?
- Can we predict popularity bias with a simple model of searching?

Popularity and PageRank



“long tail”
“scale-free”
“rich-get-richer”



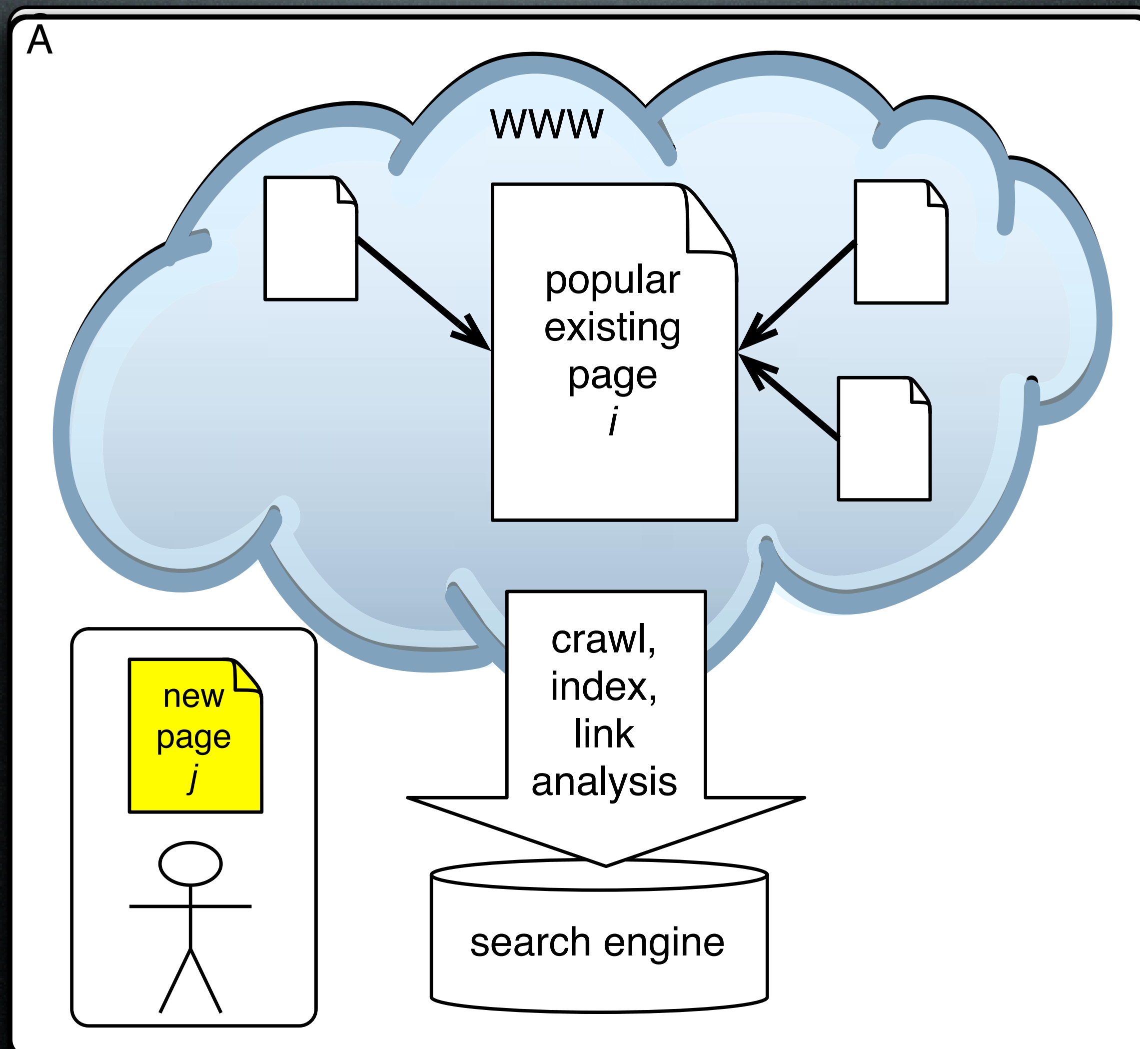
Broder & al. 2000

PageRank

$$p(i) = \frac{\alpha}{N} + (1-\alpha) \sum_{j:j \to i} \frac{p(j)}{|\ell : j \to \ell|}$$

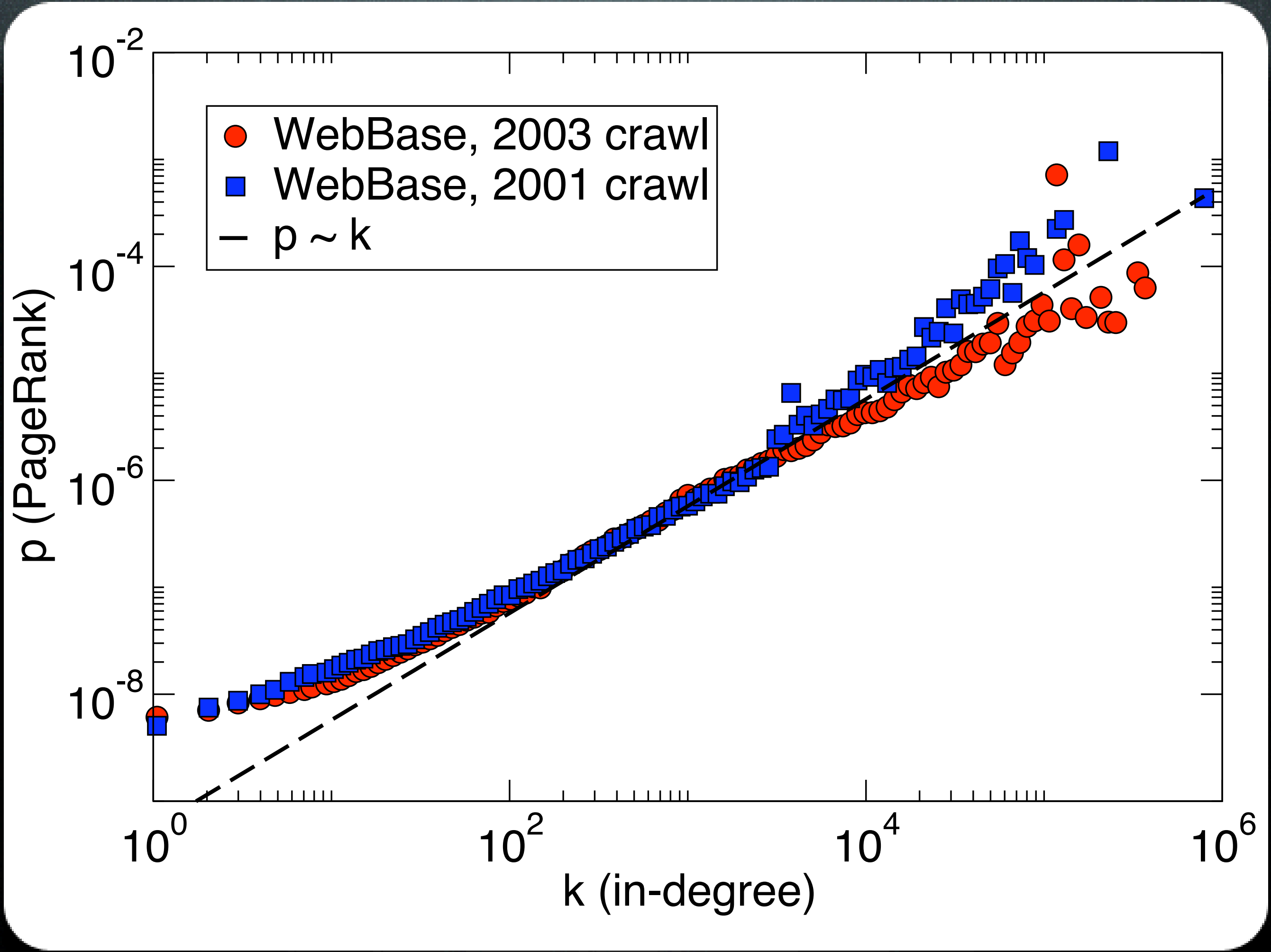
Brin & Page 1998

Popularity Bias
(“Entrenchment”, “Googlearchy”)

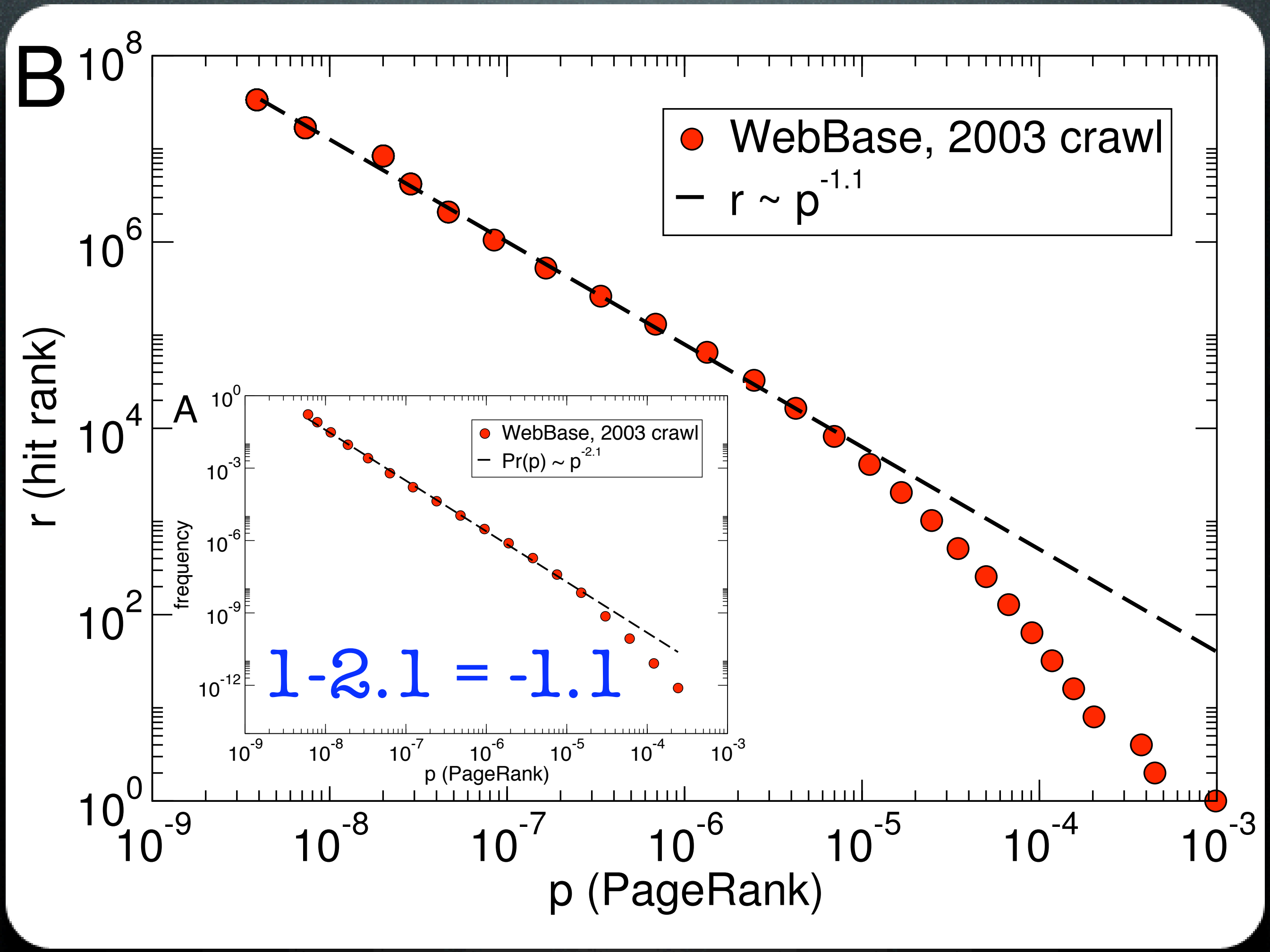


Modeling search engine bias from the relationship between indegree and traffic

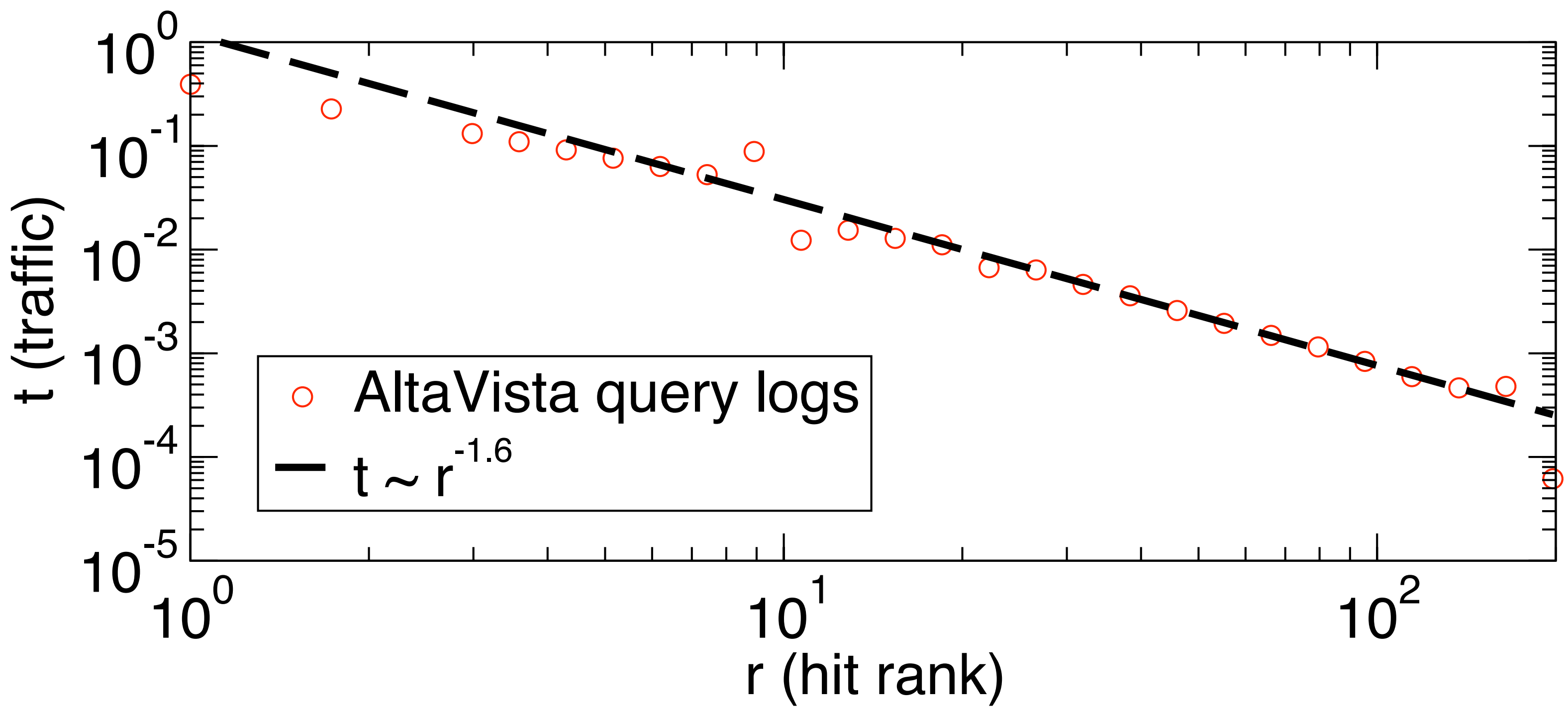
1. traffic ~ P(click)
2. P(click) ~ f(rank)
3. rank ~ f(PageRank)
4. PageRank ~ f(indegree)



4. PageRank(indegree)



3. rank(PageRank)



2. P(click|rank)

Chaining together the scaling relationships

$$t \sim r^{-1.6}$$

$$\sim (p^{-1.1})^{-1.6}$$

Without search (surfing only)

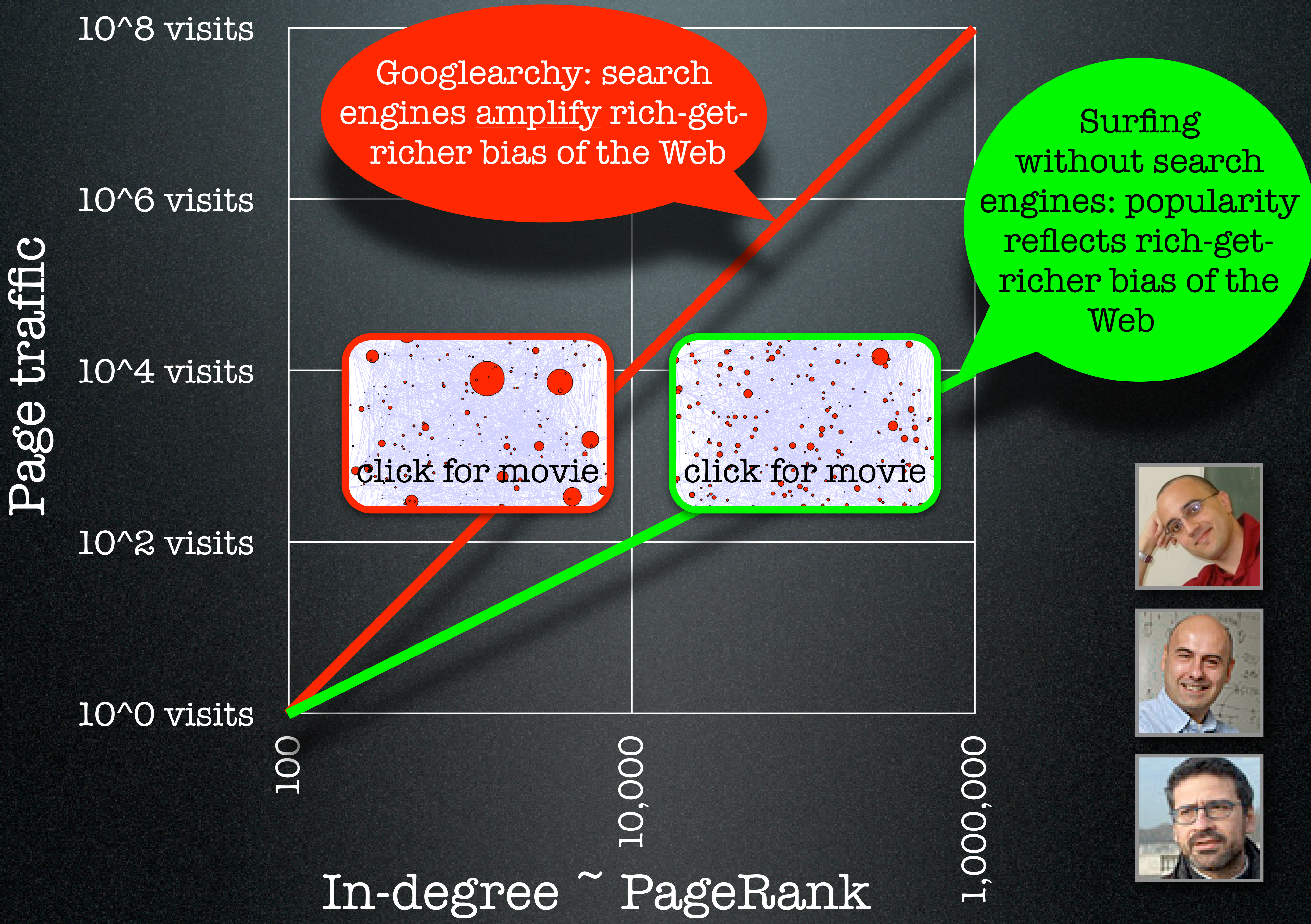
$$\sim (k^{-1.1})^{-1.6}$$

$$t \sim p \sim k$$

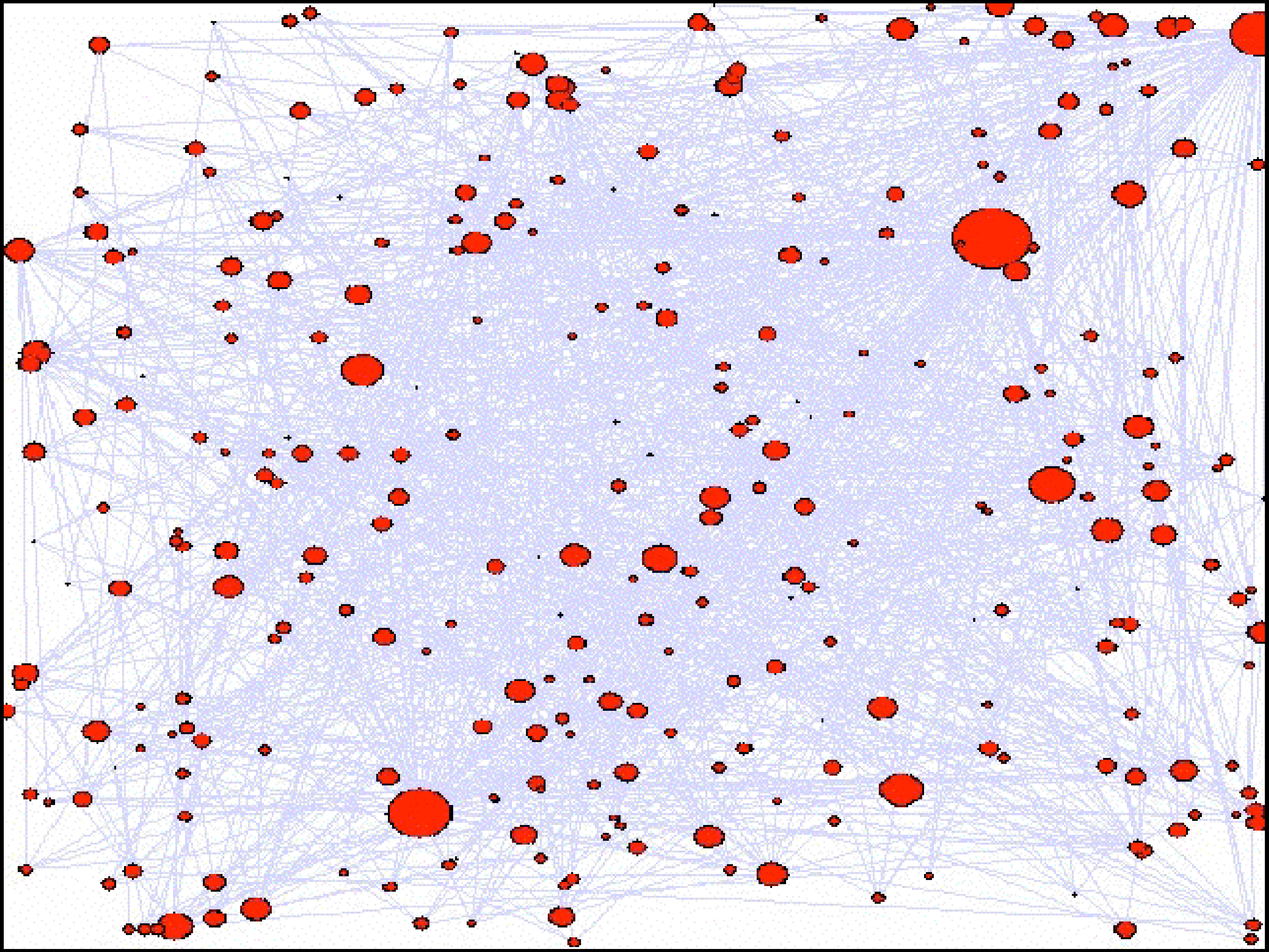
$$\sim k^{1.6 \times 1.1}$$

Googlearchy

$$\sim k^{1.8}$$

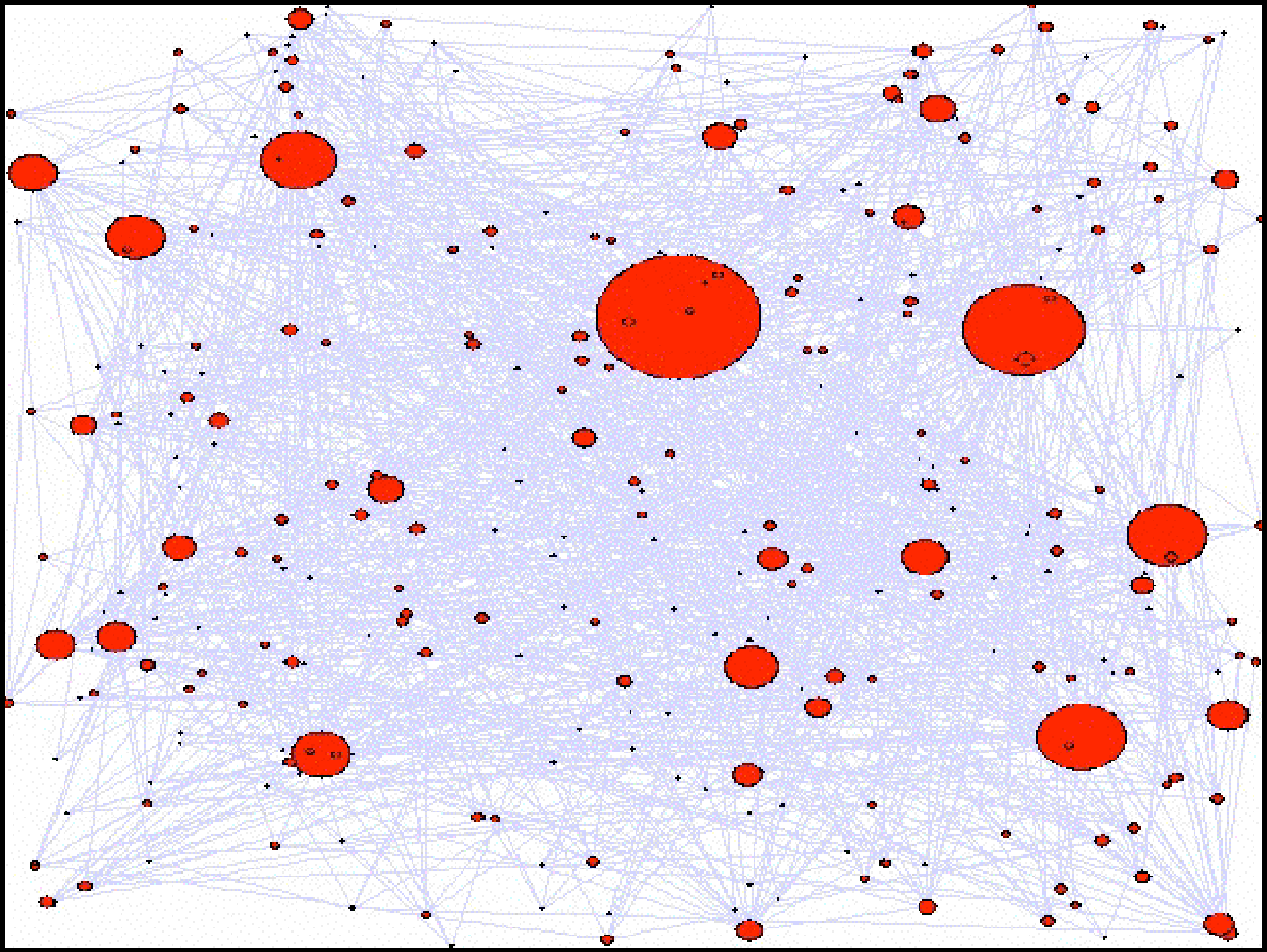

10^8 visits
10^6 visits
10^4 visits
10^2 visits
10^0 visits
Page traffic
Googlearchy: search engines amplify rich-get-richer bias of the Web
Surfing without search engines: popularity reflects rich-get-richer bias of the Web
click for movie
click for movie
100
10,000
1,000,000
In-degree ~ PageRank

Surfing without search engines: popularity reflects rich-get-richer bias of the Web



back

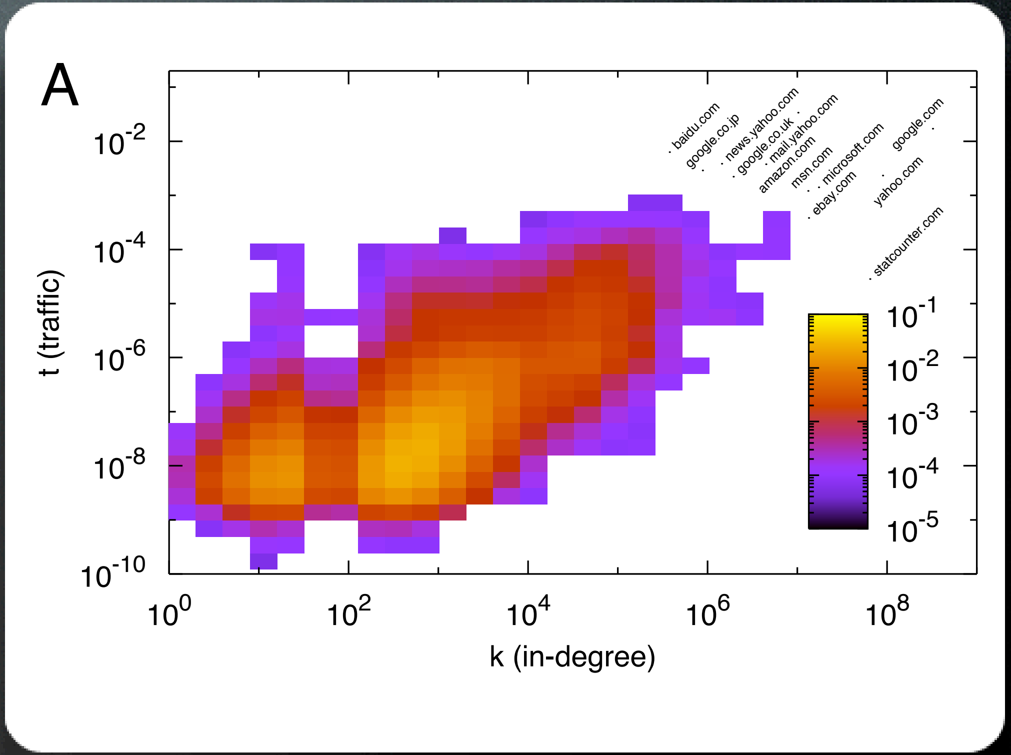
Googlearchy: search engines amplify rich-get-richer bias of the Web

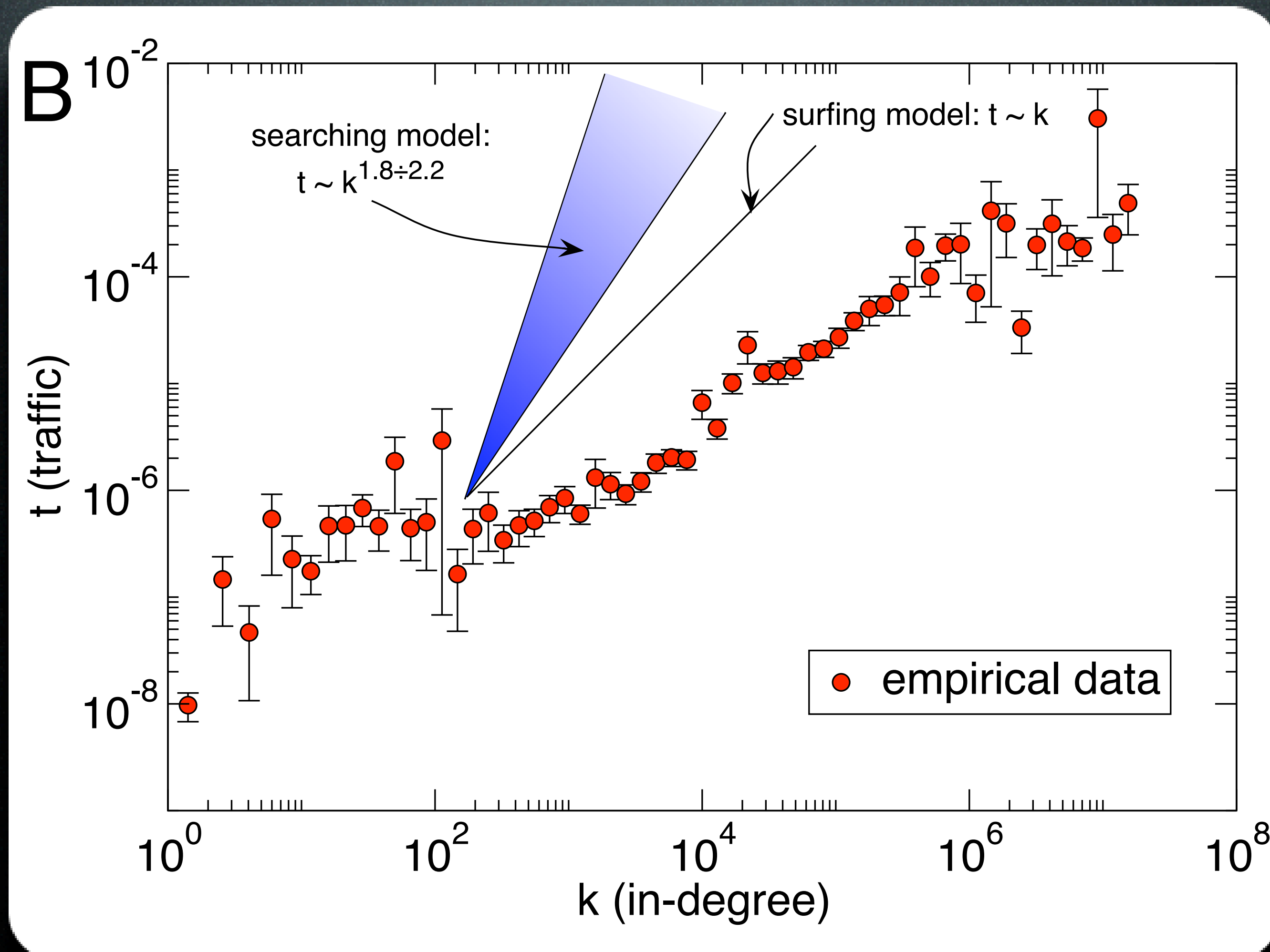


back

Empirical measurements

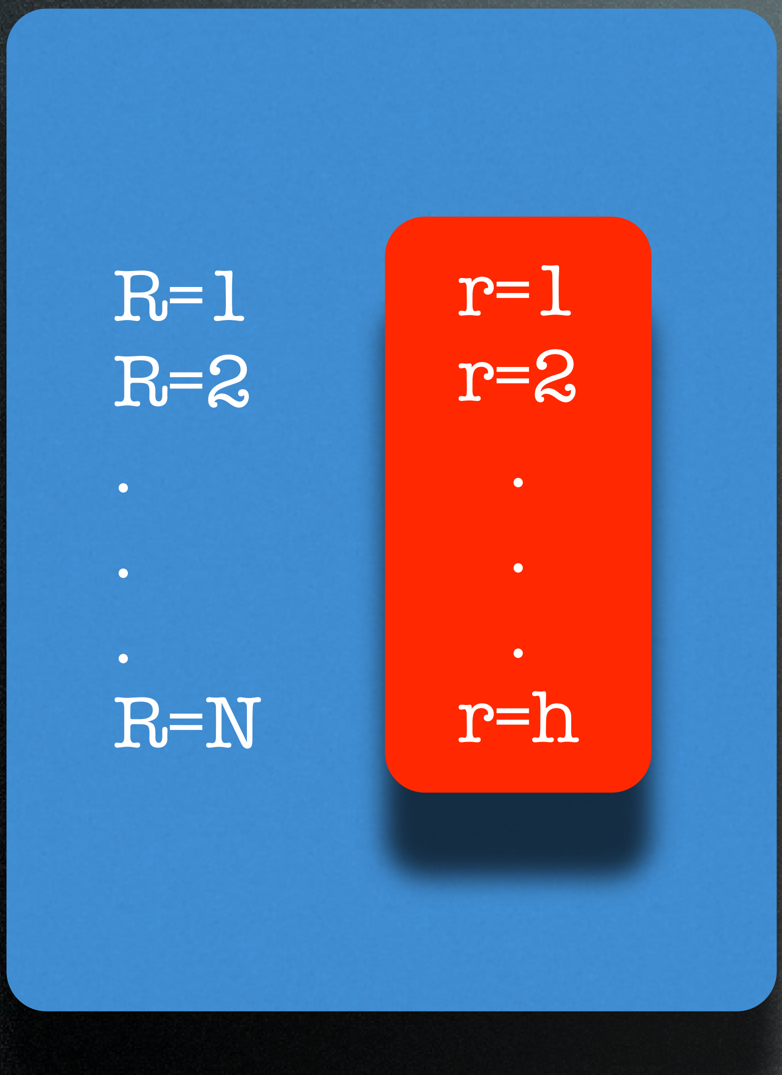
- Indegree
  - Google service
  - Yahoo service
  - Repeated a few months apart
- Traffic
  - Alexa service
  - Page views in 3 months
  - Domains vs. sites vs. pages
- 28,164 sites
  - about 2,000 popular
  - the rest random sample





Data vs. Models

What are we missing?



$$t \sim r_q^{-1.6}$$

$$r_g \sim k^{-1.1}$$

Revised model

$$t(R,r,N,n,h) = \frac{r^{-\alpha}}{\sum_{m=1}^{n} m^{-\alpha}} \Pr(R,r,N,n,h)$$

$$\Pr(R,r,N,n,h) = p_{r-1}^{R-1} p_{n-r}^{N-R} h$$

$$= h^n (1-h)^{N-n} \begin{pmatrix} R-1 \\ \\ r-1 \end{pmatrix} \begin{pmatrix} N-R \\ \\ n-r \end{pmatrix}$$

Revised model

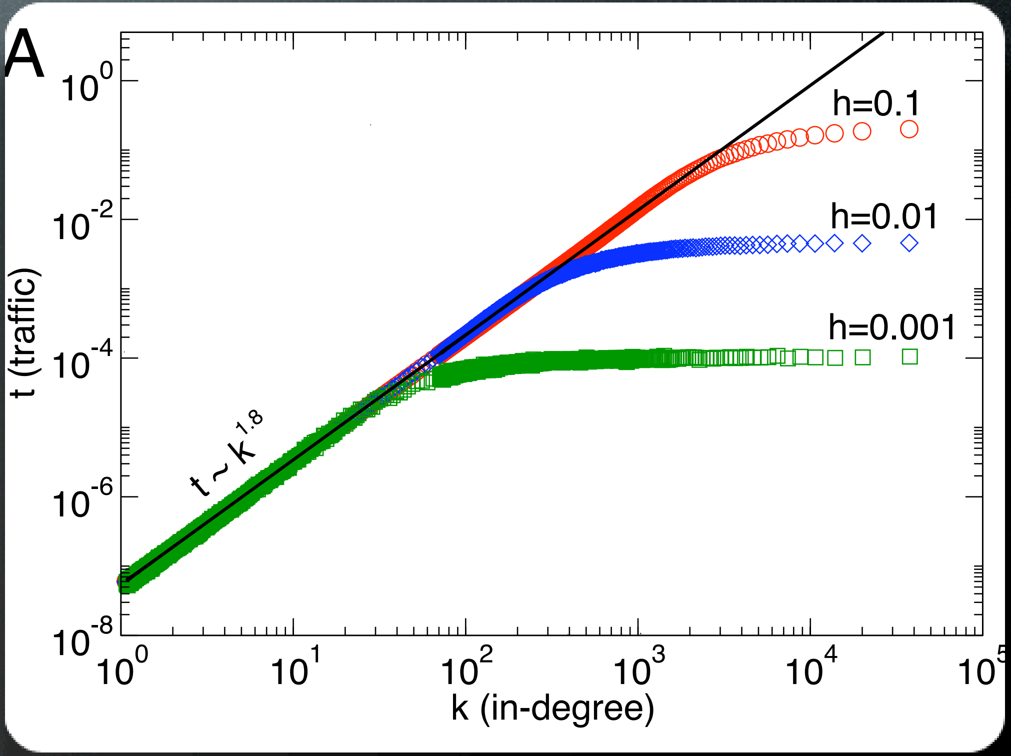
$$t(R,N,h) = \sum_{n=1}^{N} \sum_{r=1}^{n} \frac{r^{-\alpha}}{\sum_{m=1}^{n} m^{-\alpha}} h^{n}(1-h)^{N-n} \cdot \begin{pmatrix} R-1 \\ \\ r-1 \end{pmatrix} \begin{pmatrix} N-R \\ \\ n-r \end{pmatrix}$$

$$t(R,N,h) = h\,F(Rh)\,A(N)$$

$$F(Rh) \sim \begin{cases} const & \text{if } h \leq Rh \leq 1 \\ \\ (Rh)^{-\alpha} & \text{if } Rh \geq 1 \end{cases}$$

Effect of hit set size

The fewer the hits, the flatter the scaling between traffic and indegree

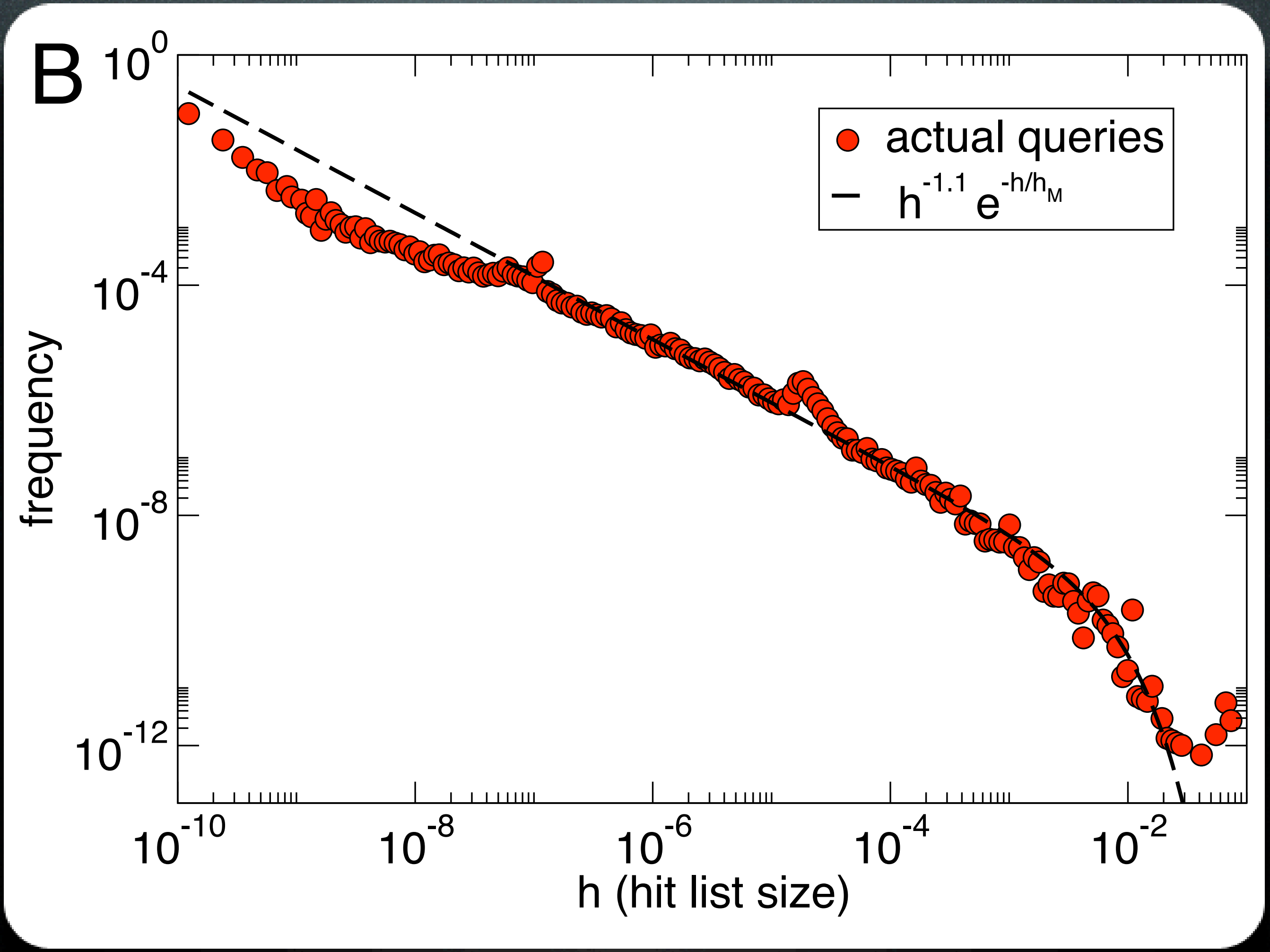


Idea: with few hits, established popular sites are less likely to be included and get a boost

Convoluting the curves

$$t_S(R, N) = \int_{h_m}^{h_M} S(h, N) t(R, N, h) dh$$

$$t_S(R, N) = \int_{1/N}^{h_M} S(h, N) h \, A(N) F(Rh) dh$$

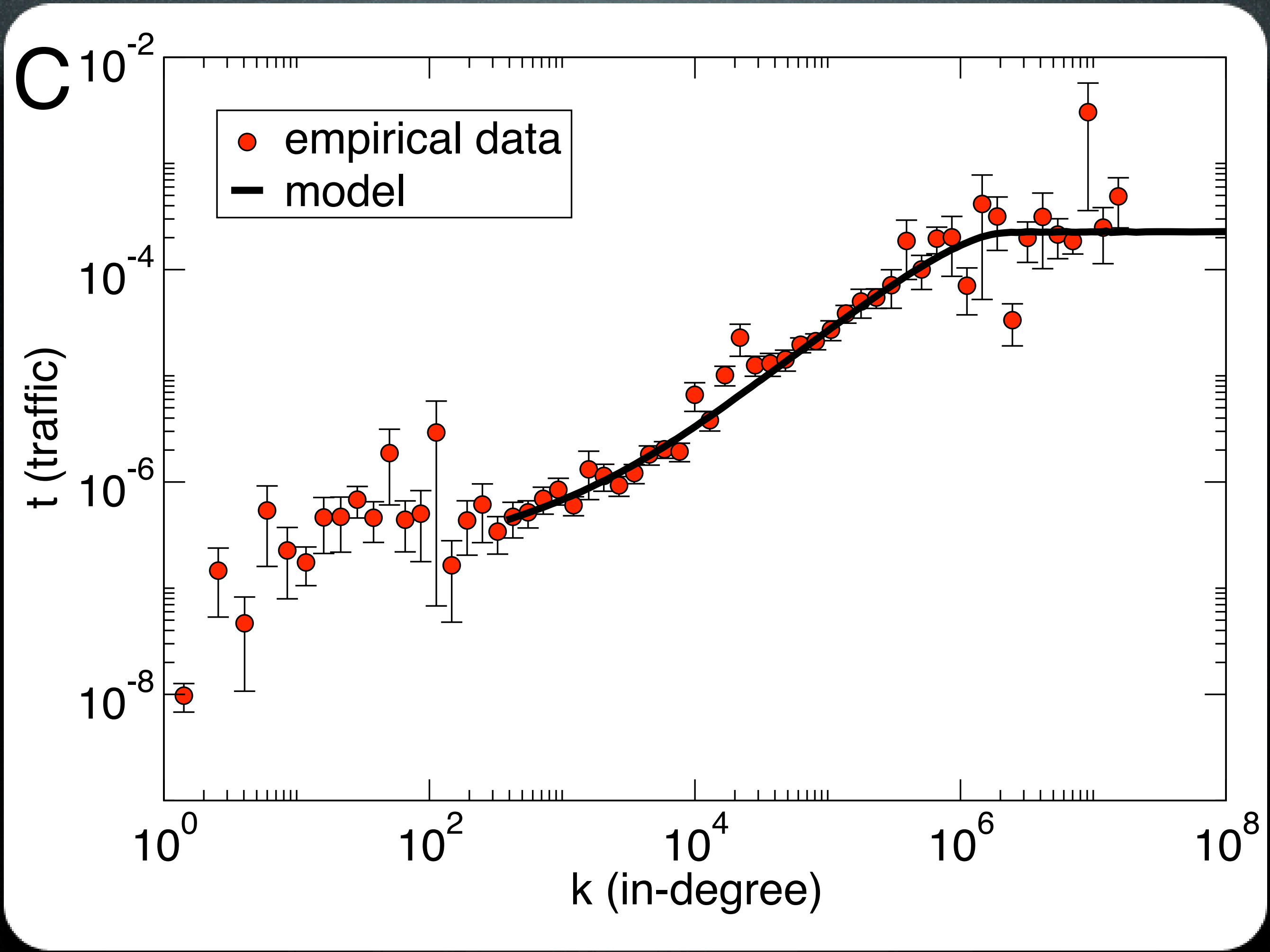


Distribution of hit set size

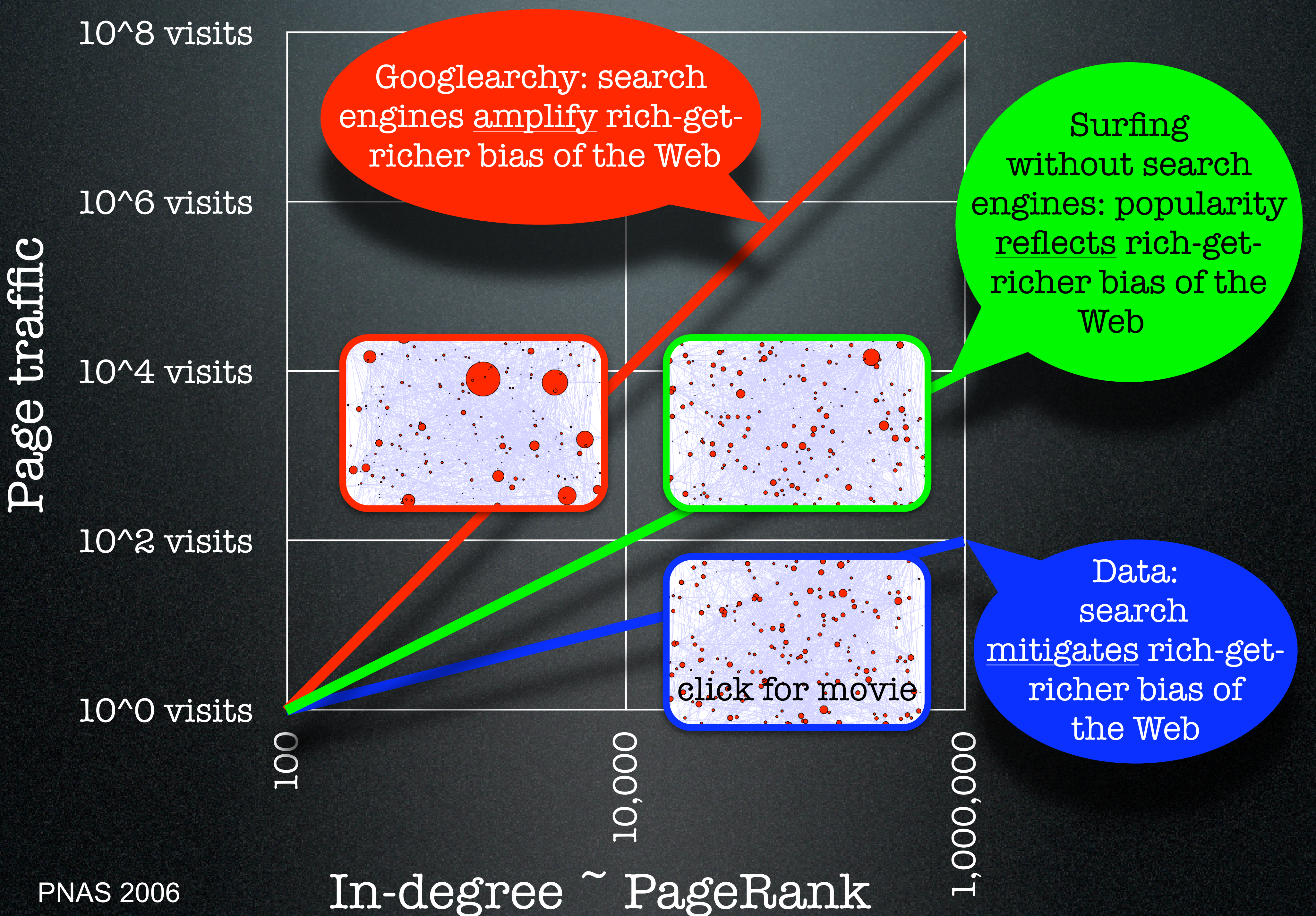
Integrating the curves by simulating the process

$$S(h, N) = B(N)h^{-\delta}$$

$$t_S(R, N) = \frac{A(N)B(N)}{N^{2-\delta}} \int_1^{h_M N} z^{1-\delta} F\left(\frac{R}{N} z\right) dz$$



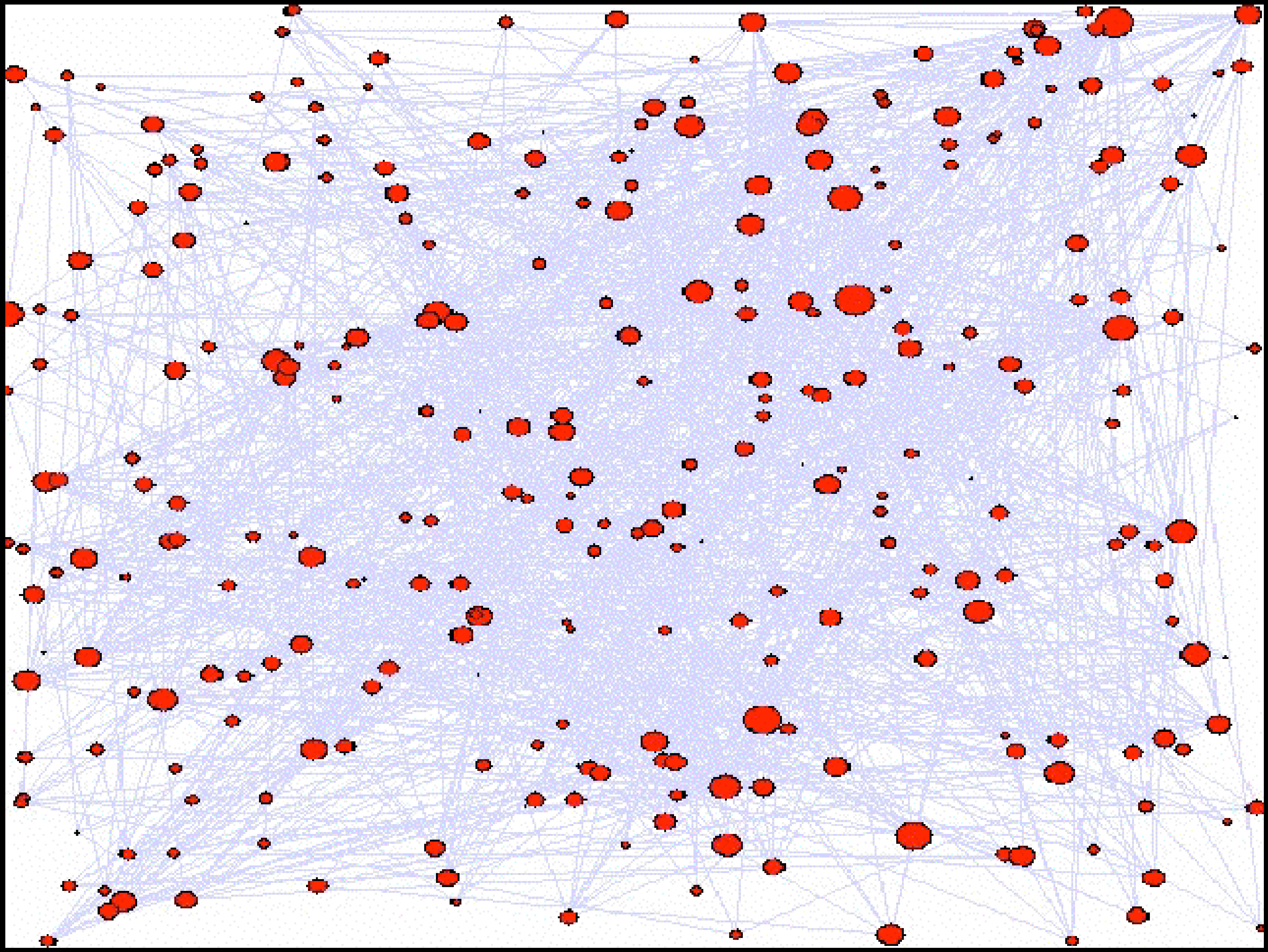
Data vs. “Semantically Correct” Model



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
Page traffic
10^8 visits
10^6 visits
10^4 visits
10^2 visits
10^0 visits
100
10,000
1,000,000
In-degree ~ PageRank

PNAS 2006

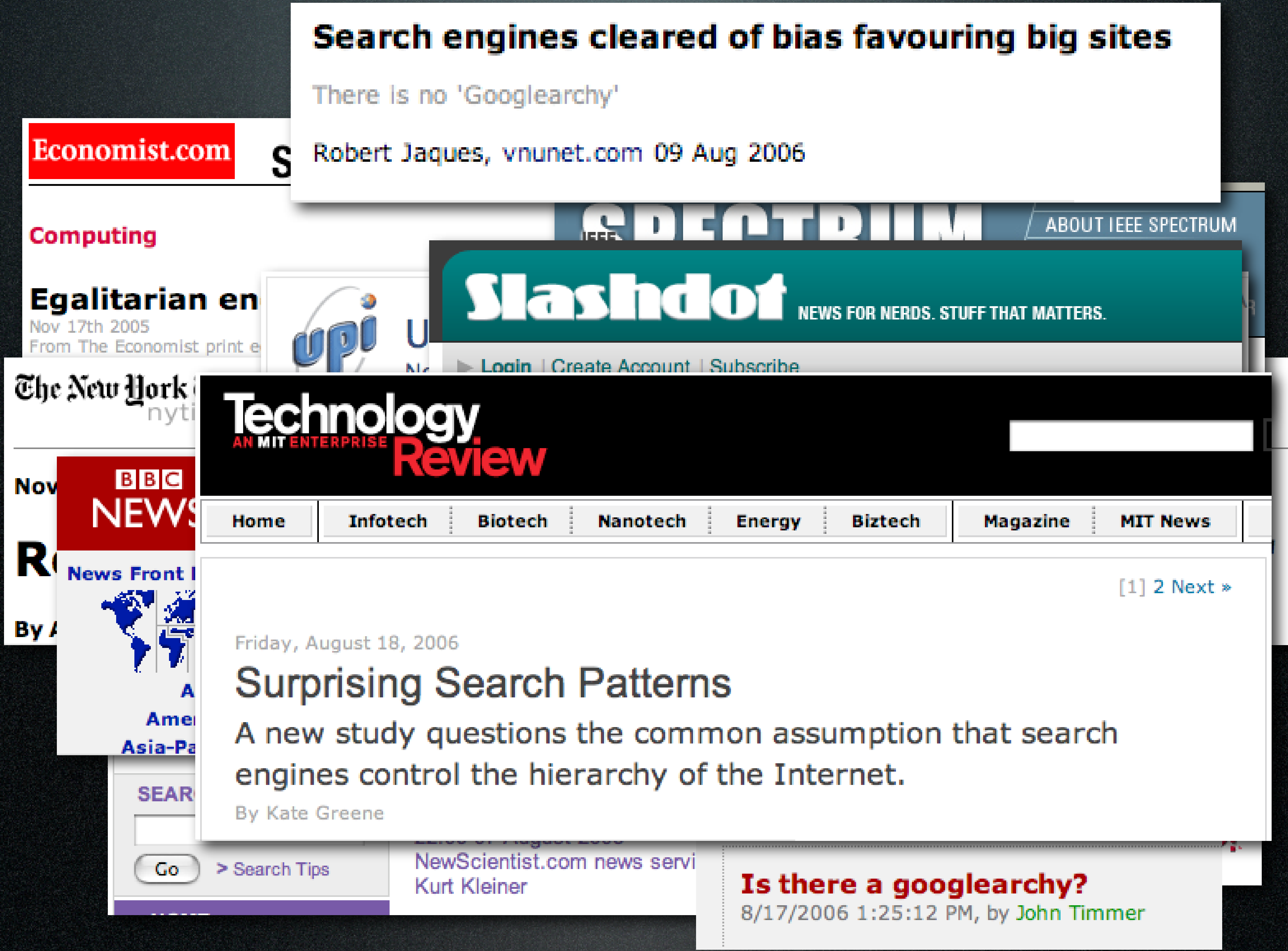
Empirical data: search mitigates rich-get-richer bias of the Web



Conclusions

- The use of search engines partially mitigates the rich-get-richer nature of the Web, giving new sites an increased chance of being discovered (compared to surfing alone), as long as they are about specific topics that match the interests of users.

- The combination of (i) how search engines index and rank results, (ii) what queries users submit, and (iii) how users view the results, leads to an egalitarian effect ("Googlocracy").


Search engines cleared of bias favouring big sites
There is no 'Googlearchy'
Robert Jaques, vnunet.com 09 Aug 2006
Economist.com
Computing
Nov 17th 2005
ABOUT IEEE SPECTRUM
Slashdot
NEWS FOR NERDS. STUFF THAT MATTERS.
Login | Create Account | Subscribe
The New York
BBC NEWS
Technology Review
AN MIT ENTERPRISE
Home
Infotech
Biotech
Nanotech
Energy
Biztech
Magazine
MIT News
[1] 2 Next »
Friday, August 18, 2006
Surprising Search Patterns
A new study questions the common assumption that search engines control the hierarchy of the Internet.
By Kate Greene
Go
> Search Tips
Kurt Kleiner
Is there a googlearchy?
8/17/2006 1:25:12 PM, by John Timmer

How do search engines affect Web growth?

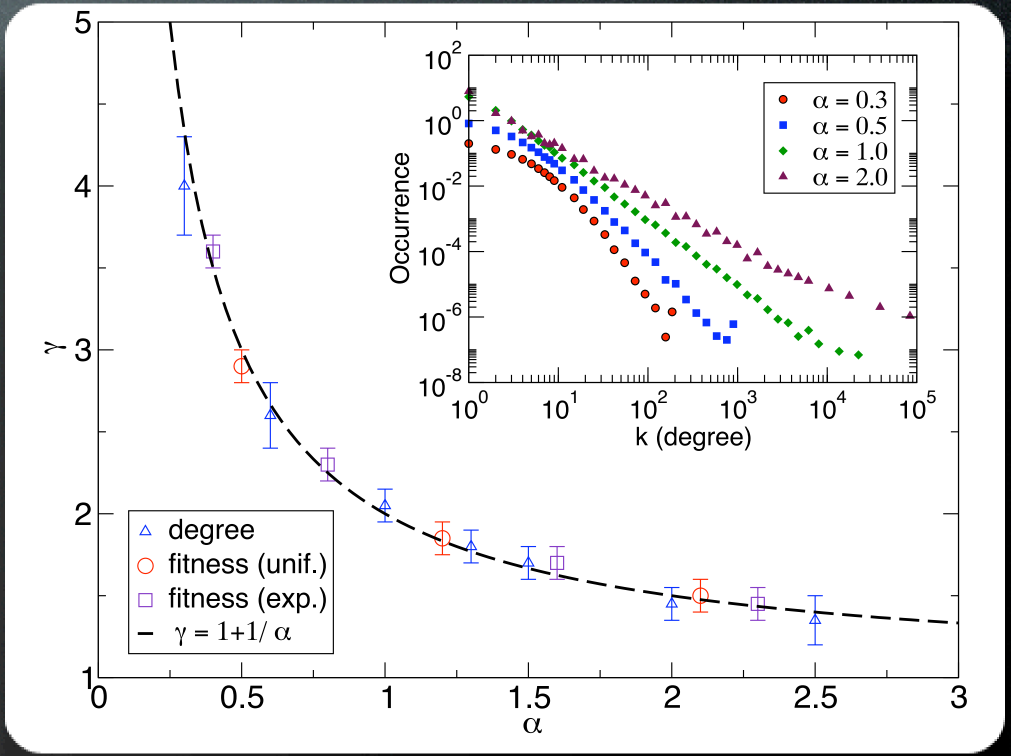
Web growth by searching

- Let’s model the evolution of the Web by assuming that pages are discovered mainly by searching
- Contrast current models:
  - Assume people link new pages to most popular ones (preferential attachment)
  - Must know degree
  - Only undirected networks
  - Only works if P(link) exactly proportional to degree
  - Disregard user interest topics, page content, etc.

General network growth model

$$p(t+1 \to j) \sim R_j^{-\alpha} \Rightarrow p(k) \sim k^{-(1+\frac{1}{\alpha})}$$

- Sort page by “prestige,” e.g., age, degree, PageRank, relevance, etc.
- No need to know values of original “prestige” measure
- R: rank (1, 2, ...)



PRL 2006

Limited information

- What if new nodes do not know global ranks, but only local ranks within a selected subset of all existing nodes?
- Preferential attachment 'breaks'...
- Two cases:

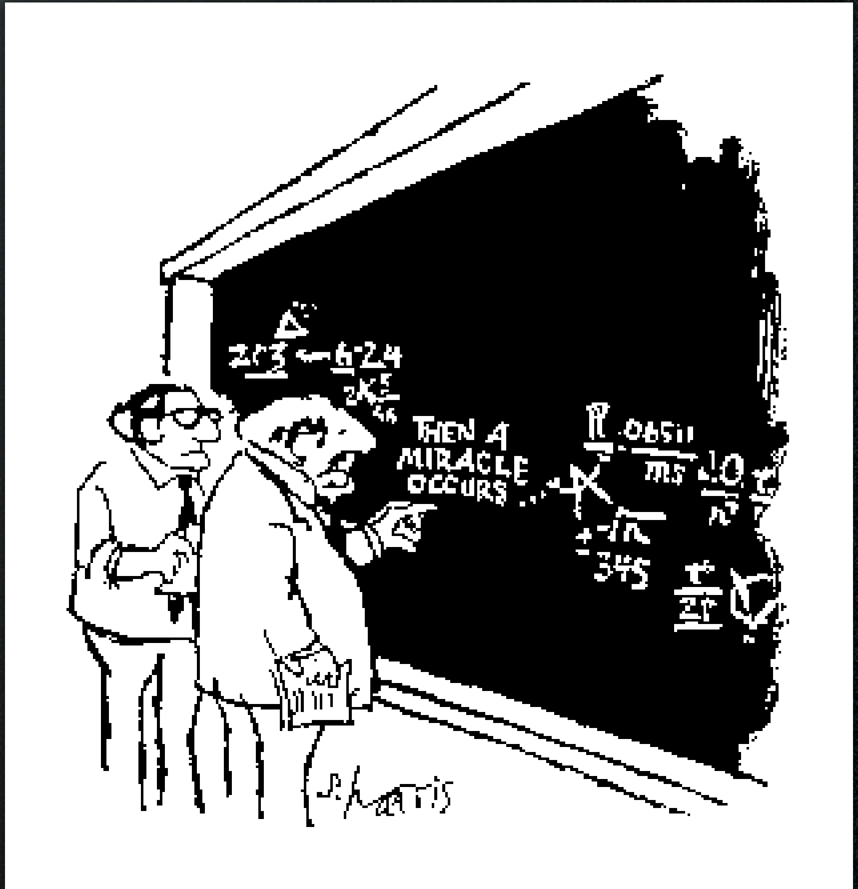
1. Each node is selected with fixed probability, h (independent of N): degree distribution still scale-free!
2. Different nodes may have different knowledge: a bit more complicated...

2. Each node is selected with probability h distributed as:

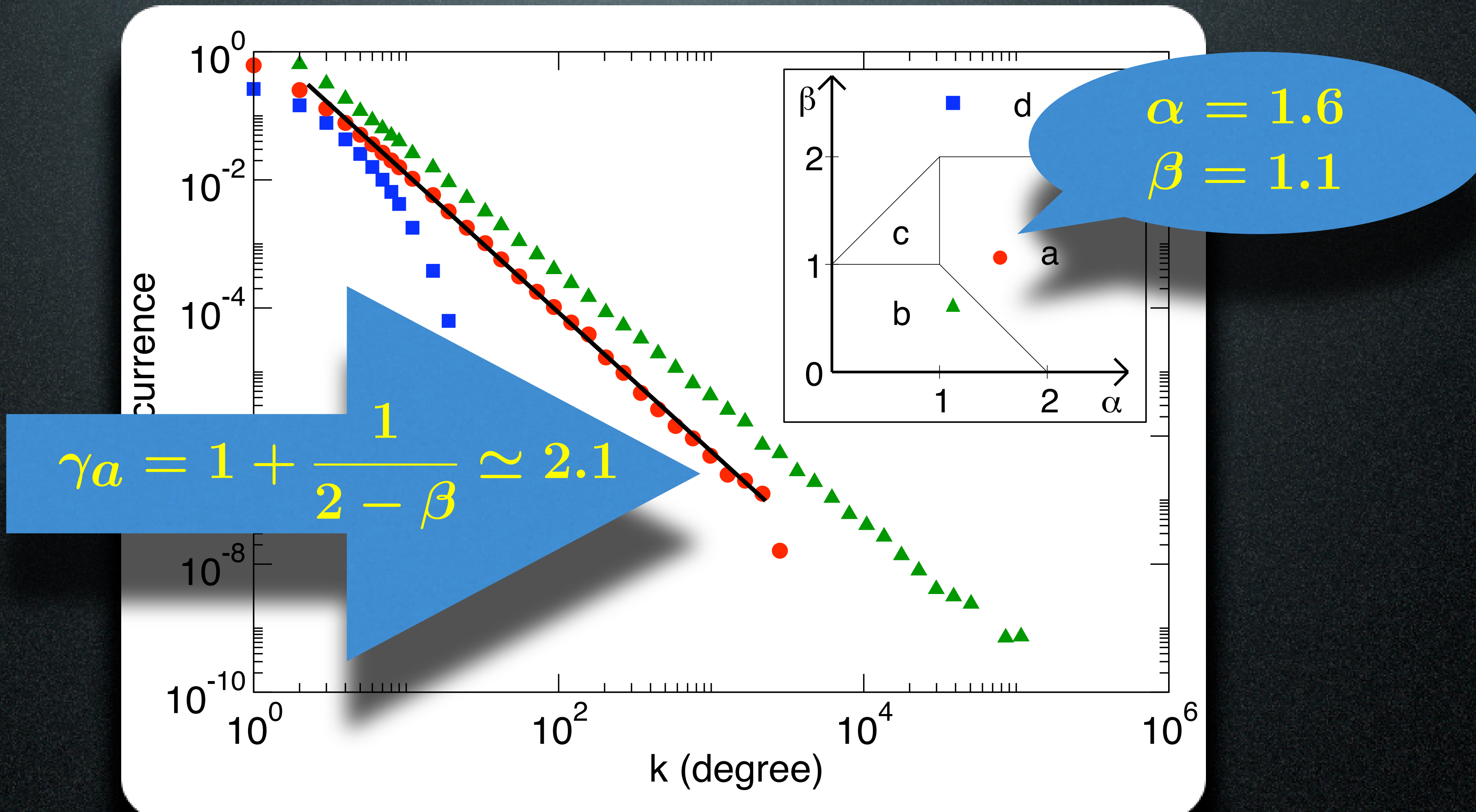
$$p(h) \sim h^{-\beta} \Rightarrow p(k) \sim k^{-f(\alpha,\beta)}$$

General model:

- uniform distribution for β=0
- exponential for β→∞



"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."



Web as special case!

Rank based growth model

- Works with many prestige measures
- No need to know degree or other prestige values, only ranks
- Works with broad class of power functions for P(link)
- Works with directed networks
- Works with limited information
- Strong stability against variation in the parameters
- Web search as special case: close prediction of Web graph's topological features

Thanks

Indiana University School of informatics

altavista

NSF

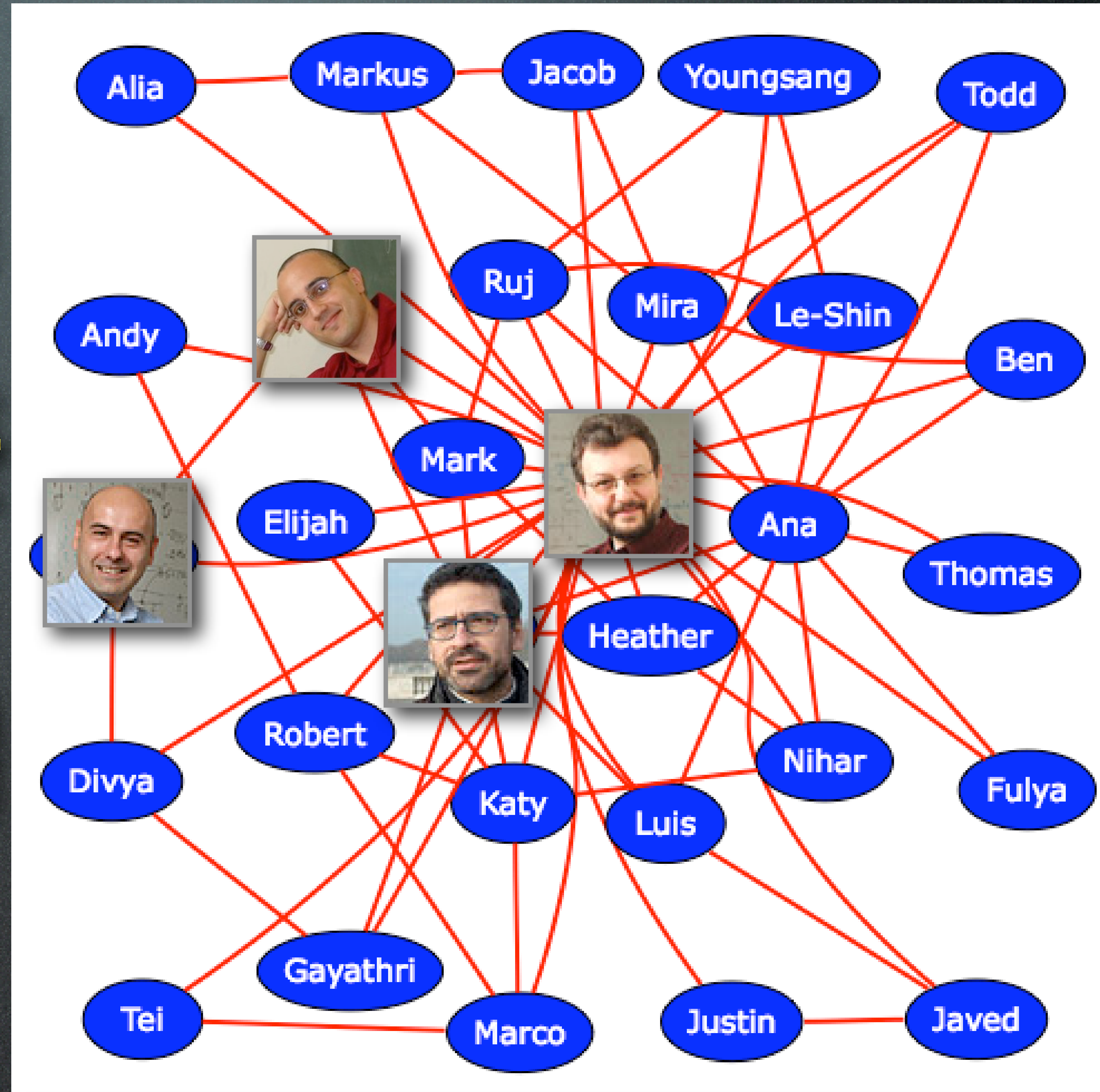
Alexa

STANFORD COMPUTER SCIENCE

YAHOO!

Google

csn.indiana.edu



NaN: Networks and Agents Network