Digital Media and Social Networks

Laurissa Tokarchuk
www.eecs.qmul.ac.uk/~laurissa
laurissa.tokarchuk@qmul.ac.uk

Lecture 5: STRUCTURE OF THE WEB, SEARCH AND POWER LAWS

Queen Mary
University of London

*some slides copyright Hamed Haddadi (QMUL) and Cecilia Mascolo (Cambridge)

IN THIS LECTURE

We describe power law networks and their properties and show examples of networks which are power law in nature, including the web.

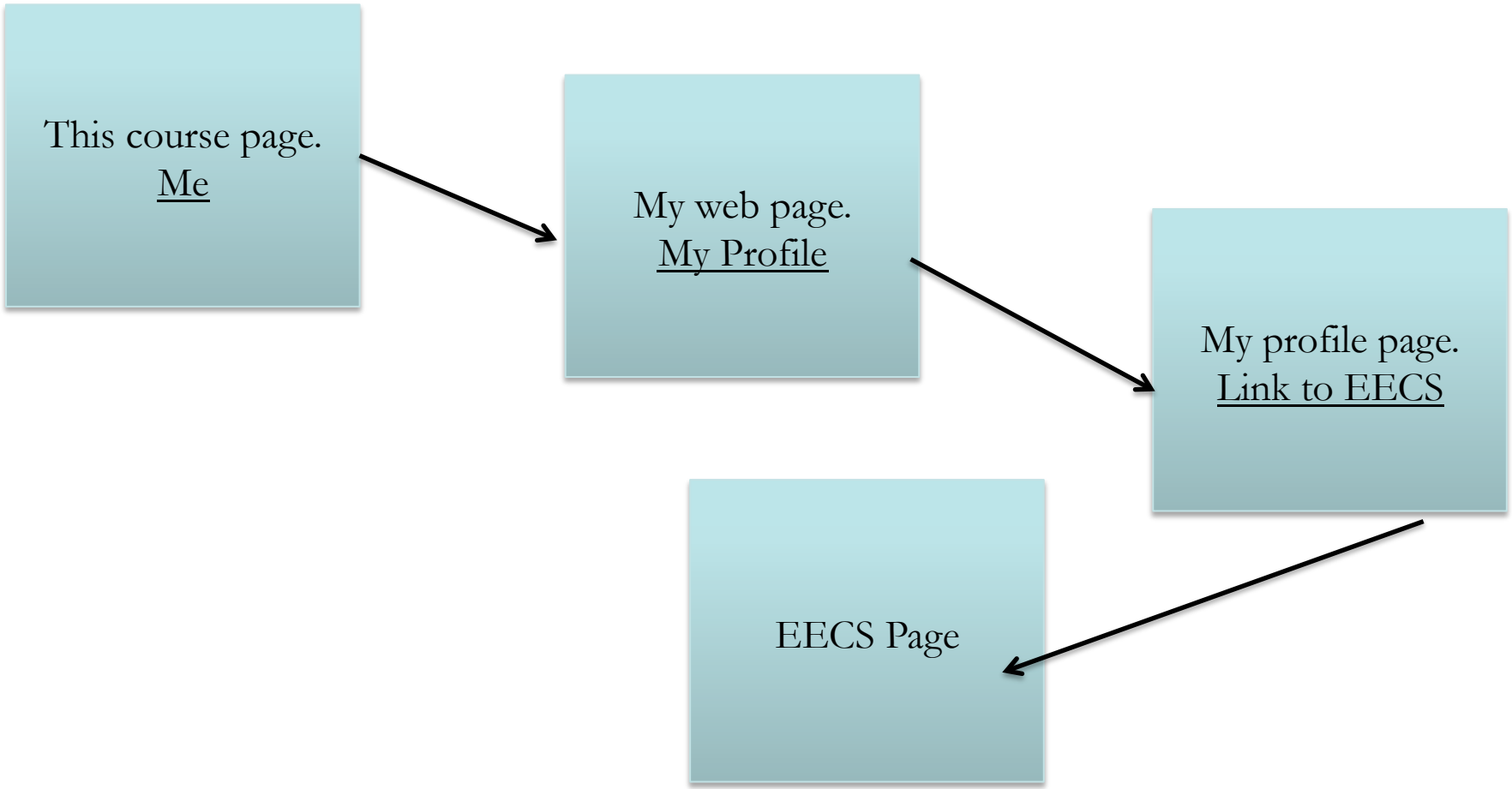
We present the preferential attachment model which allows the generation of power law networks.

We study prediction of power laws

We introduce search and PageRank

THE WEB IS A GRAPH…

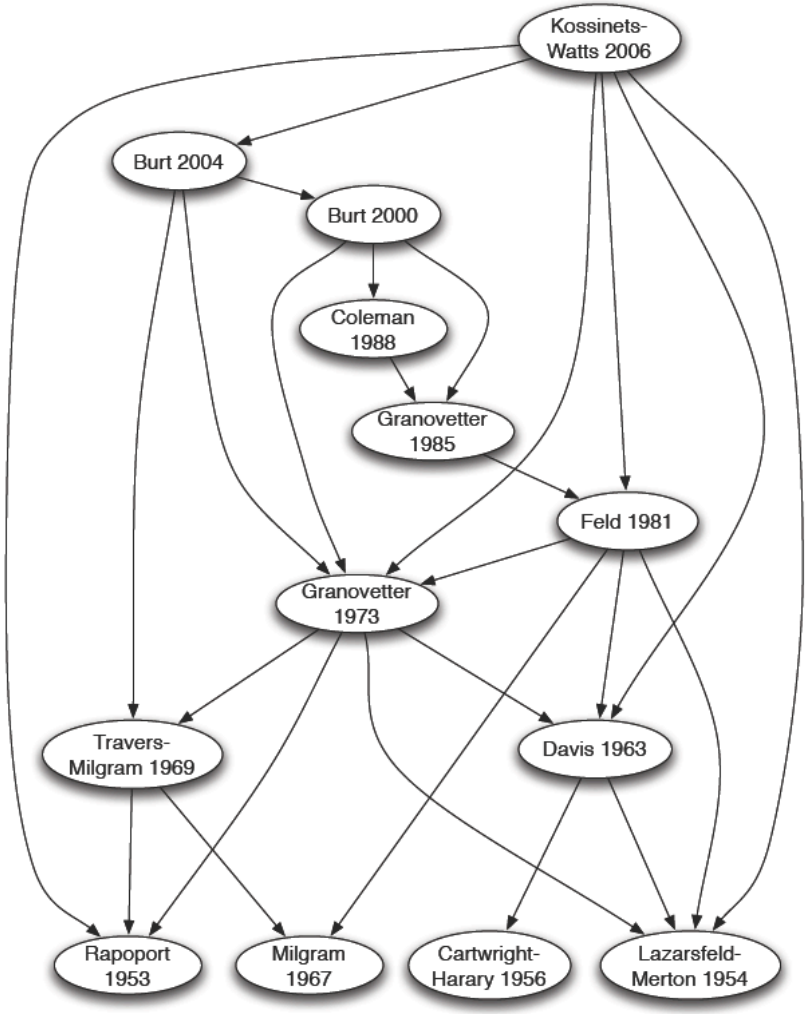
[CH 13, EASLEY & KLEINBERG]



PRECURSOR OF HYPERTEXTS

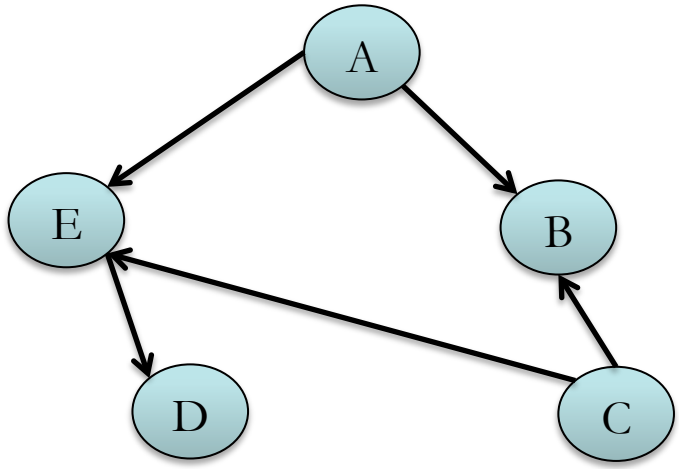
Citation networks of books and articles.

Difference: links point only backwards in time



Web is a **Directed** Graph

- **Path**: A path from A to B exists if there is a sequence of nodes beginning with A and ending with B such that each consecutive pair of nodes is connected by an edge pointing in the forward direction.

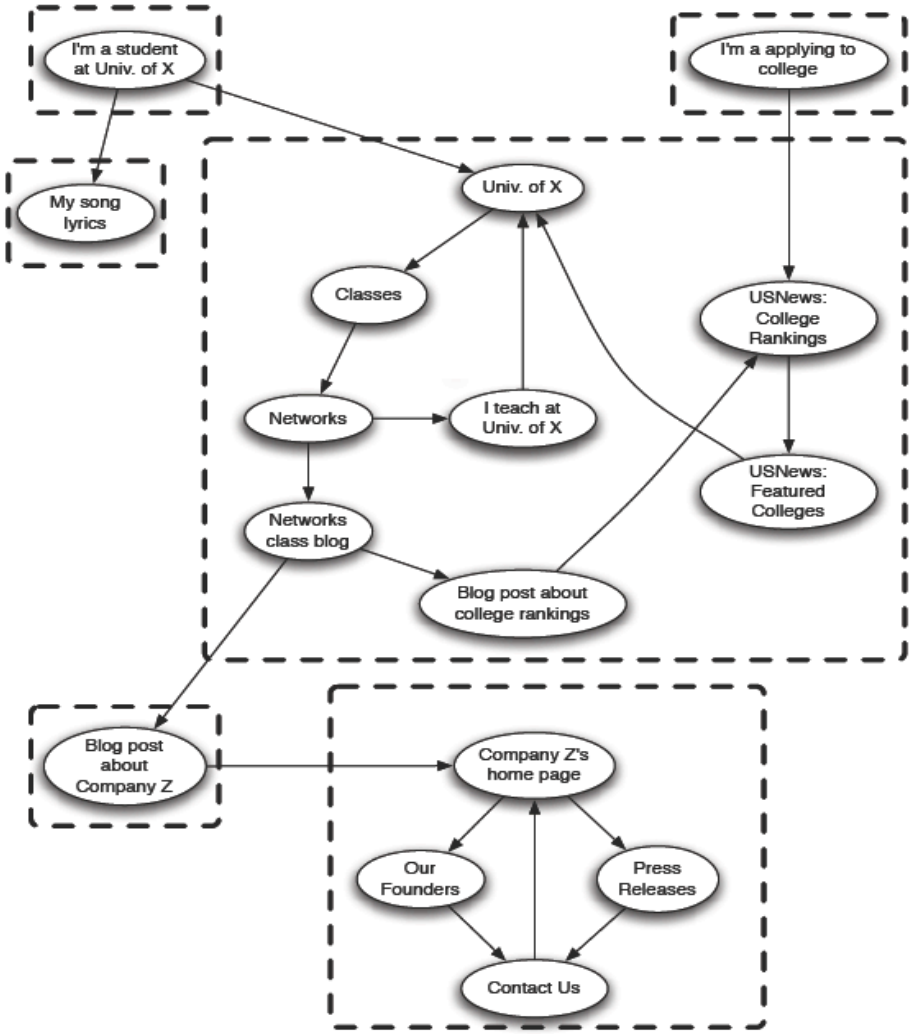


Strongly Connected Component

- A *strongly connected component (SCC)* in a directed graph is a subset of nodes such that:
  - i) Every pair in the subset has a path to each other
  - ii) The subset is not part of some larger subset with property i)

- *Weakly connected component (WCC)* is the connected component in **the undirected** graph derived from the directed graph.
  - i) Two nodes can be in the same WCC even if there no directed path between them.

SCC example

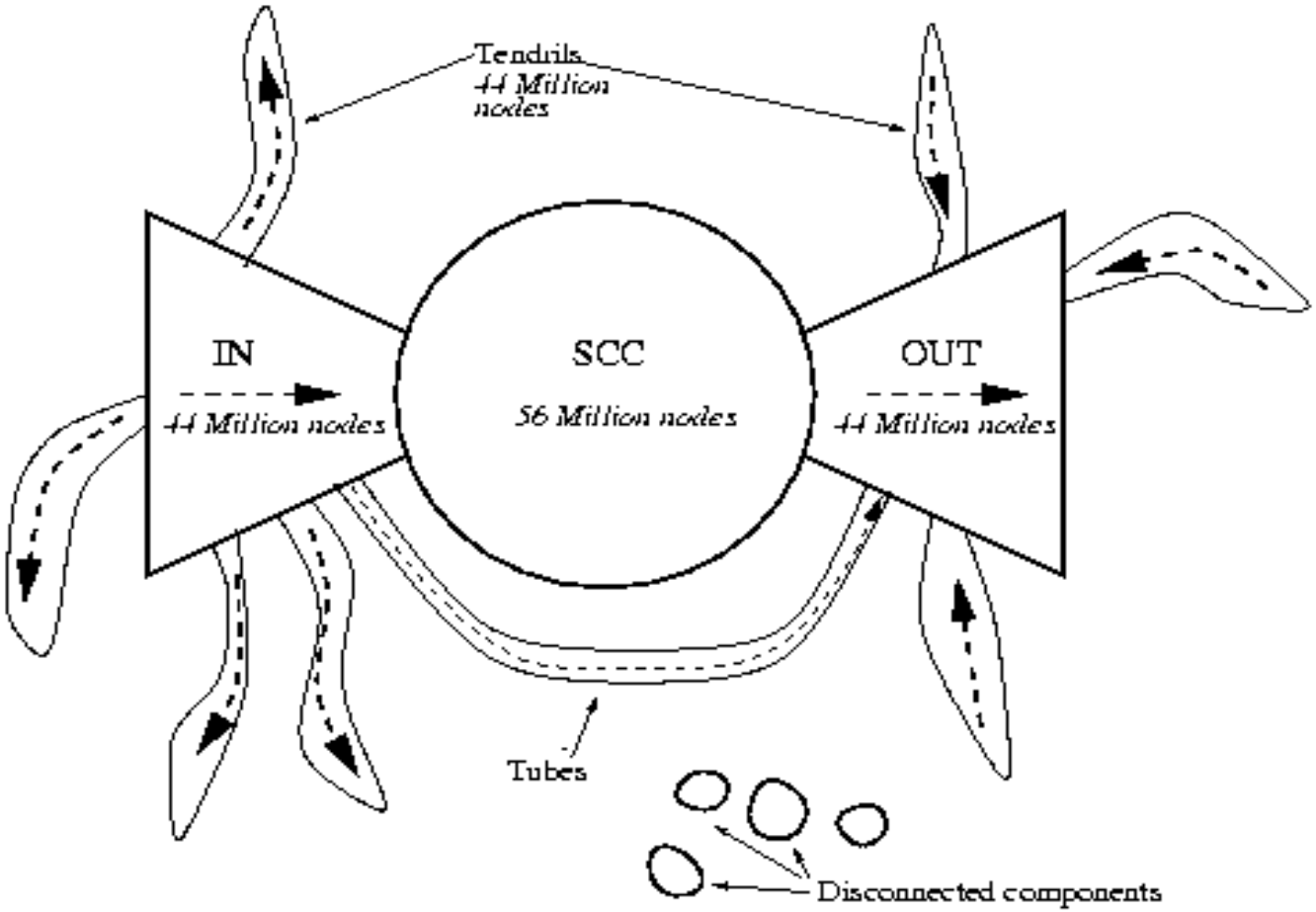

I'm a student at Univ. of X
I'm a applying to college
My song lyrics
Univ. of X
Classes
USNews: College Rankings
Networks
I teach at Univ. of X
USNews: Featured Colleges
Networks class blog
Blog post about college rankings
Blog post about Company Z
Company Z's home page
Our Founders
Press Releases
Contact Us

The Web *(Broder 2000)*

- Study the web as a graph. Why?
  - Design crawling strategies
  - Understand sociology of content creation
  - Analyse behavior of web algorithms that use link information
  - (many other reason!)
- Web isn't like a highly-connected bowl of spaghetti, but connectively is limited by a high level global structure.

The Web

- Broder'00
- Data from Altavista (200 million pages)
- 186M nodes in the WCC (90% of links)



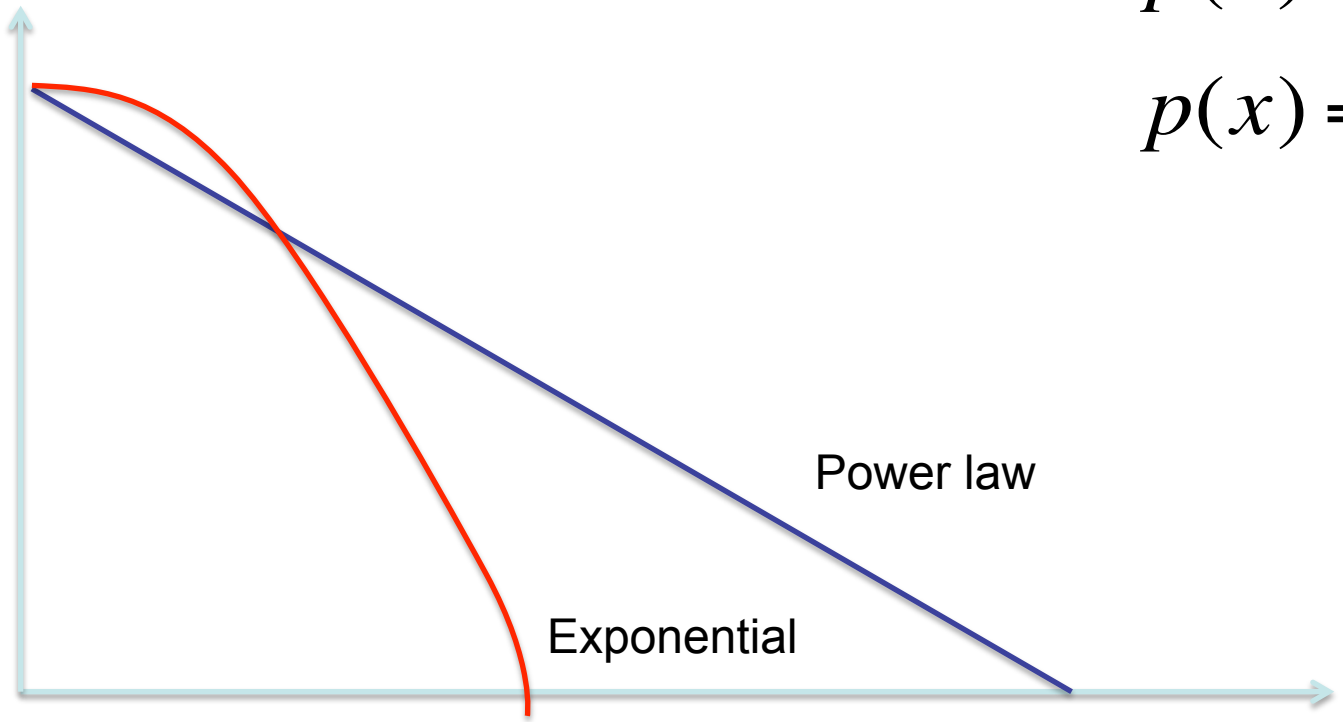
Popularity of Web Pages

- How do we expect the popularity of web pages to be distributed?
  - What fraction of web pages have k in-links?
    - If each page decides independently at random whether to link to any given other page then the n of in-links of a page is the sum of independent random quantities -> normal distribution
    - In this case, the number pages with k in-links decreases exponentially in k
    - Is this true for the Web?

Power Law vs Exponential

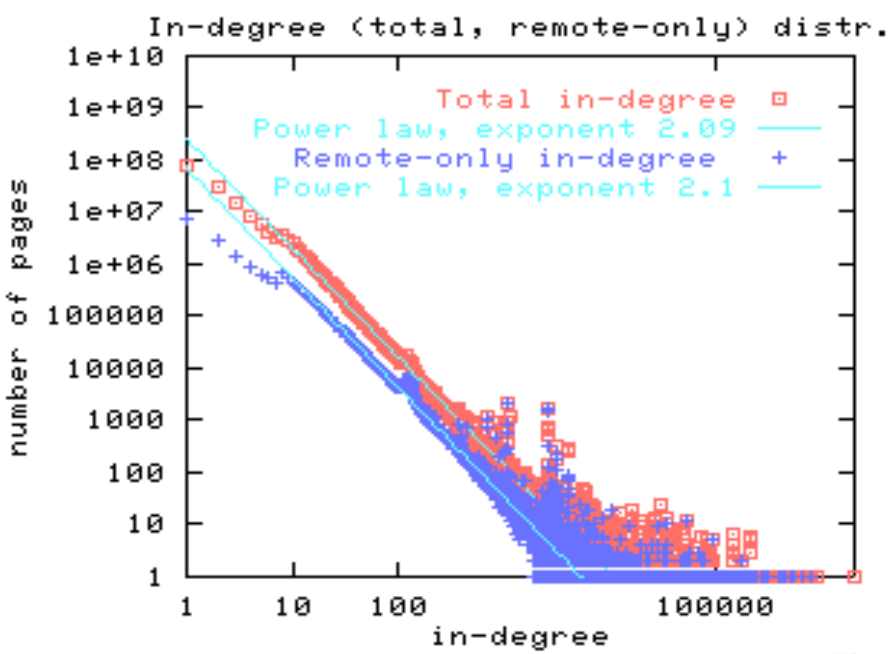
$$p(x) = x^{-\alpha}$$

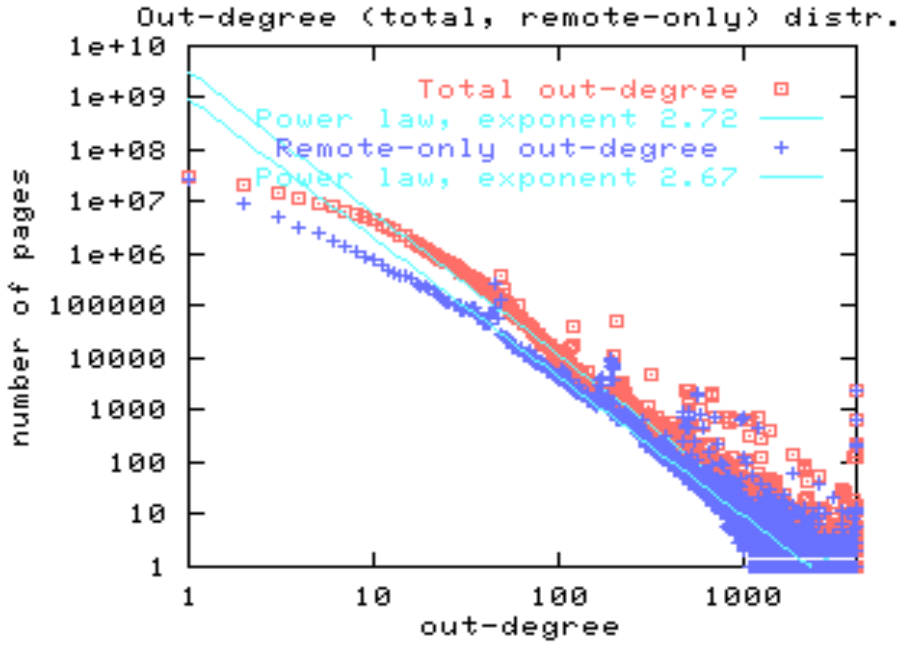
$$p(x) = e^{-\lambda x}$$



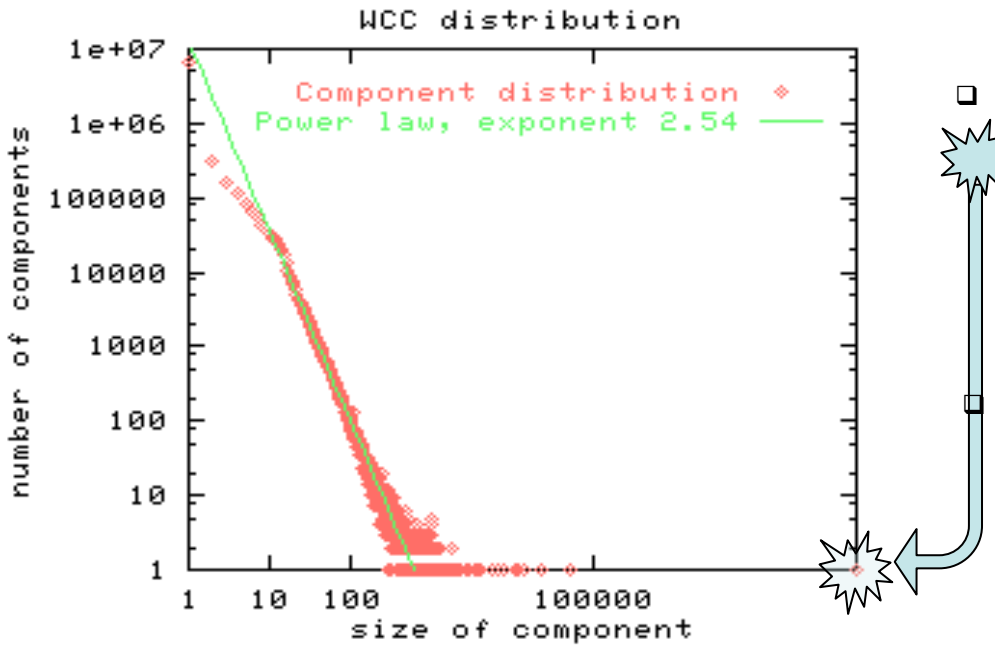
Degree distribution for the Web

- Finding: degree distr. proportional to $\sim 1/k^2$
- $1/k^2$ decreases much more slowly than a normal distribution





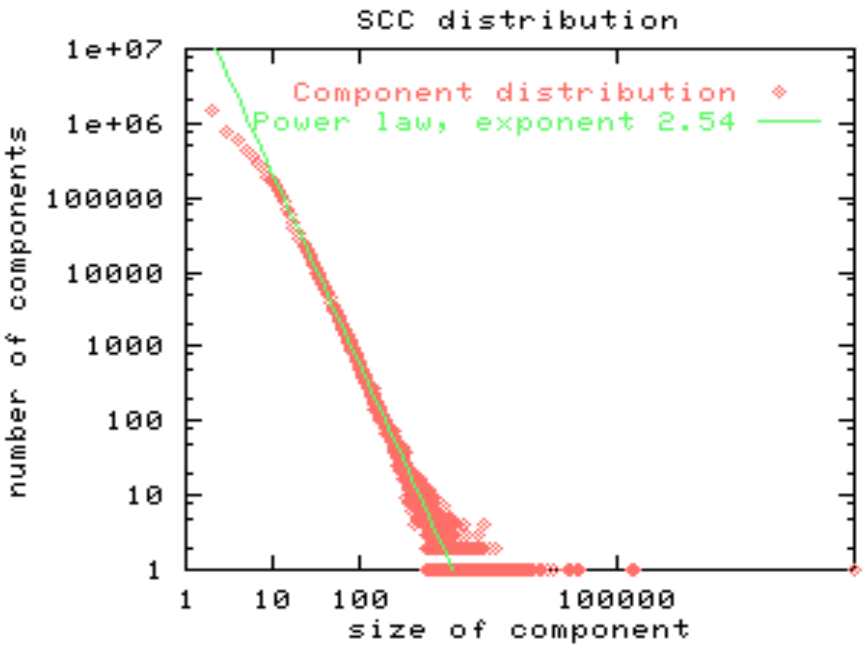
Distribution of WCC



- Treat web graph as undirected.
- Giant component → 186 million nodes.
  - 91% reachable!
  - Very well connected graph *(if you could go forwards and backwards)*
- Does connectivity = few high degree junctions?
  - NO
  - Remove all nodes d >= 5, giant component = 59 million.
    - Web is extremely resilient.
    - Hub nodes are embedded in graph that is well connected without them.

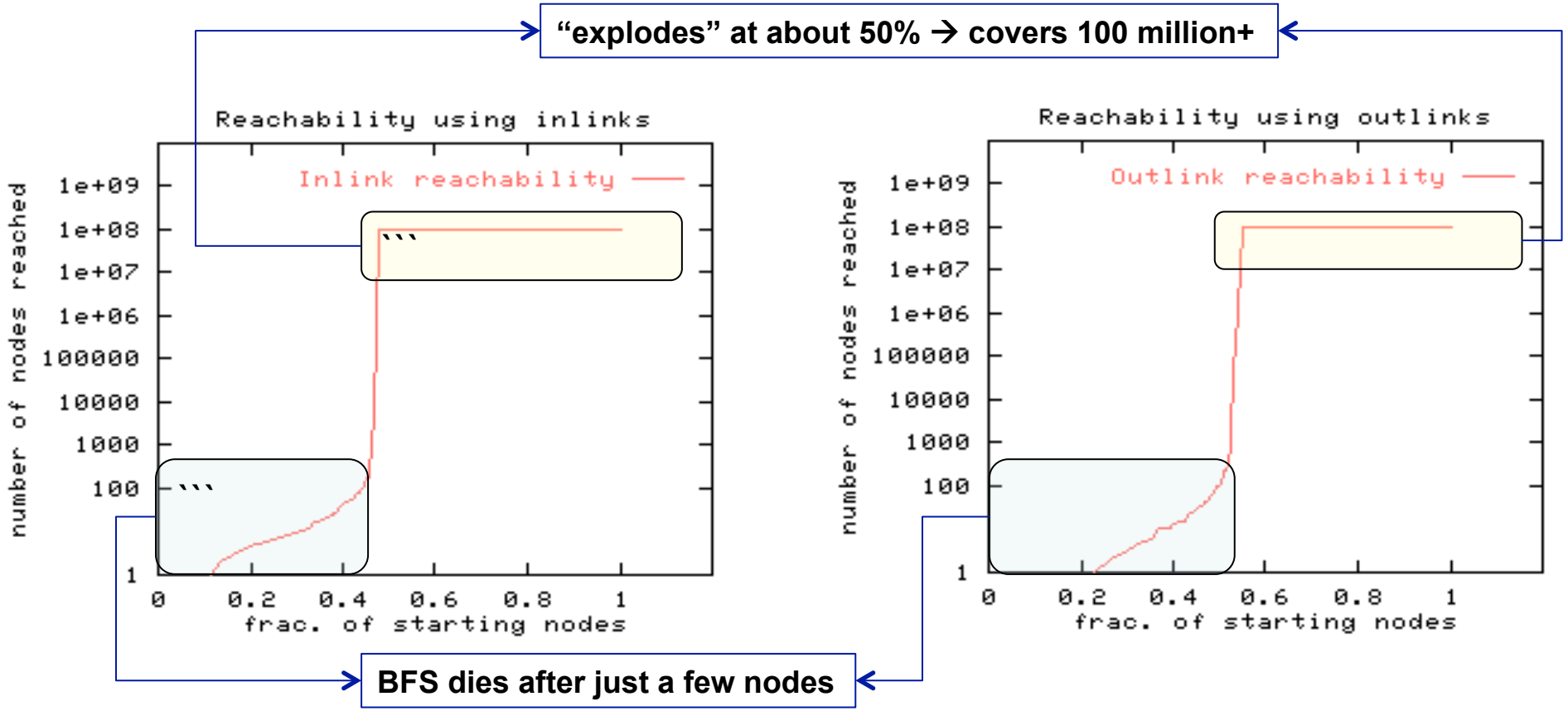
Distribution of SCC

- Single SCC of about 56 million pages → only 28% of all pages.
  - Where have all the pages gone?



Reachability

- Followed links backwards and forward
- BFS twice over 570 randomly chosen nodes.

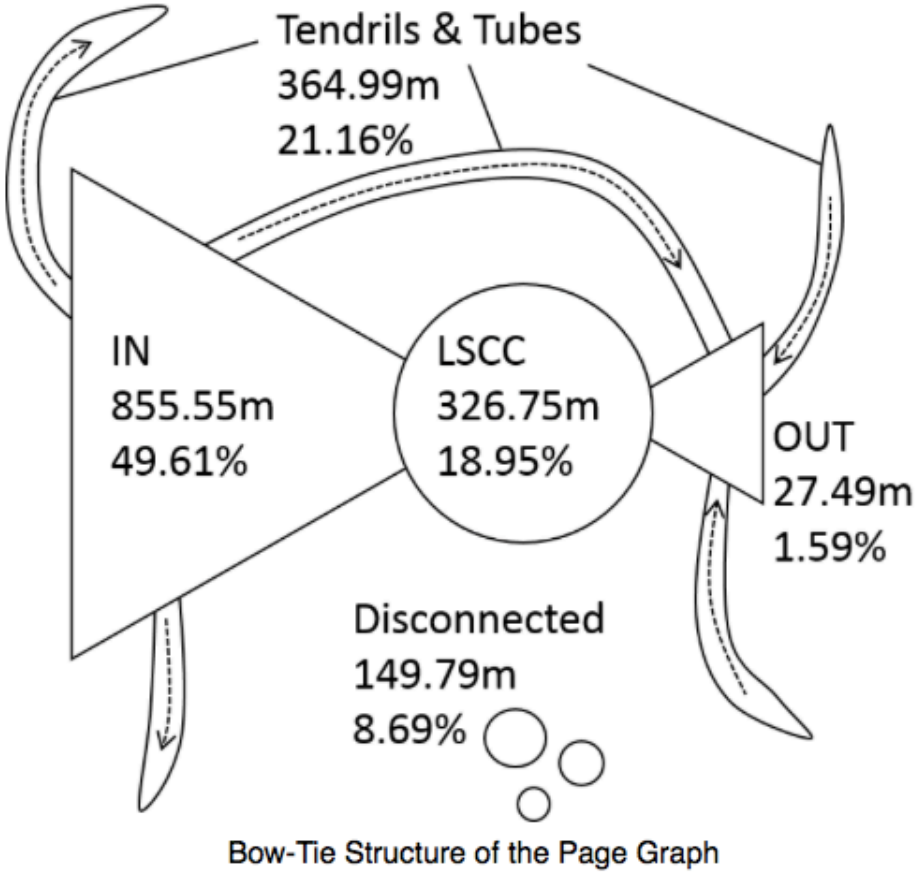


Diameter of the Web

- 75% of the time there is no directed path between two random nodes
- Average distance of existing paths: 16
- Average distance of undirected paths: 6.83

- Diameter in the SCC is at least 28

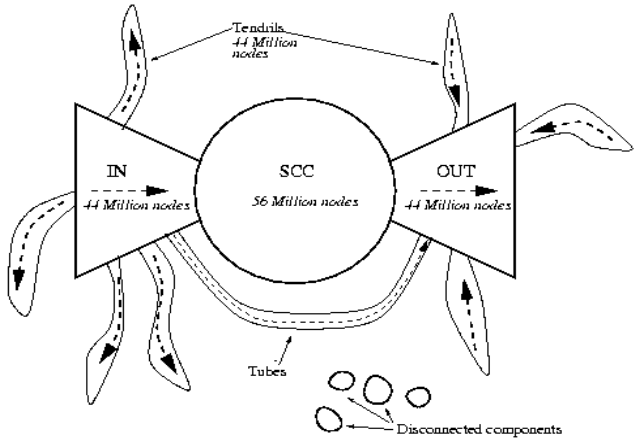
How things change over time - 2014?



Bow-Tie Structure of the Page Graph

Web Data Commons - Hyperlink Graph Crawl:

- 1.7 billion web pages
- 64 billion hyperlinks between these pages.



http://webdatacommons.org/hyperlinkgraph/2014-04/topology.html

Power Laws aka Scale Free Networks

[Ch 18, Easley & Kleinberg]

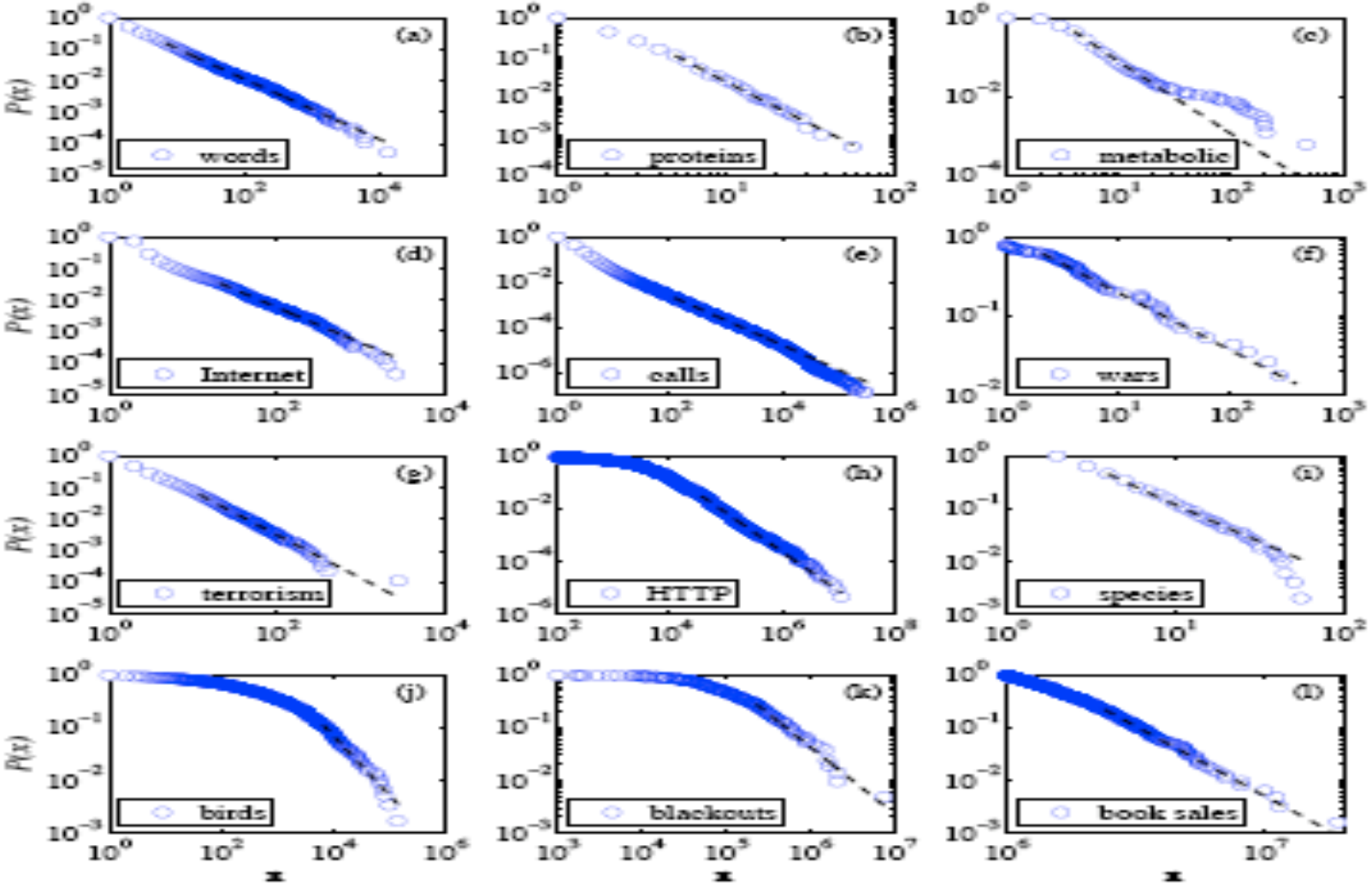
- We have seen that the degree distribution followed a straight line in log-log

$$\ln p_k = -\alpha \ln k + c$$

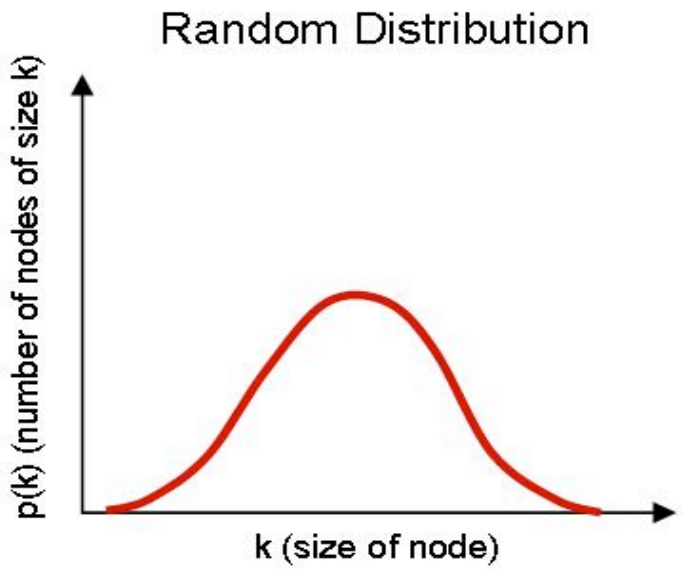
$$p_k = Ck^{-\alpha}$$

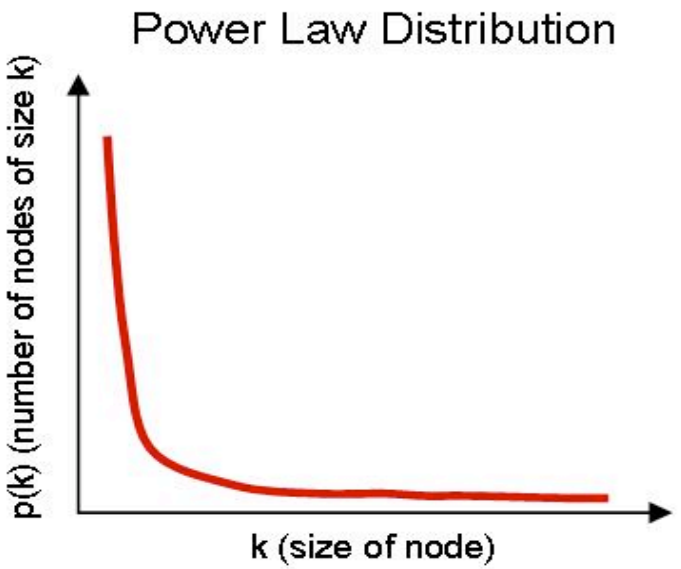
- α defines the slope of the curve
- α is typically between 2 and 3.

Power Laws in various domains

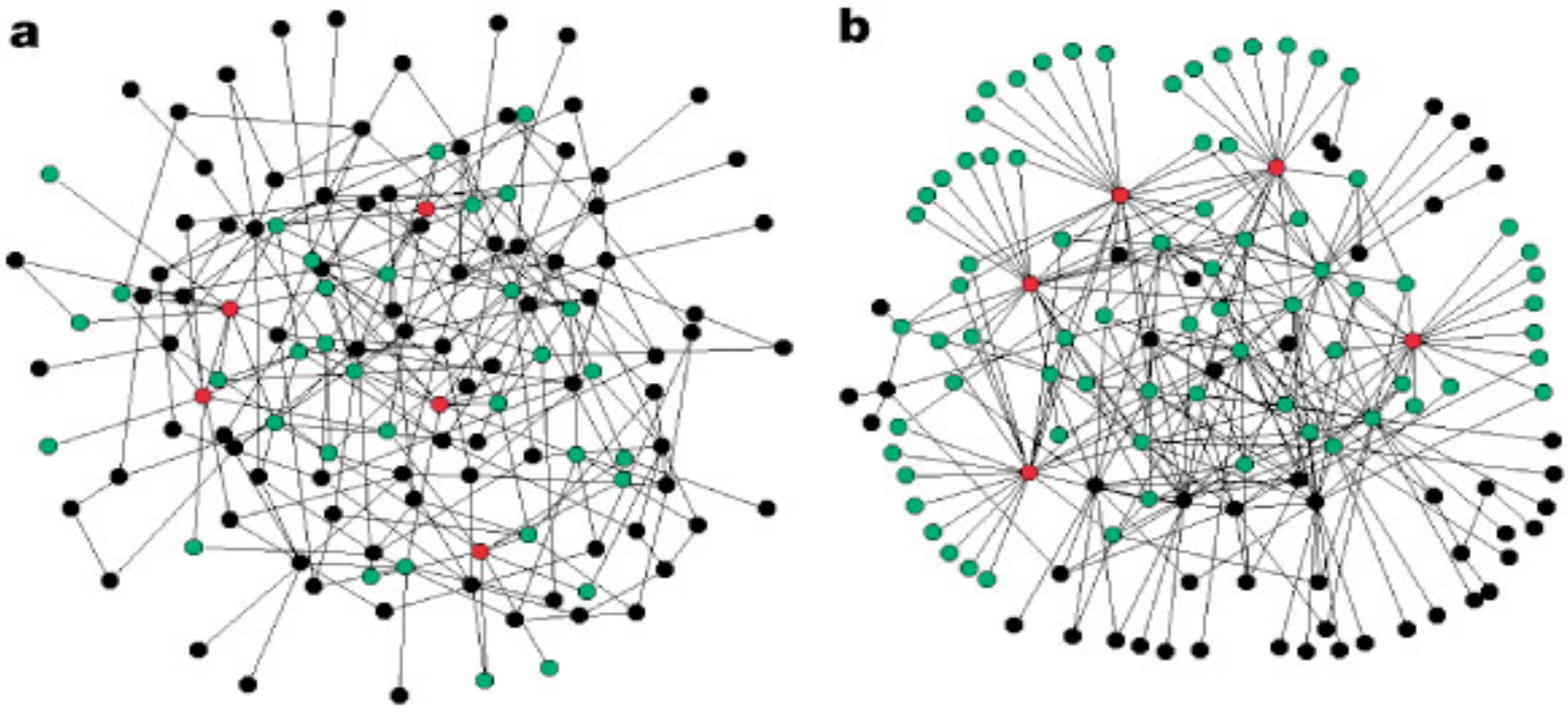


What does it mean?





Random vs Power Law Networks



What's a good model for scale free networks

- Let's use the web network as example:
- Pages are created in order (1,2,3..)
- Page j created and it links to an earlier page in the following way:
  - With prob. p, j chooses page i at random and links it;
  - With prob. 1-p, j chooses page i and links to the page i points to.
  - Repeat.
- The middle step is essentially a copy of the node i behaviour…

Remember – the web is just one example

Preferential attachment

- Pages are created in order (1,2,3..)
- Page j created and it links to an earlier page in the following way:
  - With prob. p, j chooses page i at random and links it;
  - **With prob. 1-p, j chooses a page z with prob. proportional to z's current number of in-links and links to z (ie proportional to degree).**
  - Repeat.

Rich-get-richer model

If we run this for many pages the fraction of pages with k in-links will be distributed approximately according to a power law $1/k^c$
c depends on p

Intuition

- With probability 1-p page j chooses a page l with probability proportional to l's number of inlinks and creates a link to l.

- This mechanism predicts that the growth happens so that
  - A page's popularity growth at a rate proportional to its current value.
  - The rich get richer effect amplifies the larger values

Preferential Attachment

- What have we shown?
- There is a “copying” behaviour happening in these networks where node seem to emulate other nodes.
- This is shown true for selection of books, songs, web pages, movies etc.

How predictable is the rich-get-richer process?

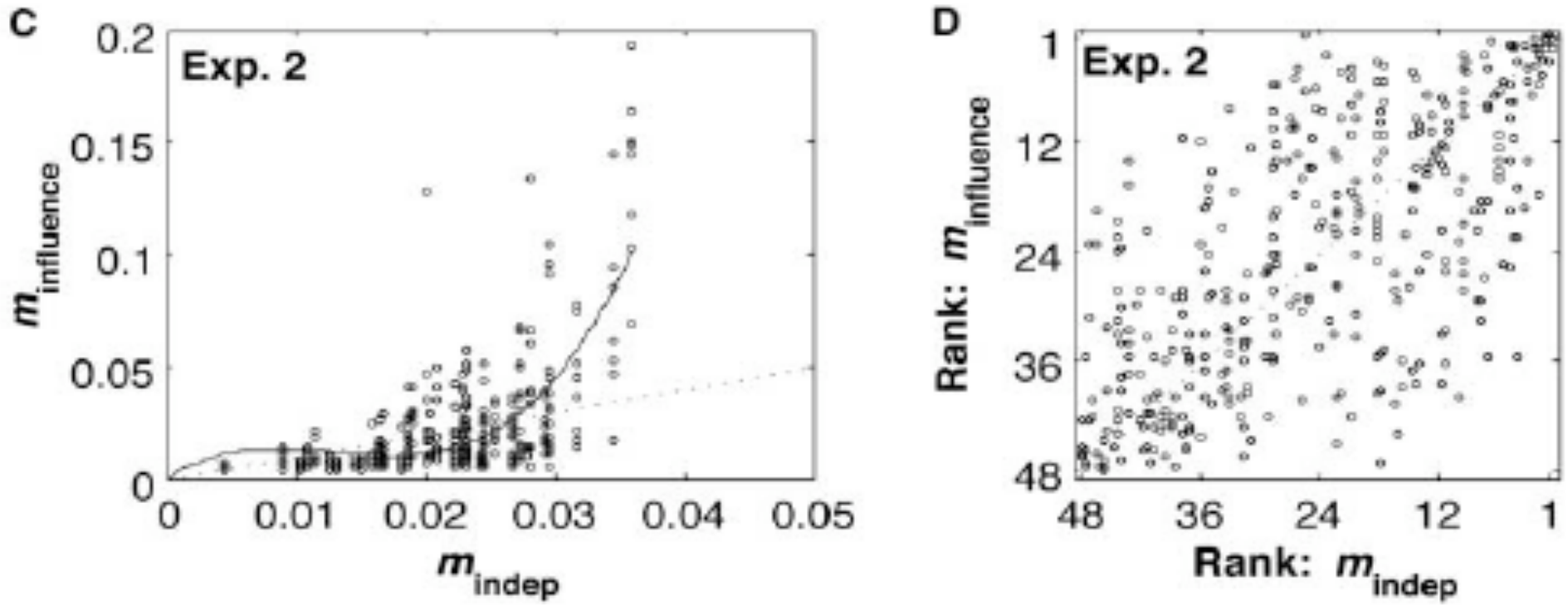
- Is the popularity of items in the power law predictable?
- Would a popular book still be popular if we go back in time and start the process again?
- Experiments show it would not…
  - The prediction problem: Experts often get it wrong…

Artificial Music Market [Salganik et al 06]

- Experiment:
  - 14,341 participants downloaded previously unknown songs either with or without knowledge of previous participants' choices.
  - Increasing the strength of social influence increased both inequality and unpredictability of success.
  - Success was also only partly determined by quality: The best songs rarely did poorly, and the worst rarely did well, but any other result was possible.

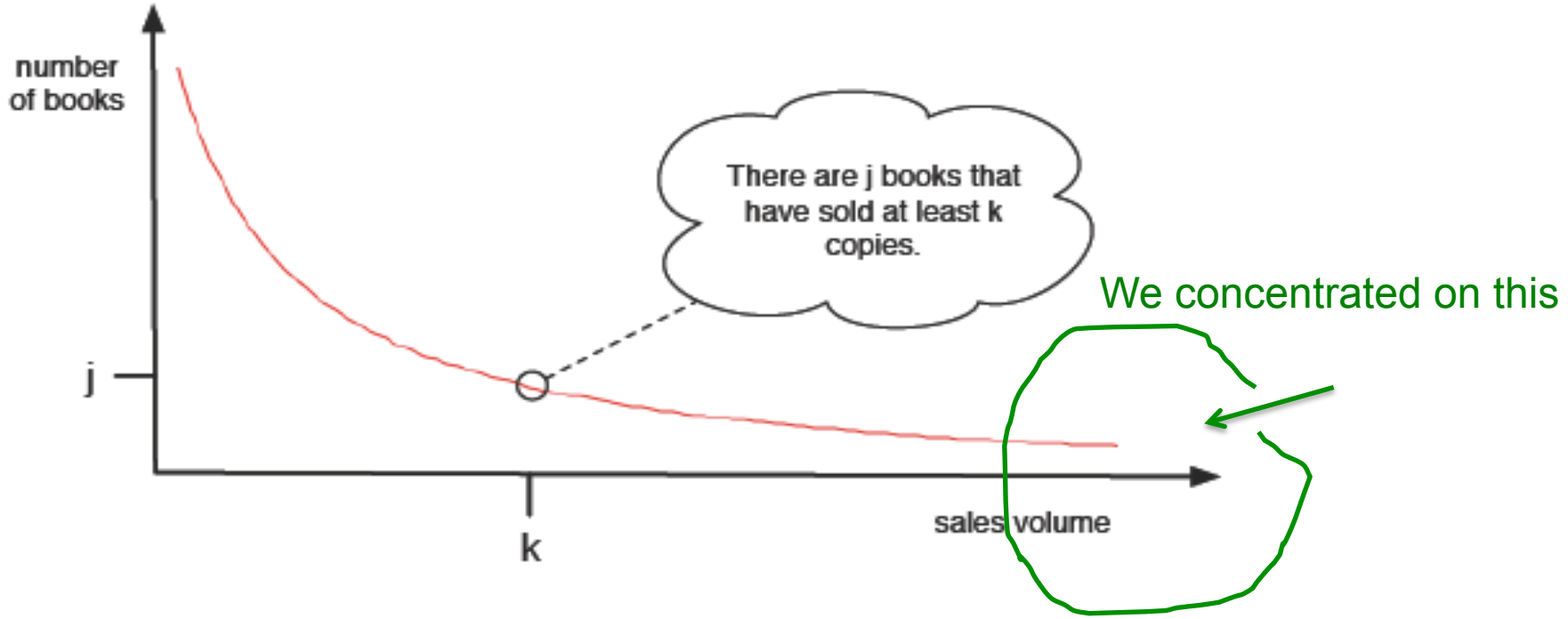
Unpredictability [Salganik et al 06]

- 48 songs, 14,000 participants, 8 servers



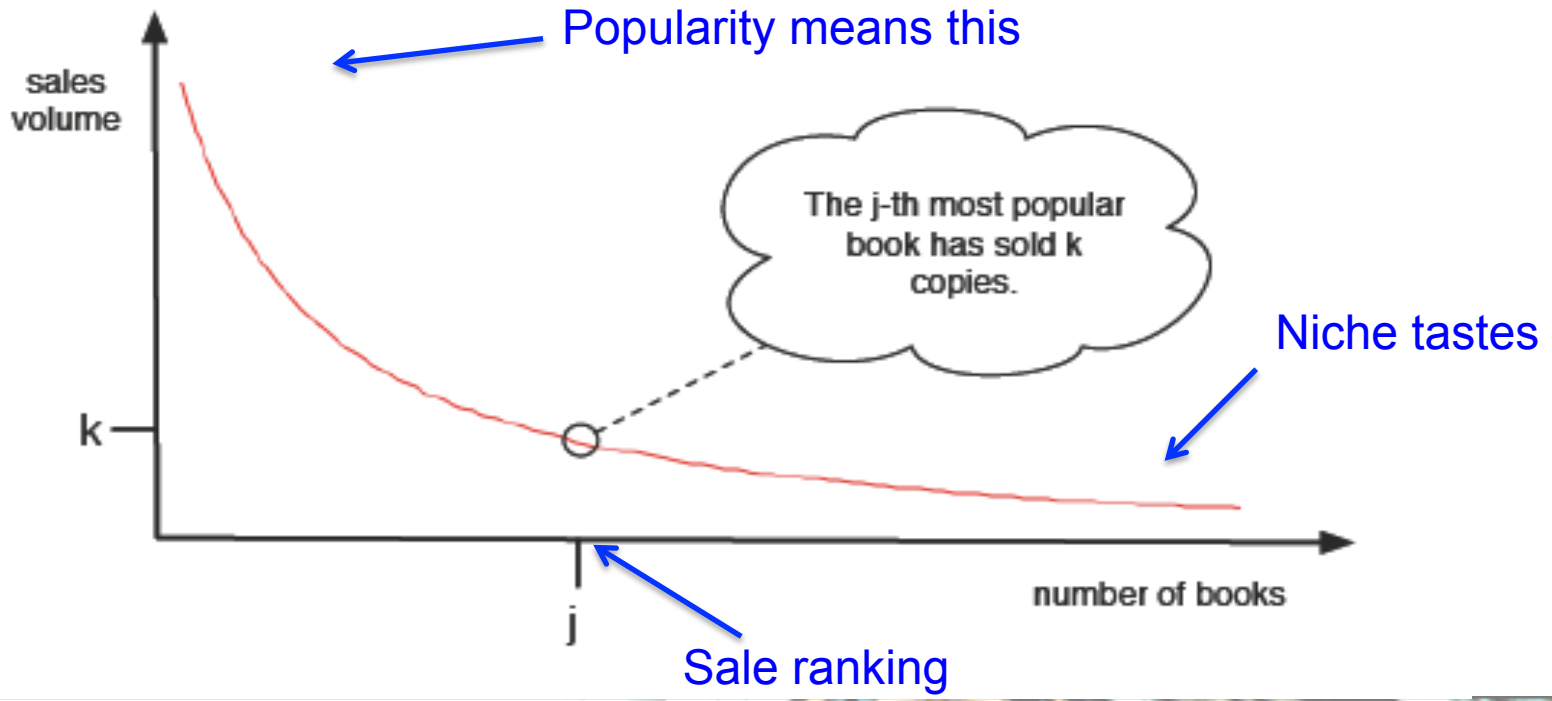
View of the curve

- The way we have seen the curve so far…



Let's transform the function

- If the initial function is a power law, this one is too (we do not prove this)



Search [Ch. 14 Easley & Kleinberg]

- Information retrieval problem: synonyms (jump/leap), polysemy (Leopard), etc
- Now with the web: diversity in authoring introduces issues of common criteria for ranking documents
- The web offers abundance of information: whom do we trust as source?

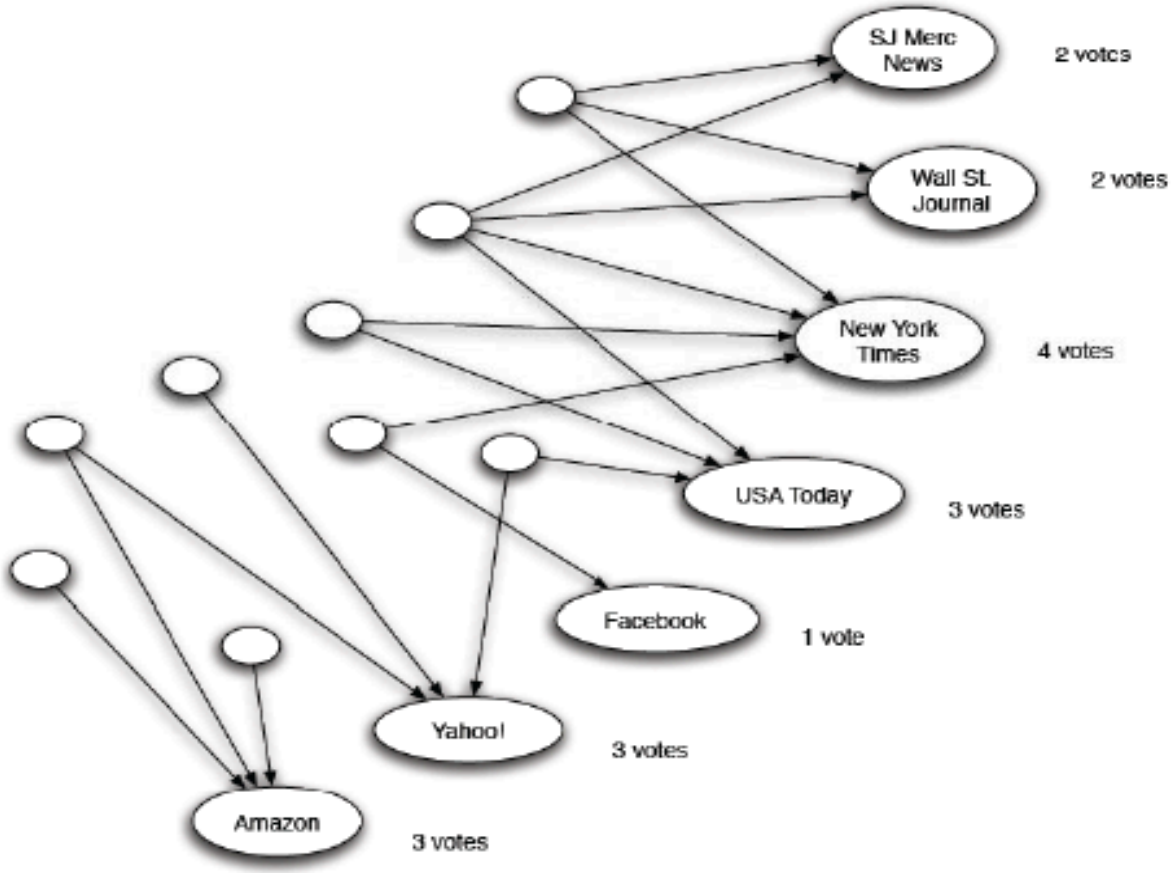
- Still one issue: static content versus real time
  - World trade center query on 11/9/01
  - Twitter helps solving these issues these days

Automate the Search

- When searching "EECS QMUL" on Google the first link is for the department's page.
- How does Google know this is the best answer?
- We could collect a large sample of pages relevant to "EECS QMUL" and collect their votes through their links.
- The pages receiving more in-links are ranked first.
- But if we use **the network structure** more deeply we can improve results.

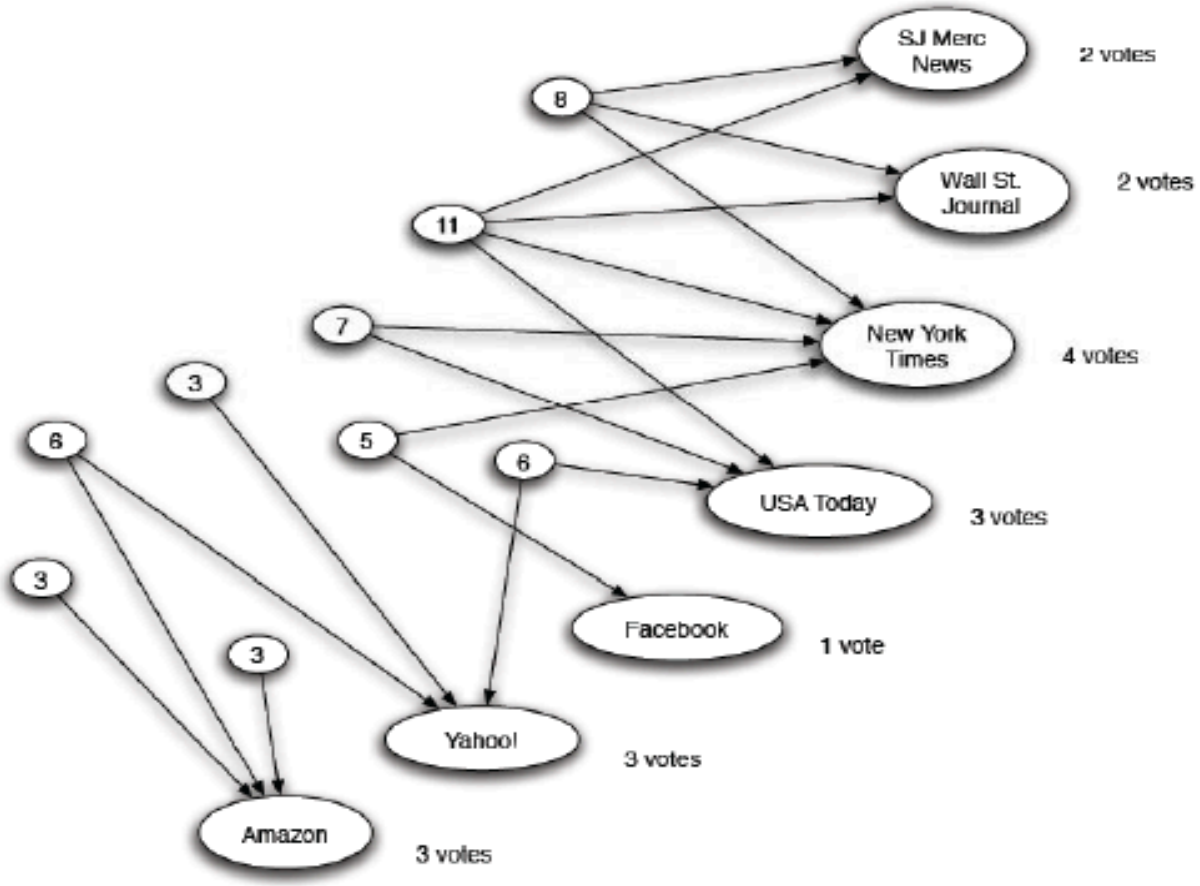
Example: Query "newspaper" Authorities

- Links are seen as votes.
- **Authorities** are established: the highly endorsed pages



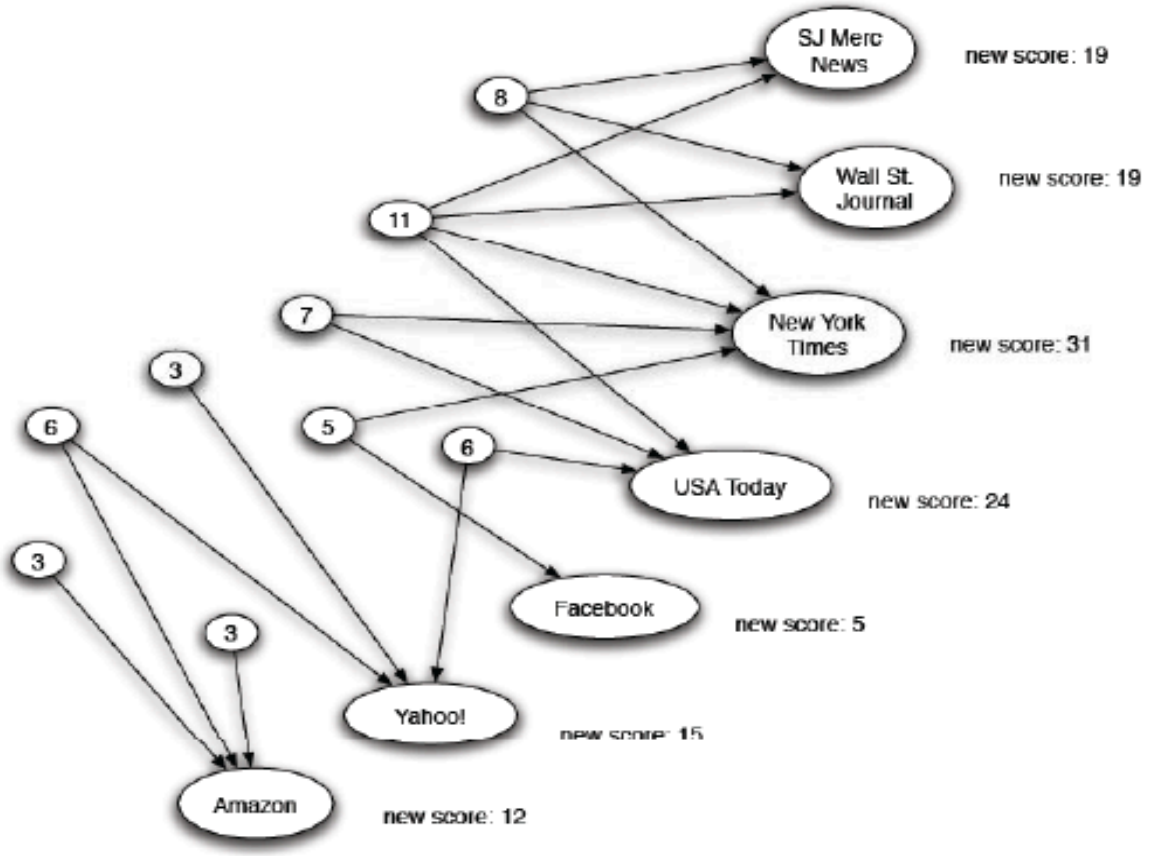
A Refinement: Hubs

- Numbers are reported back on the source page and aggregate.
- Hubs are high value lists



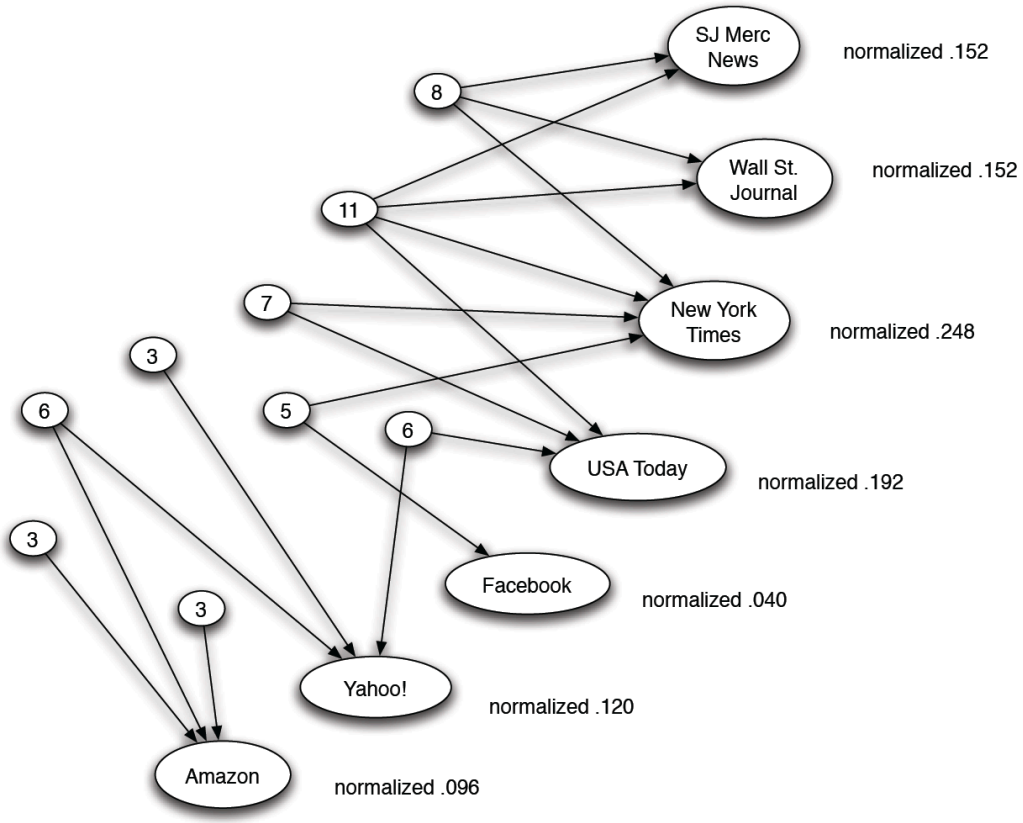
Principle of Repeated Improvement

- And we are now reweighting the authorities
- When do we stop?



Repeating and Normalizing

- The process can be repeated
- Normalization:
  - Each authority score is divided by the sum of all authority scores
  - Each hub score is divided by the sum of all hub scores



More Formally: does the process converge?

- Each page has an authority a_i and a hub h_i score
- Initially $a_i = h_i = 1$
- At each step

$$a_i = \sum_{j->i} h_j$$

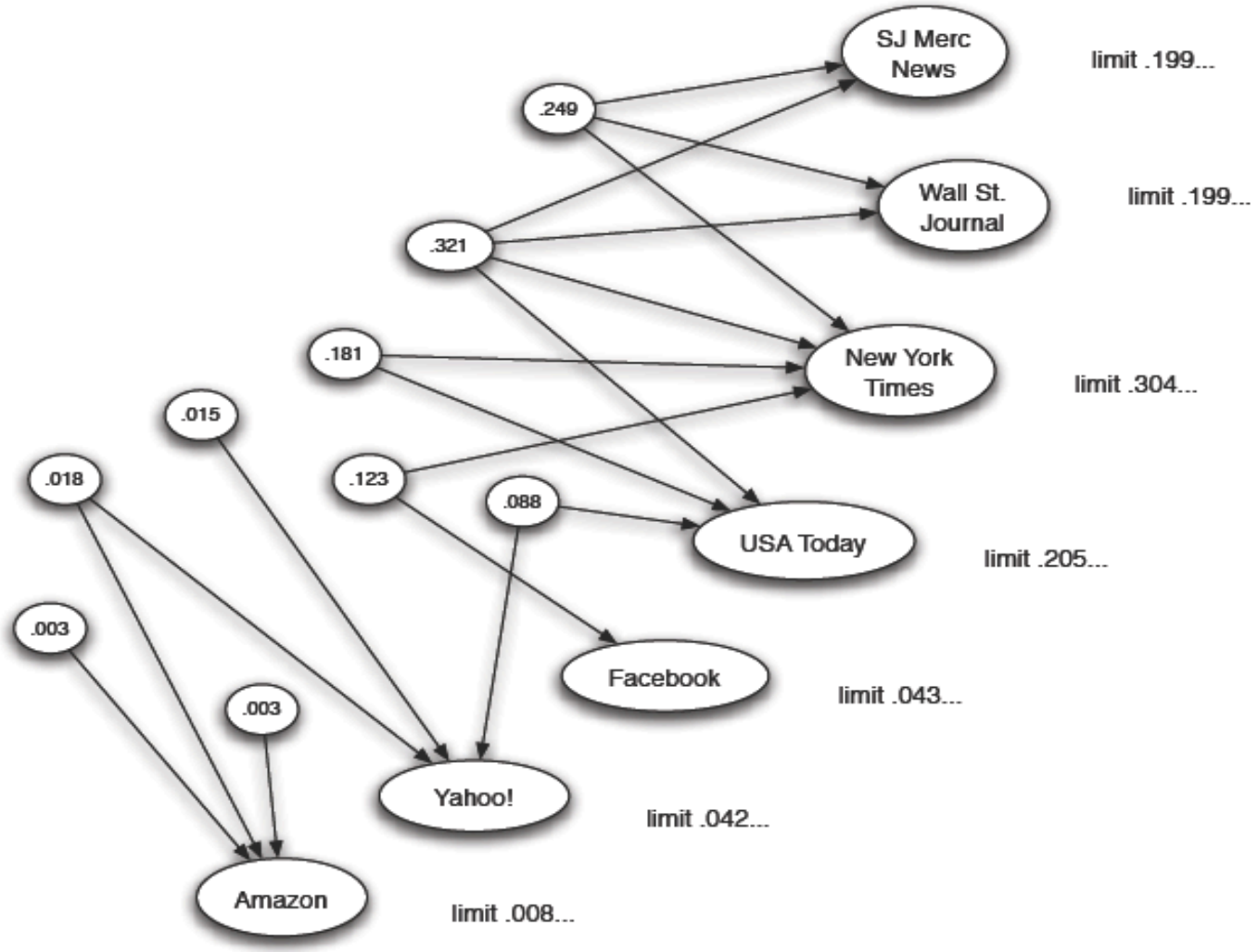
$$h_j = \sum_{j->i} a_i$$

- Normalize

$$\sum a_i = 1$$

$$\sum h_j = 1$$

The process converges



PageRank

- We have seen hubs and authorities
  - Hubs can “collect” links to important authorities who do not point to each others
  - There are other models: better for the web, where one prominent can endorse another.
- The **PageRank** model is based on transferrable importance.

PageRank Concepts

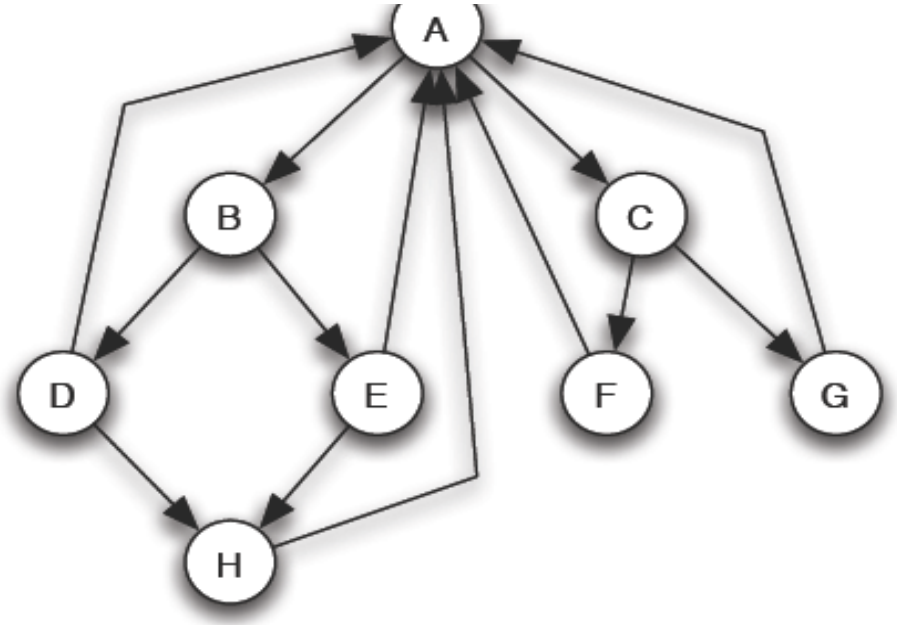
- Pages pass endorsements on outgoing links as fractions which depend on out-degree
- Initial PageRank value of each node in a network of n nodes: 1/n.
- Choose a number of steps k.
- **[Basic] Update rule**: each page divides its pagerank equally over the outgoing links and passes an equal share to the pointed pages. Each page's new rank is the sum of received pageranks.

Example

- All pages start with PageRank= 1/8

Step	A	B	C	D	E	F	G	H
1	1/2	1/16	1/16	1/16	1/16	1/16	1/16	1/8
2	3/16	1/4	1/4	1/32	1/32	1/32	1/32	1/16

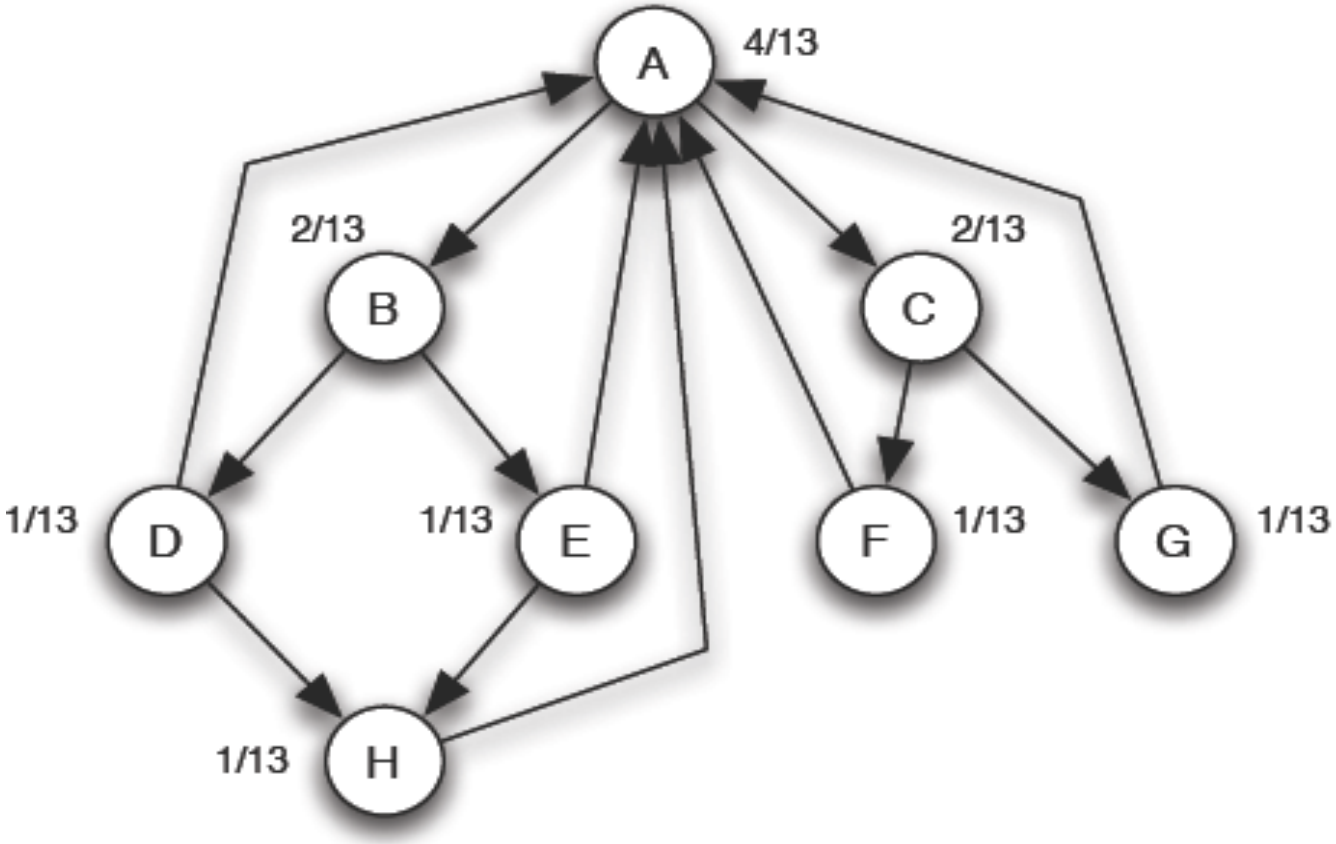
A becomes important and
B,C benefit too at step 2



Convergence

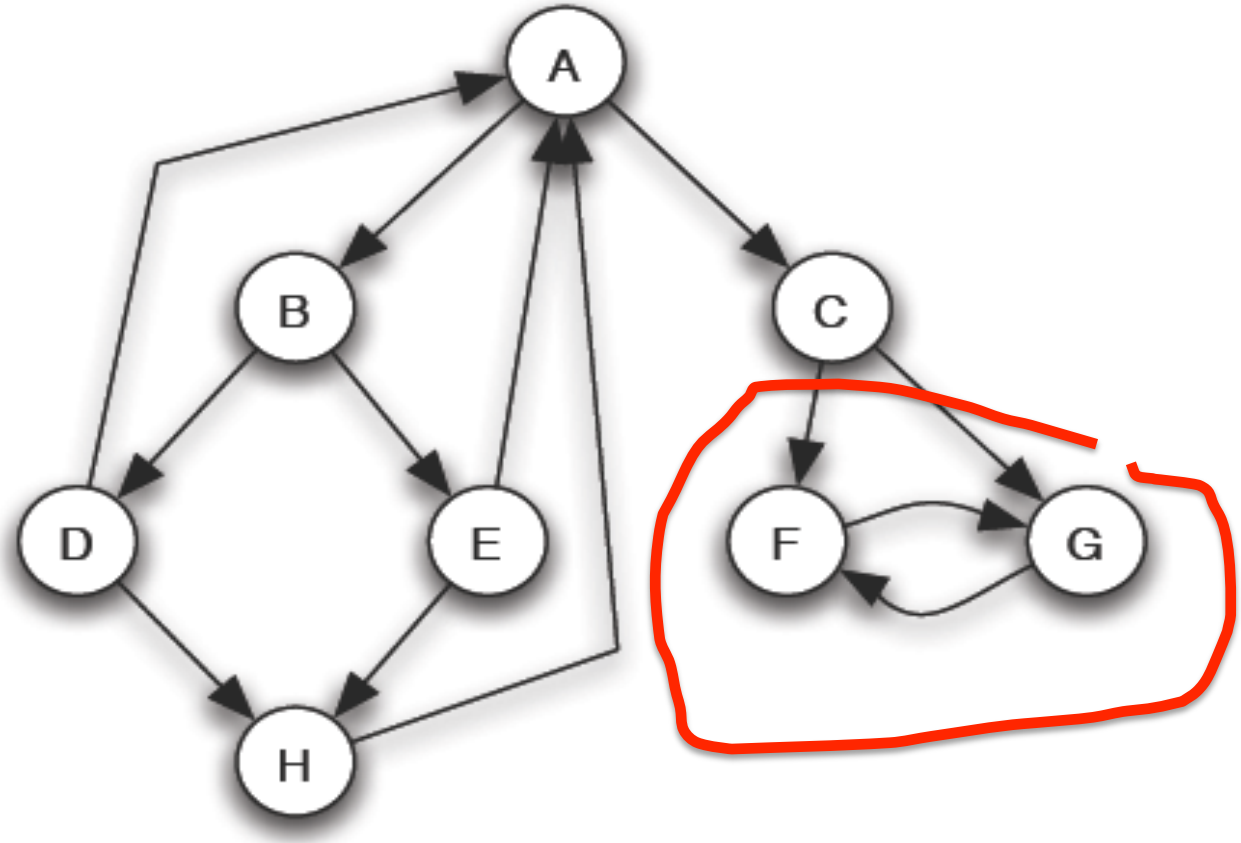
- Except for some special cases, PageRank values of all nodes converge to limiting values when the number of steps goes to infinity.
- The convergence case is one where the PageRank of each page does not change anymore, i.e., they regenerate themselves.

Example of Equilibrium



Problems with the basic PageRank Dead ends

- F,G converge to ½ and all the other nodes to 0



Solution: The REAL PageRank

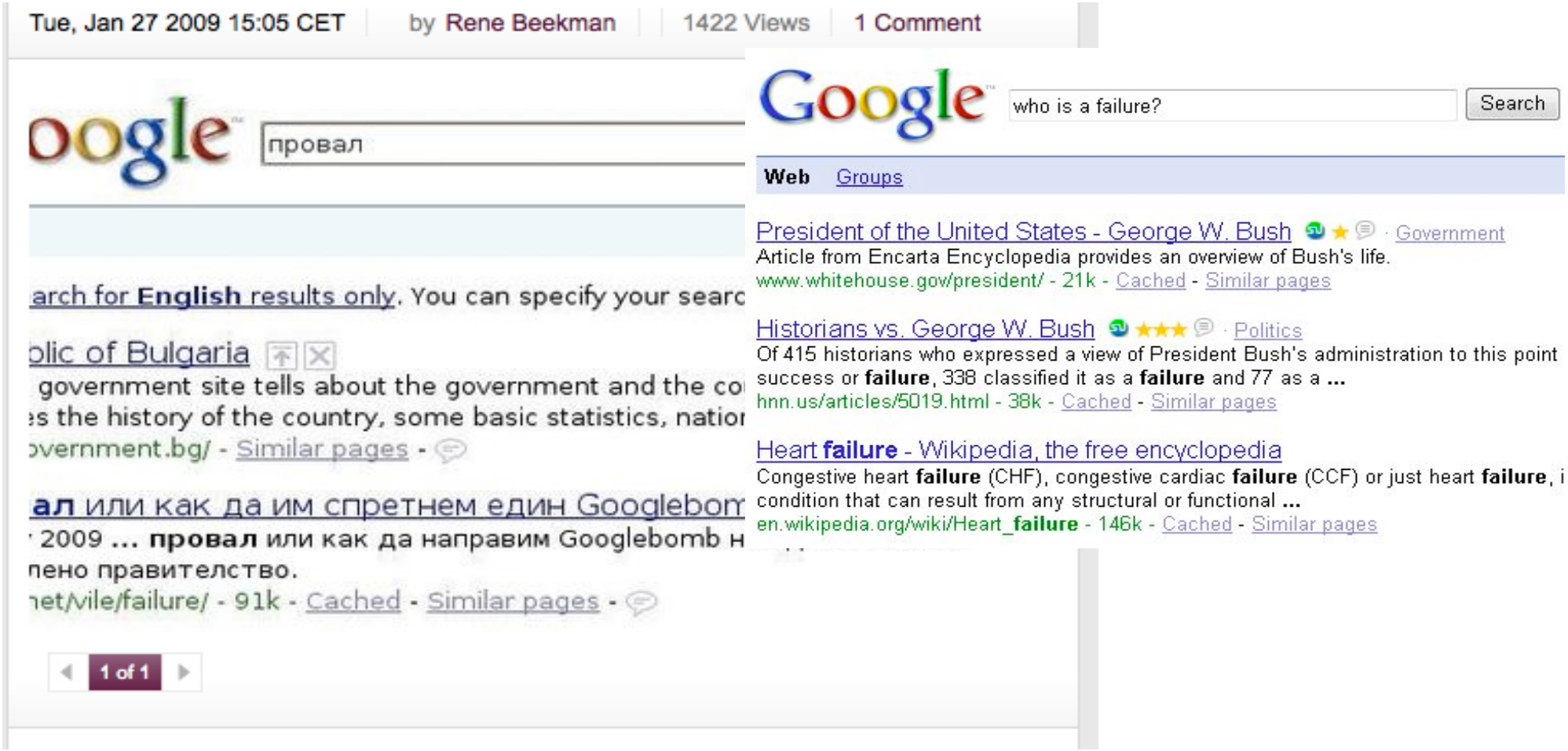
- **[Scaled] Update Rule:**
  - Apply basic update rule. Then, scale down all values by scaling factor s [chosen between 0 and 1].
  - **[Total network PageRank value changes from 1 to s]**
  - Divide 1-s residual units of PageRank equally over all nodes: (1-s)/n each.

- It can be proven that values converge again.
- Scaling factor usually chosen between 0.8 and 0.9

Search Ranking is very important to business

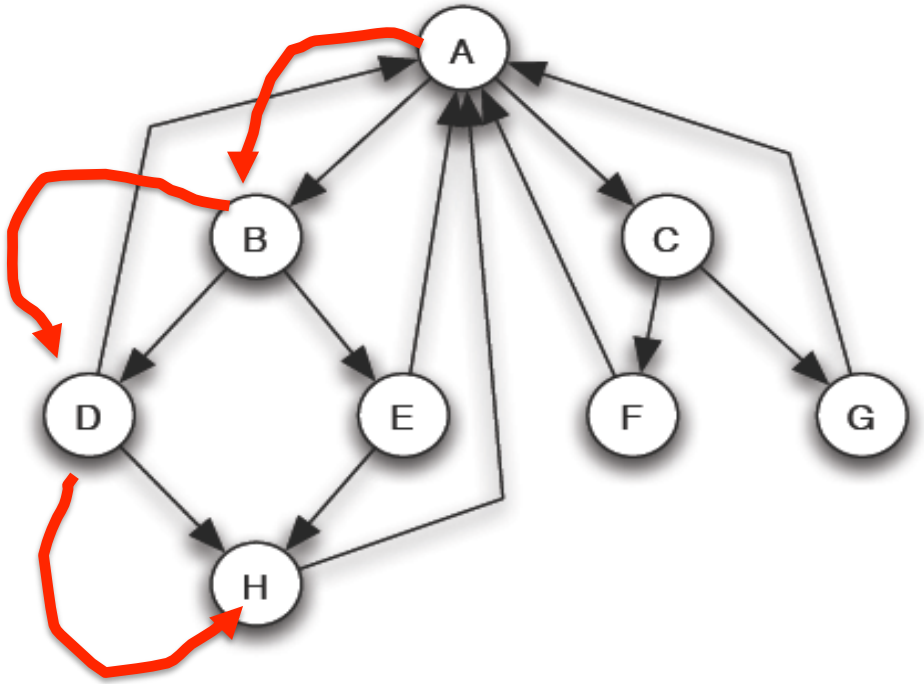
- A change in results in the search pages might mean loss of business
  - I.e., not appearing on first page.
- Ranking algorithms are kept very secret and changed continuously.

Examples of Google Bombs



Random Walks

- Starting from a node, follow one outgoing link with an equal probability



PageRank as Random Walk

- The probability of being at a page X after k steps of a random walk is precisely the PageRank of X after k applications of the Basic PageRank Update Rule
- Scaled Update Rule equivalent: follow a random outgoing link with probability s while with probability 1-s jump to a random node in the network.

References

- Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the Web. In Proc. 9th International World Wide Web Conference, pages 309-320,2000.
- A. Clauset, C. R. Shalizi and M. E. J. Newman, 2009. “Power-law distributions in empirical data.” SIAM Review Vol. 51, No. 4. (2 Feb 2009), 661.
- Barabási, Albert-László and Réka Albert, "Emergence of scaling in random networks", *Science*, 286:509-512, October 15, 1999
- Matthew Salganik, Peter Dodds, and Duncan Watts. Experimental study of inequality and unpredictability in an artificial cultural market. Science, 311:854-856, 2006.