Social and Technological Network Data Analytics

Lecture 5: Structure of the Web, Search and Power Laws

Prof Cecilia Mascolo

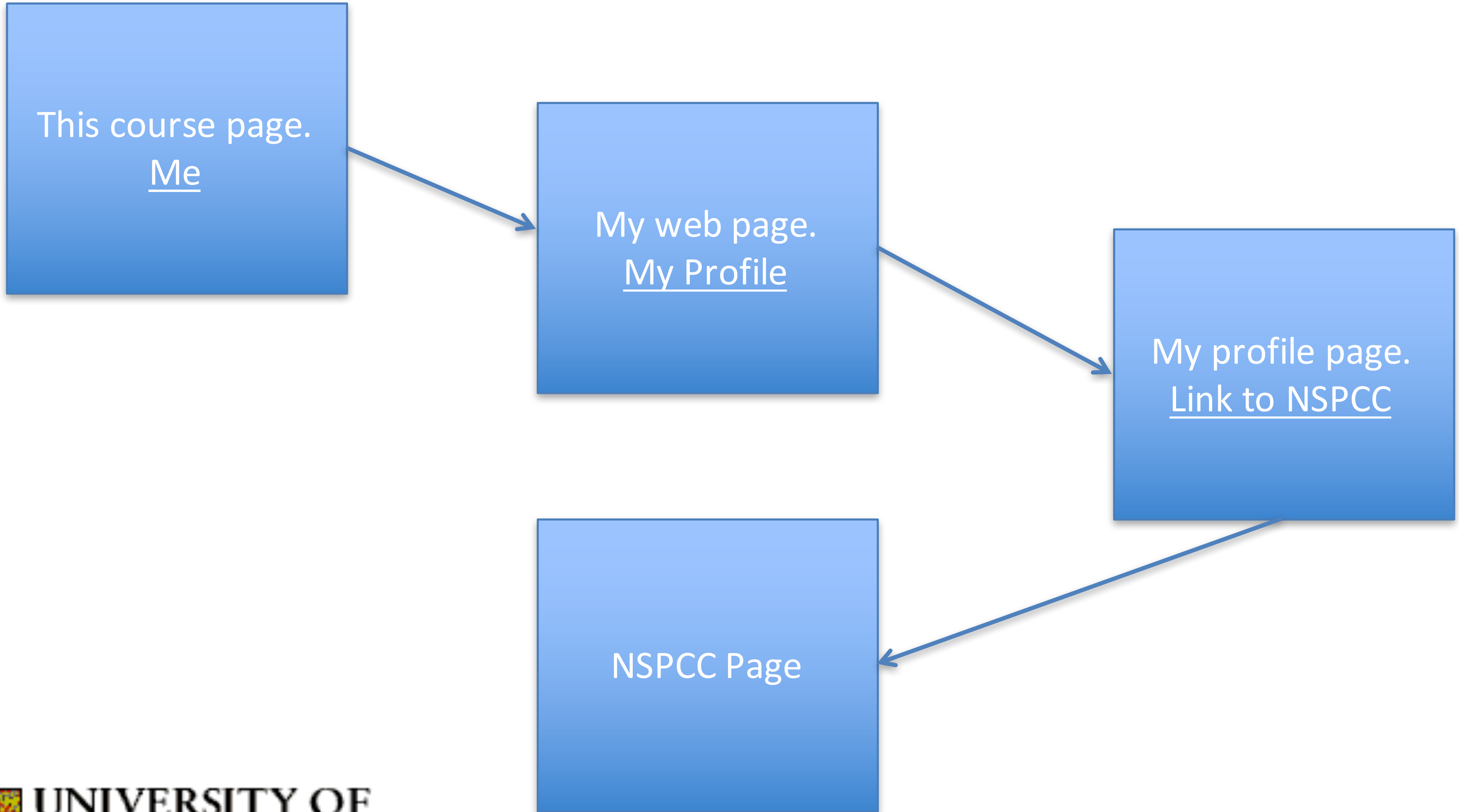
UNIVERSITY OF CAMBRIDGE

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

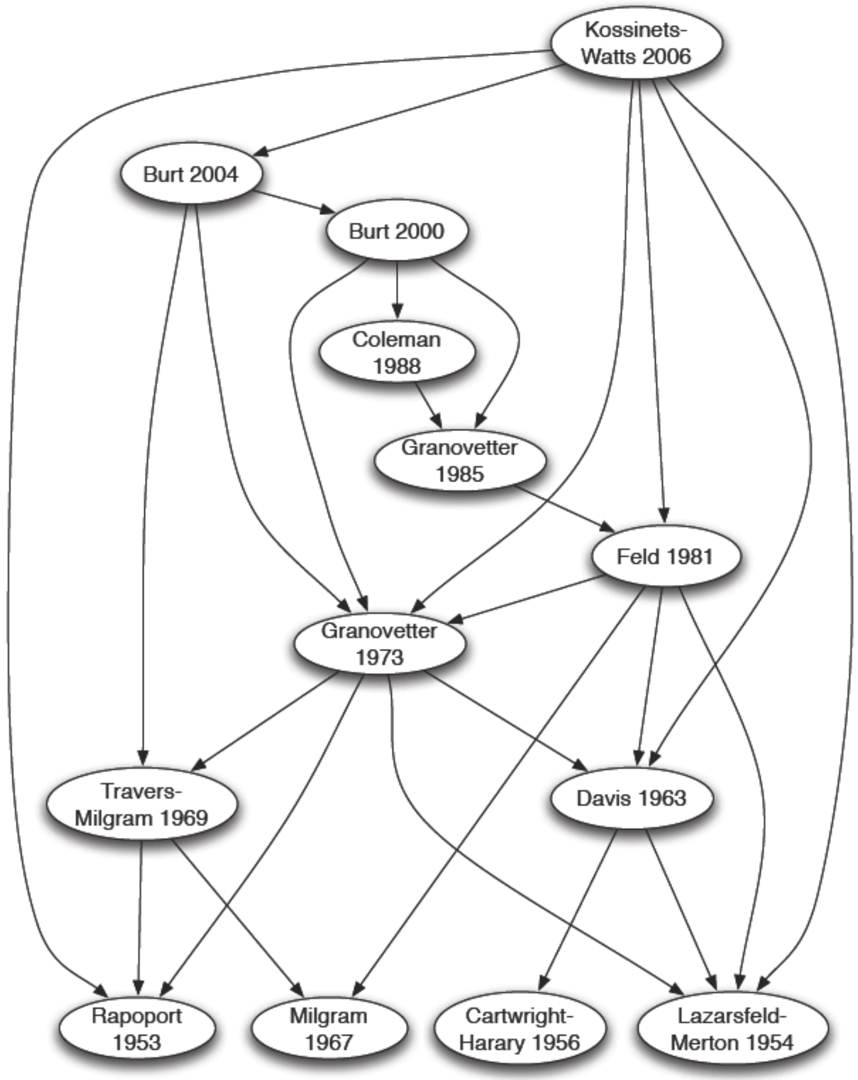
UNIVERSITY OF CAMBRIDGE

The Web is a Graph…



Precursor of hypertexts

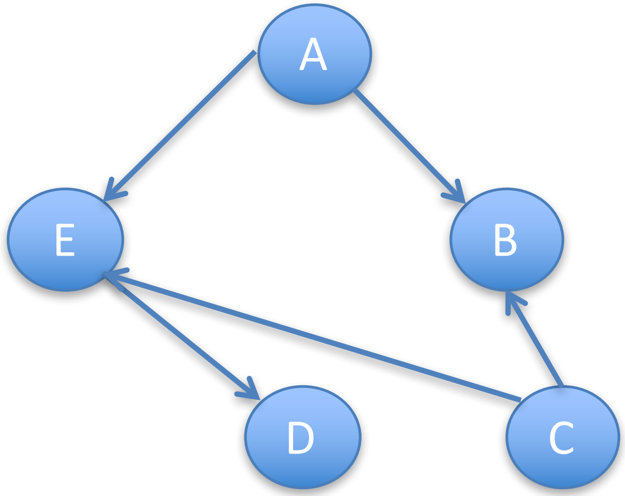
- Citation networks of books and articles.
- Difference: links point only backwards in time



UNIVERSITY OF CAMBRIDGE

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

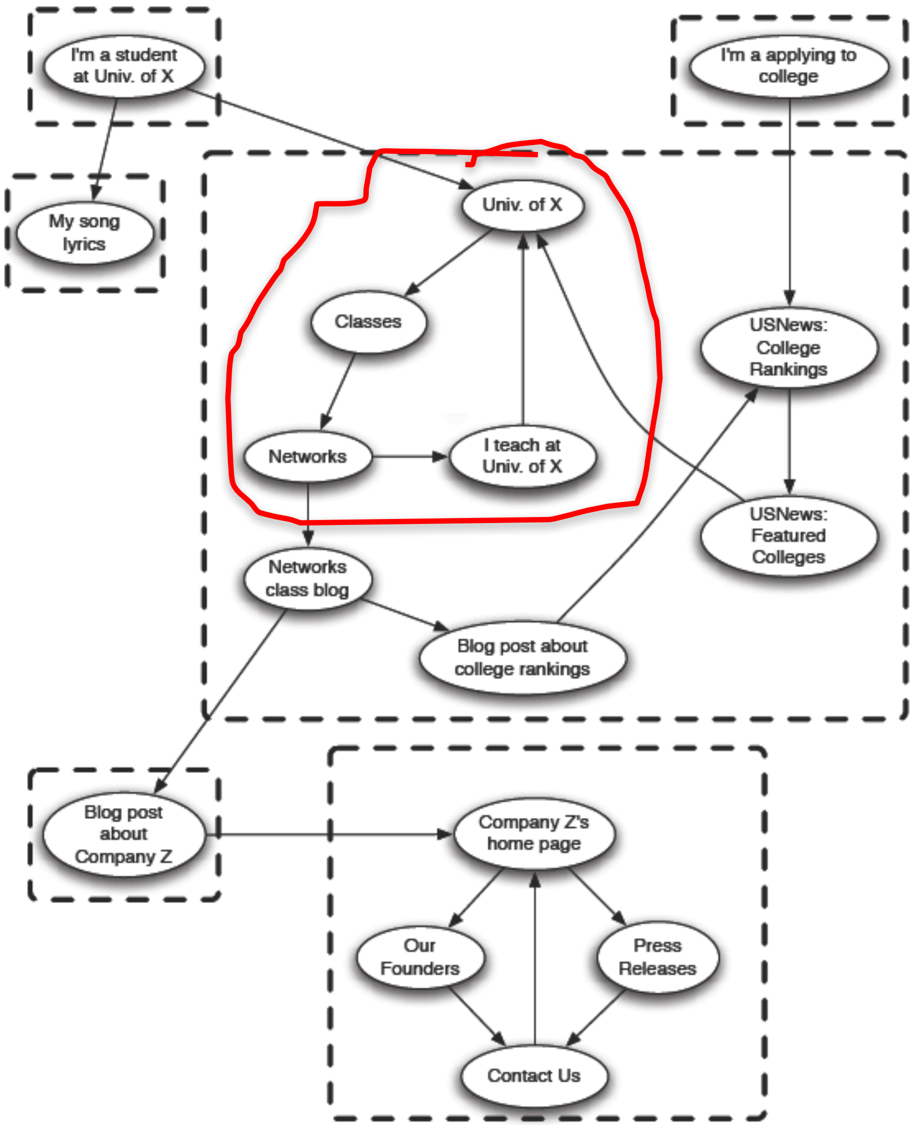
- A strongly connected component (SCC) in a directed graph is a subset of nodes such that:

  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.
  – Two nodes can be in the same WCC even if there no directed path between them.

UNIVERSITY OF CAMBRIDGE

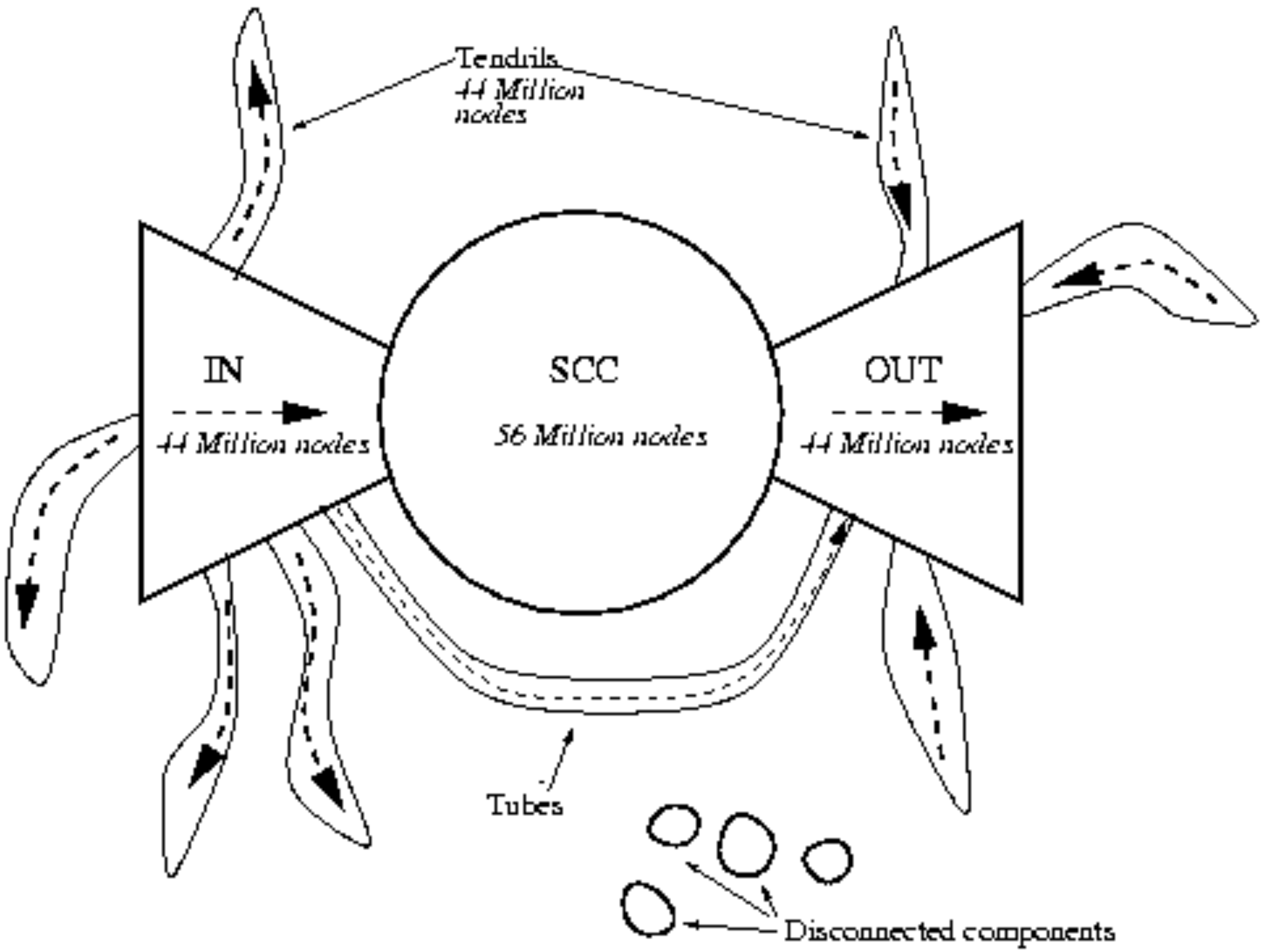
SCC example



UNIVERSITY OF CAMBRIDGE

The Web

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



UNIVERSITY OF CAMBRIDGE

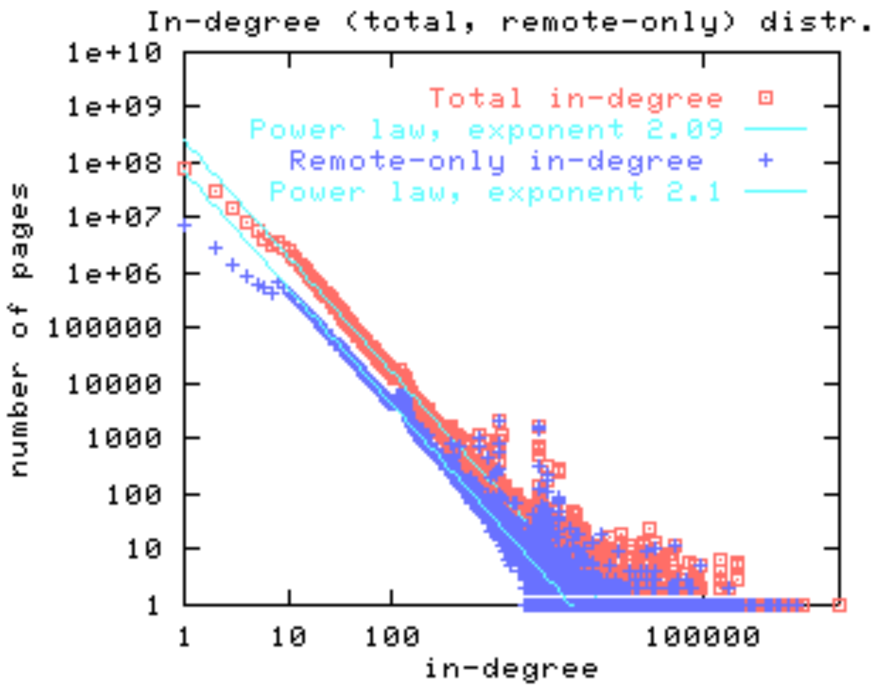
Popularity of Web Pages

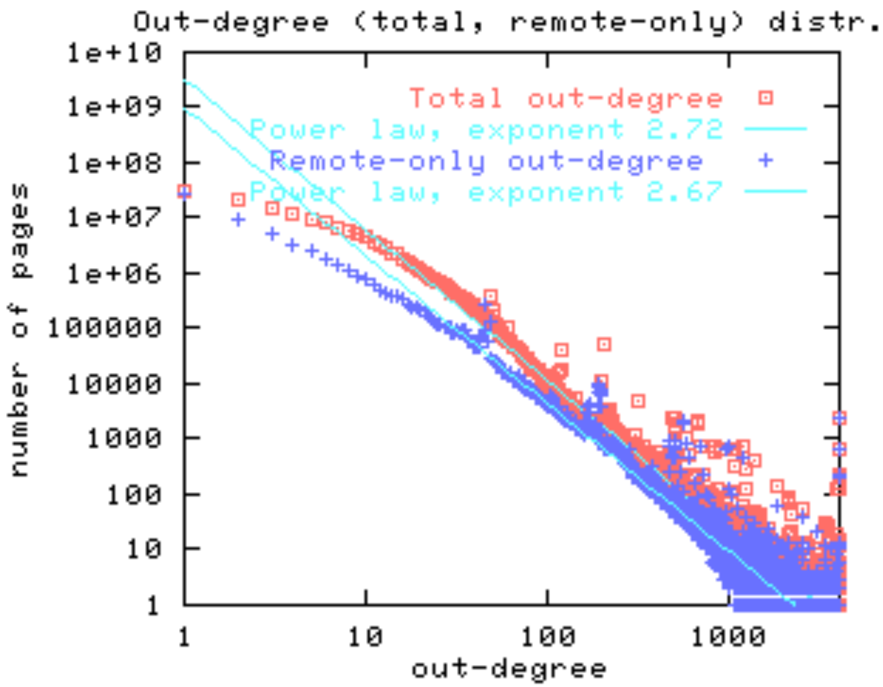
- How do we expect the popularity of web pages to be distributed?
  - What fraction of web pages have k in-links?
  - If each page decides independently at random whether to link to any given other page then the n of in-links of a page is the sum of independent random quantities -> normal distribution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

UNIVERSITY OF CAMBRIDGE

Degree distribution for the Web

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



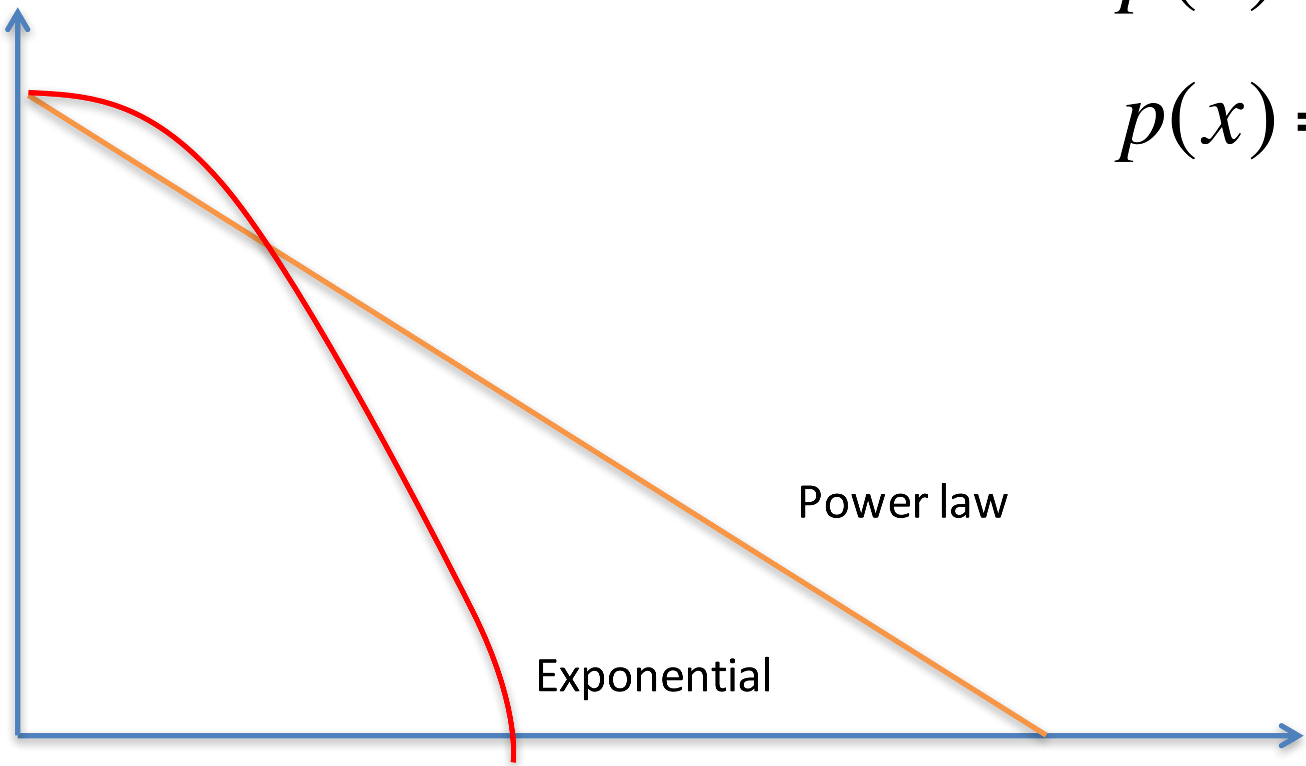


UNIVERSITY OF CAMBRIDGE

Power Law vs Exponential

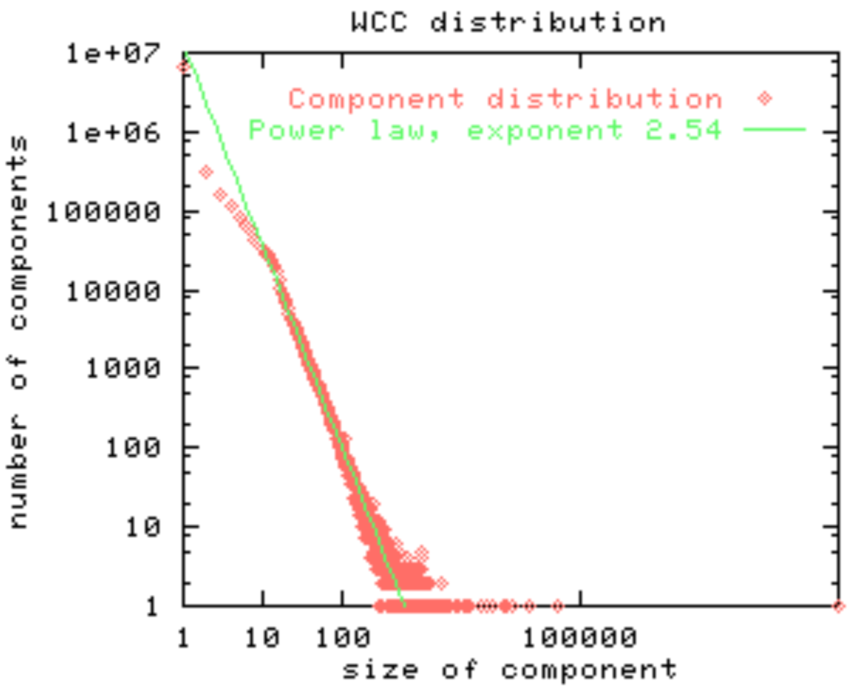
$$p(x) = x^{-\alpha}$$

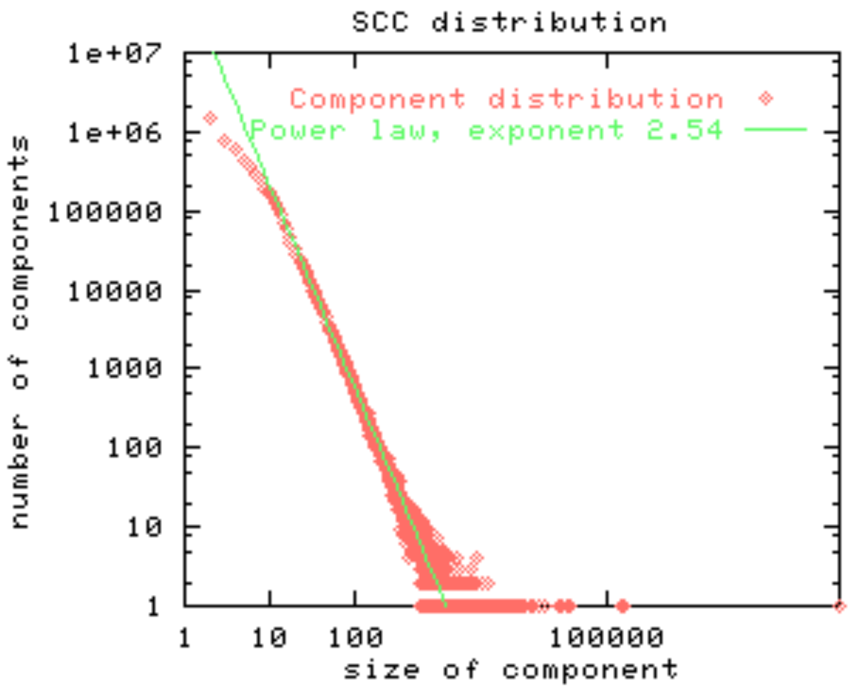
$$p(x) = e^{-\lambda x}$$



UNIVERSITY OF
CAMBRIDGE

Distribution of WCC and SCC

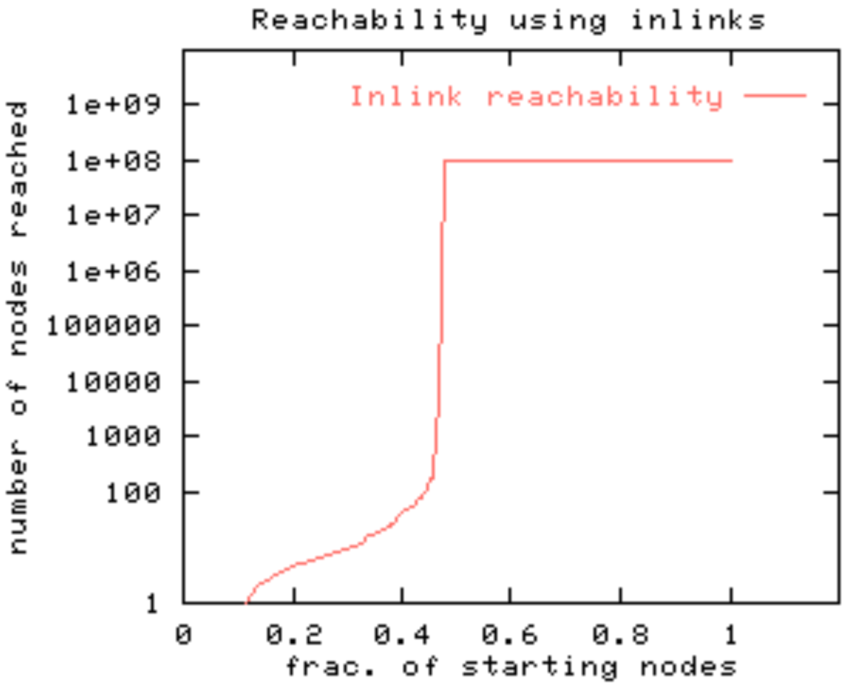


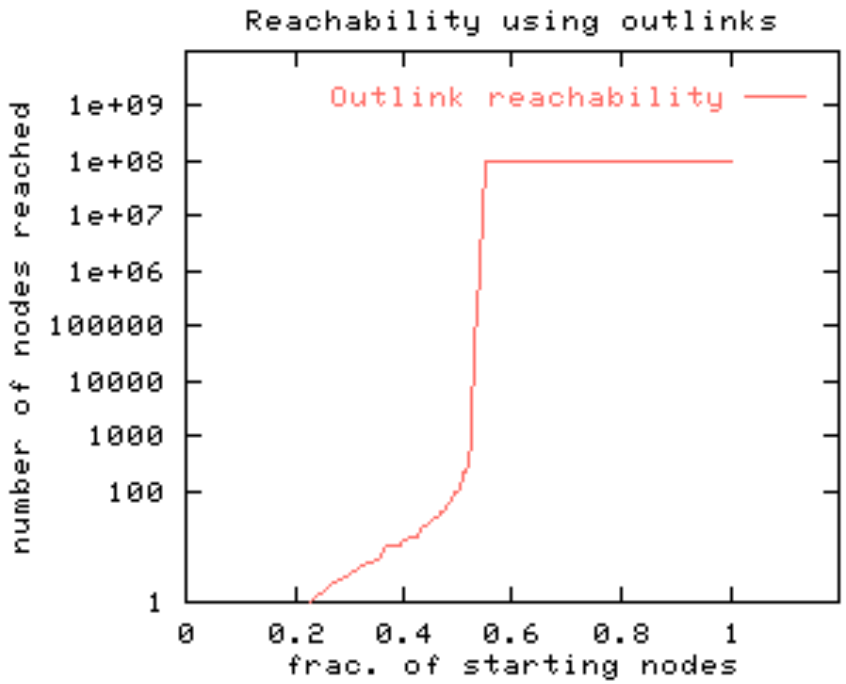


UNIVERSITY OF CAMBRIDGE

Reachability

- Followed links backwards and forward





UNIVERSITY OF CAMBRIDGE

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

Power Laws aka Scale Free Networks

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

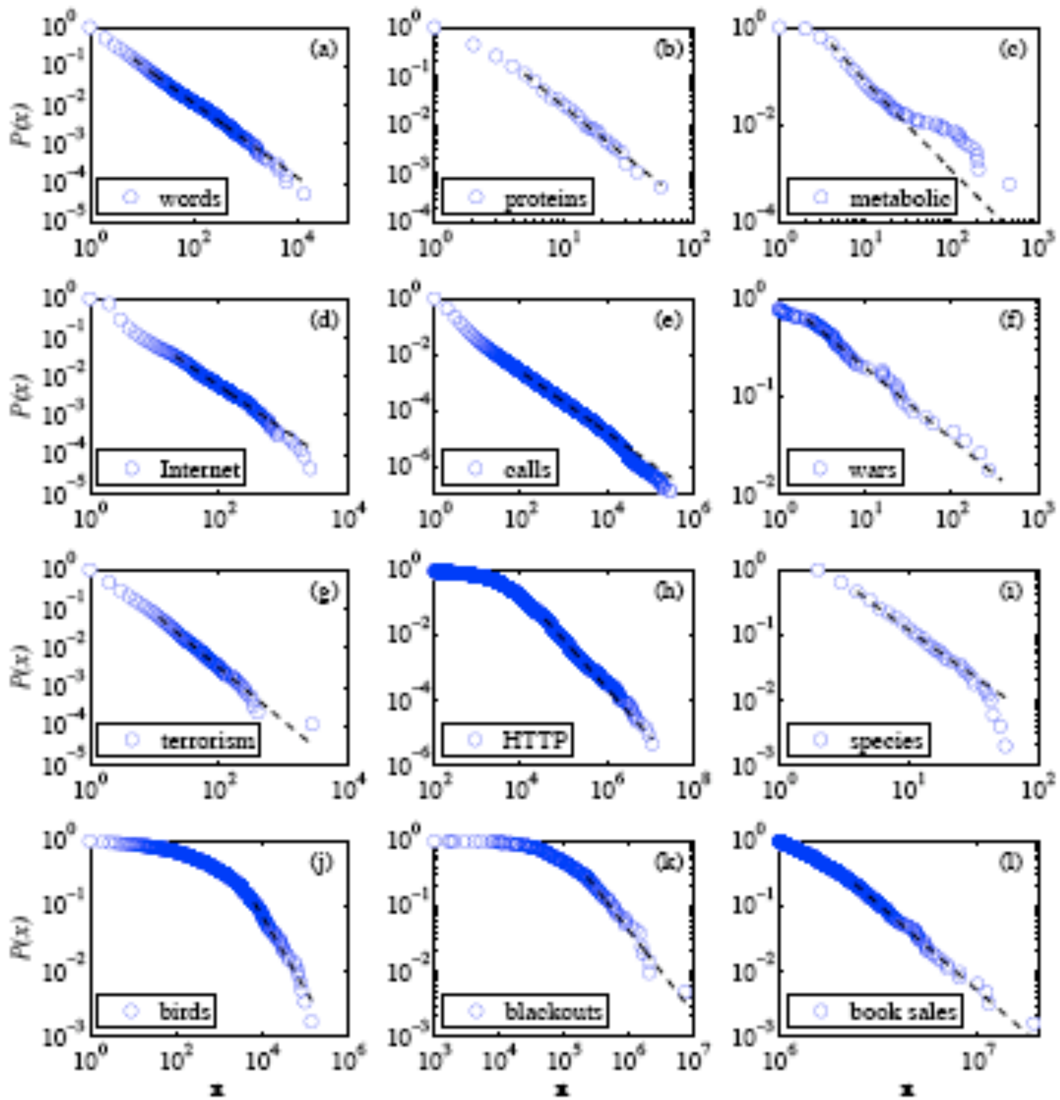
$$\ln p_k = -\alpha \ln k + c$$

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

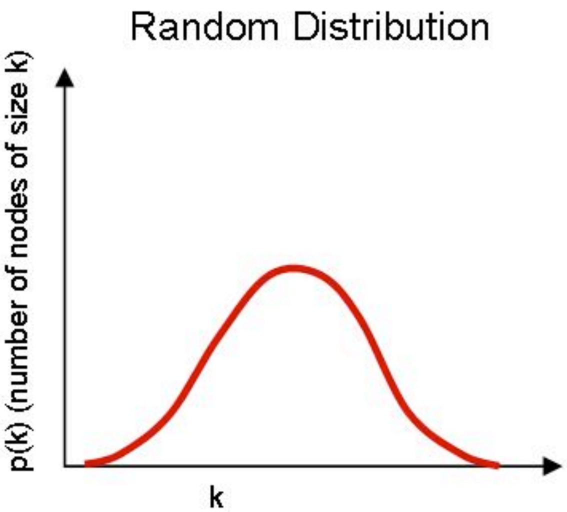
- α defines the slope of the curve
- α is typically between 2 and 3.

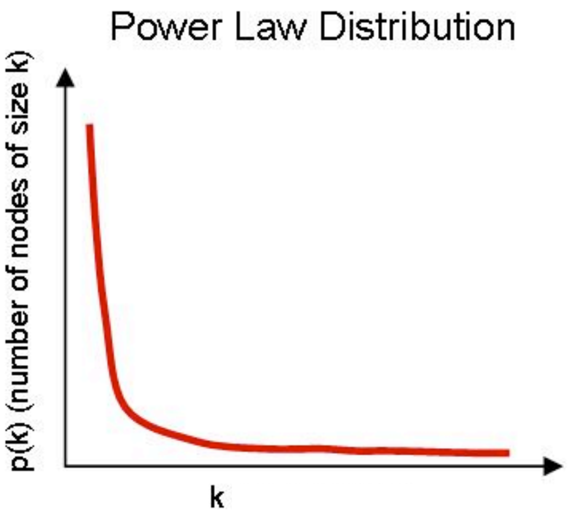
UNIVERSITY OF CAMBRIDGE

Power Laws in various domains



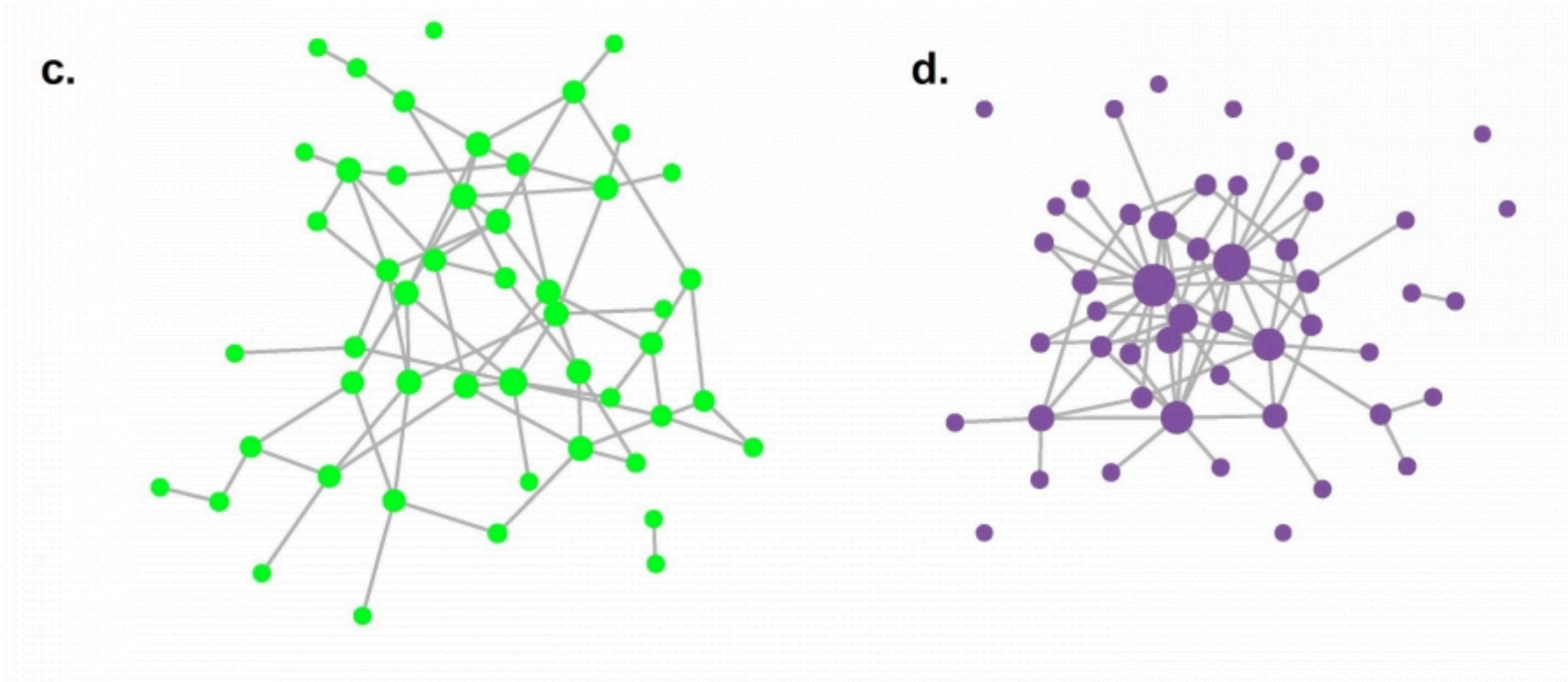
What does it mean?





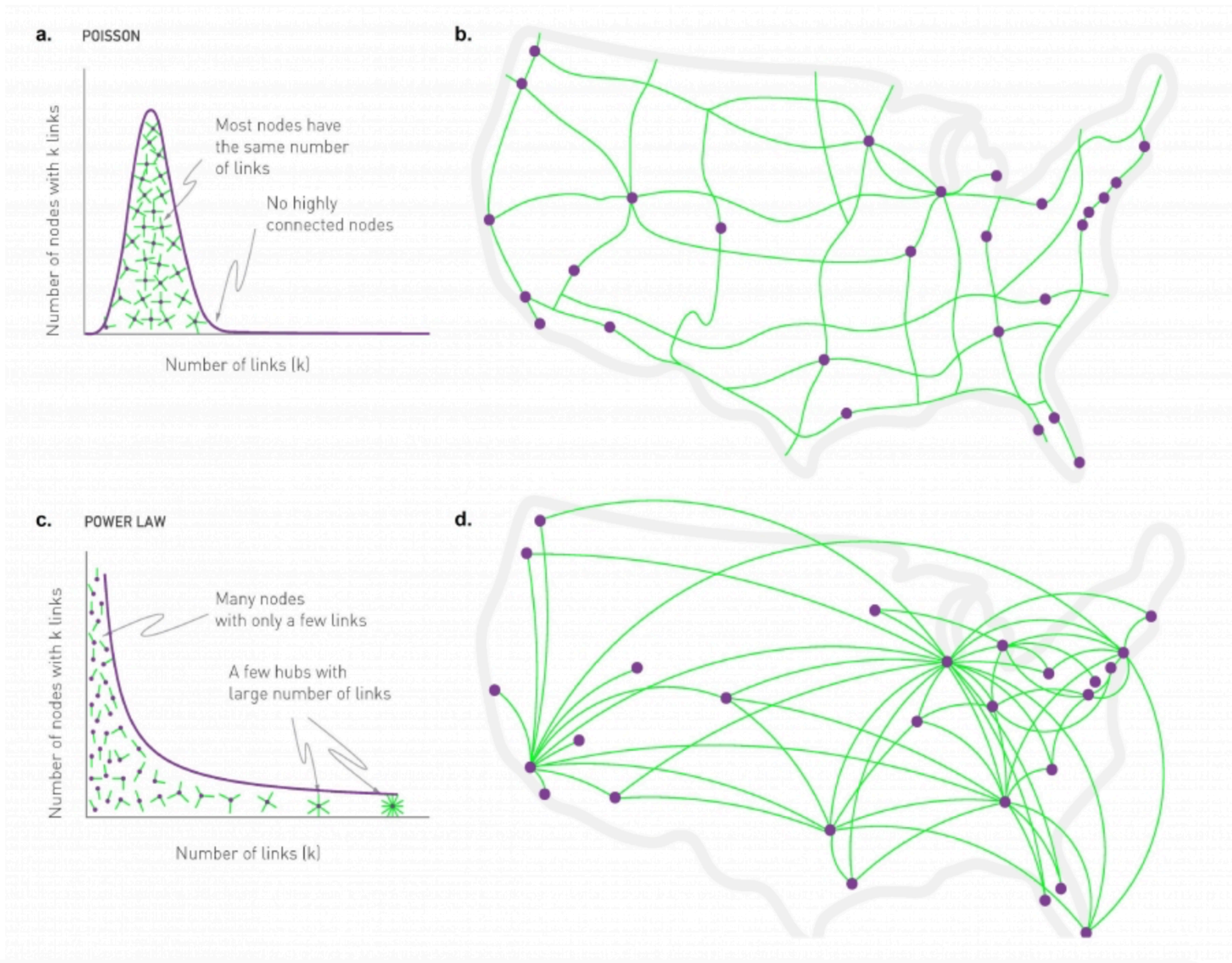
UNIVERSITY OF CAMBRIDGE

Random vs Power Law Networks



UNIVERSITY OF CAMBRIDGE

Example



UNIVERSITY OF CAMBRIDGE

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 at random **and links to the page i points to**.
  - Repeat.
- The middle step is essentially a copy of the node i behaviour…

UNIVERSITY OF CAMBRIDGE

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

UNIVERSITY OF CAMBRIDGE

Intuition

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

- 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

UNIVERSITY OF CAMBRIDGE

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.

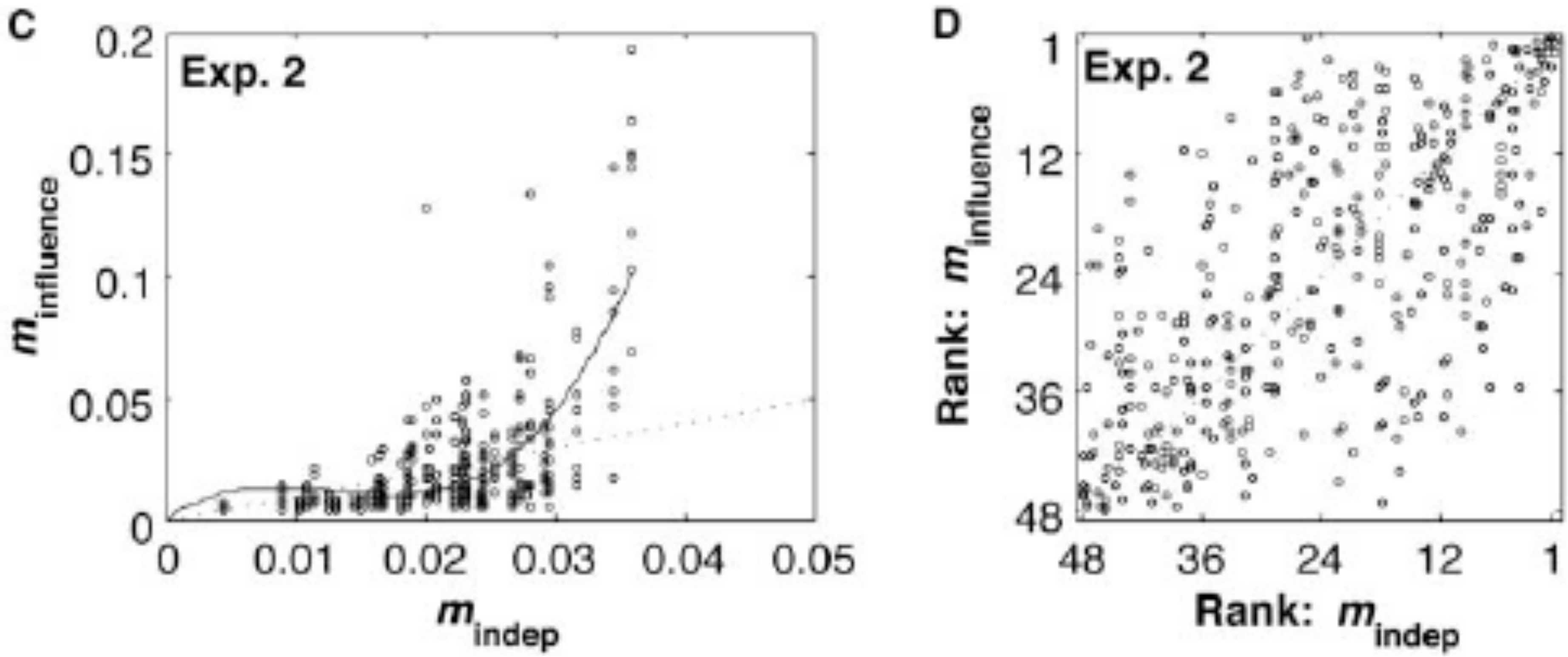
UNIVERSITY OF CAMBRIDGE

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…

Unpredictability [Salganik et al 06]

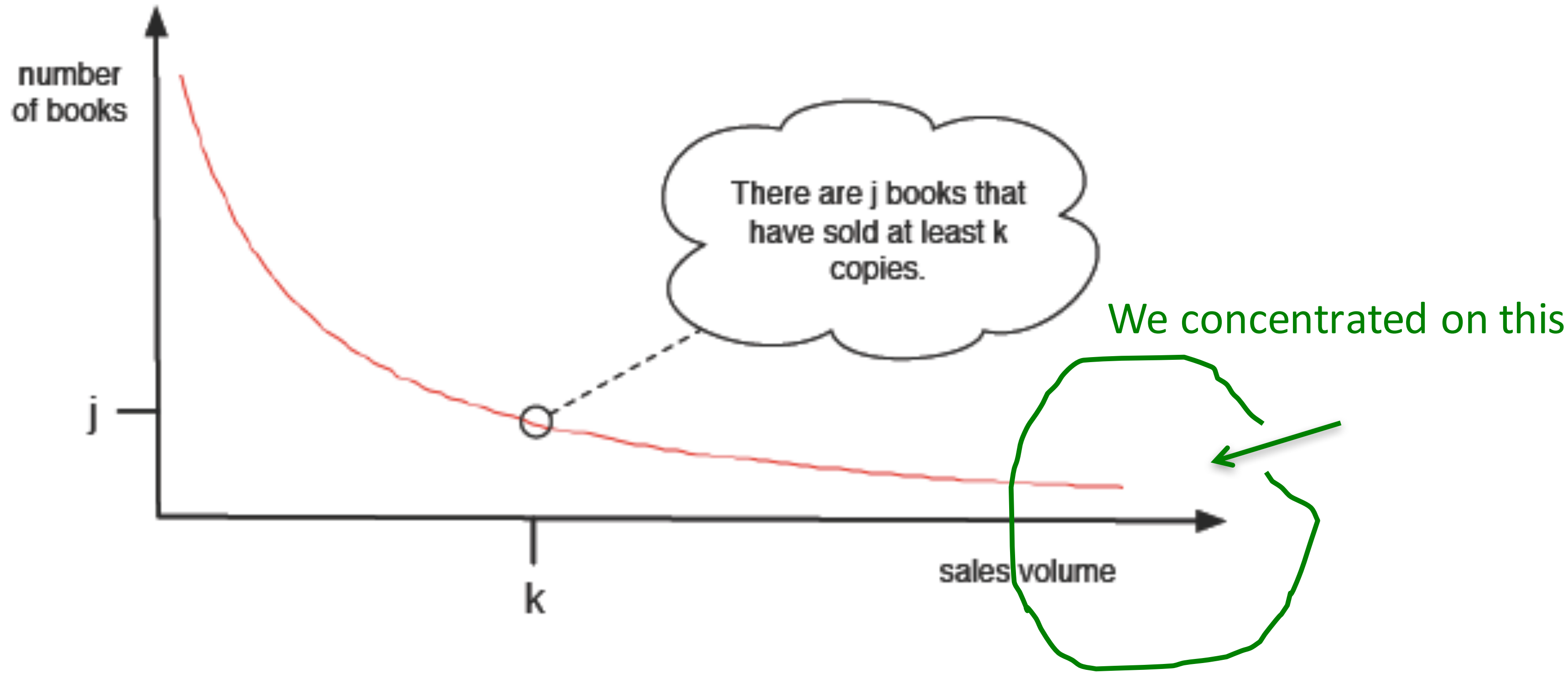
- 48 songs, 14,000 participants, 8 servers



UNIVERSITY OF CAMBRIDGE

View of the curve

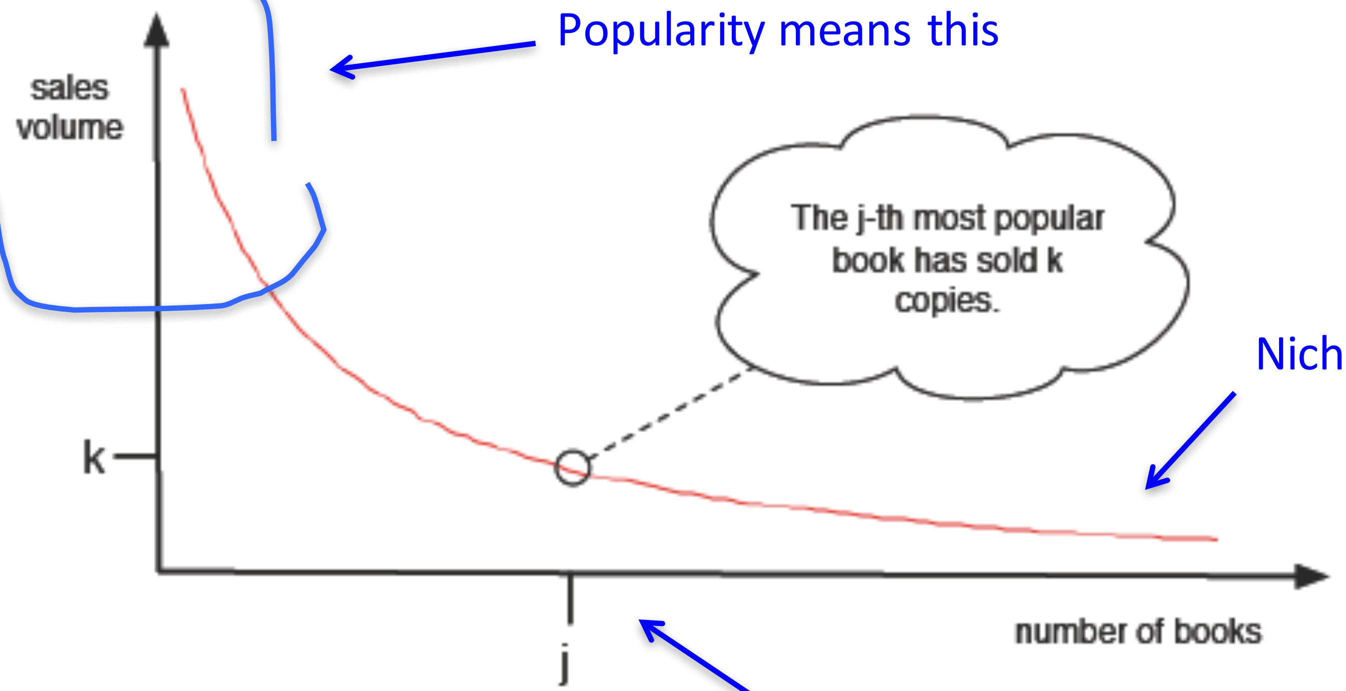
- The way we have seen the curve so far…



UNIVERSITY OF CAMBRIDGE

Let’s transform the function

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



Popularity means this

Niche tastes

Sale ranking

Search

- – 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

UNIVERSITY OF CAMBRIDGE

Automate the Search

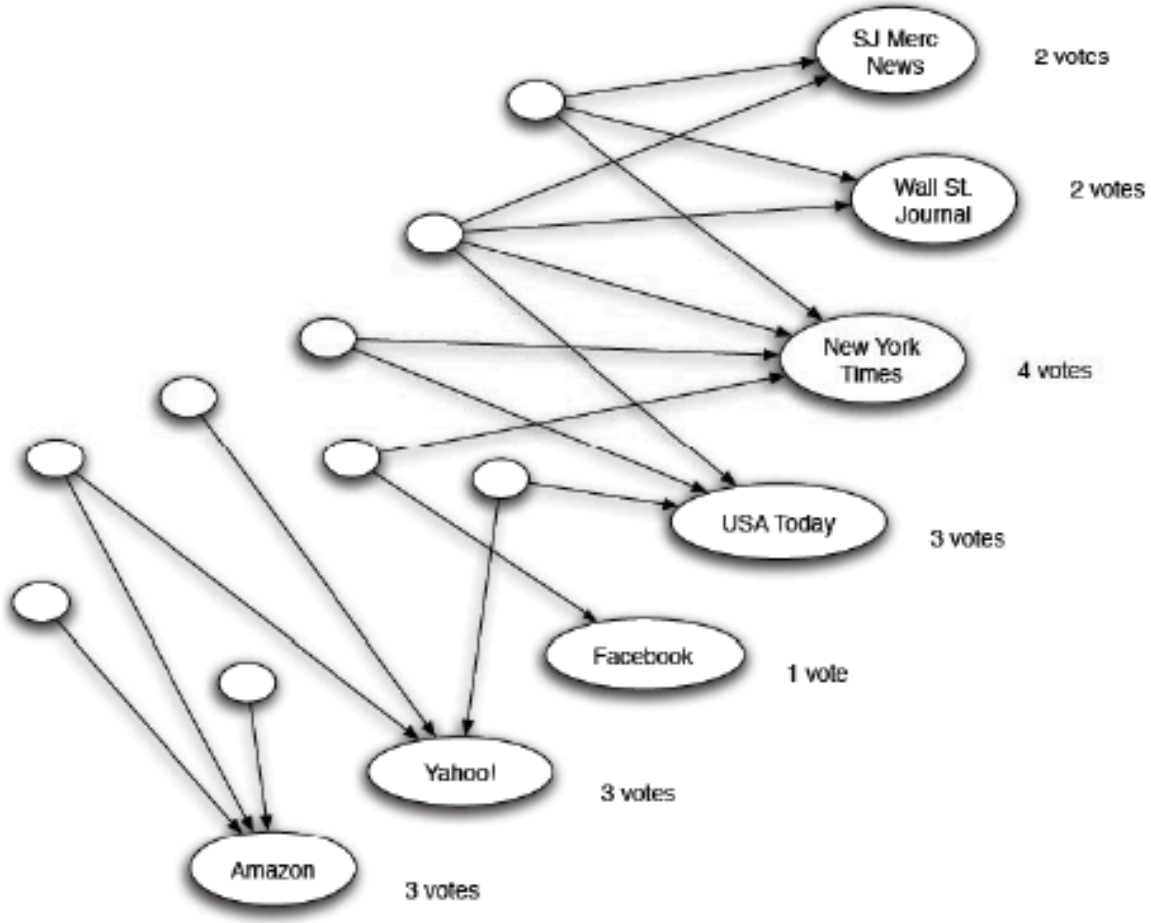
- When searching "Computer Laboratory" on Google the first link is for the department's page.

- How does Google know this is the best answer?
- We could collect a large sample of pages relevant to "computer laboratory" 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.

UNIVERSITY OF CAMBRIDGE

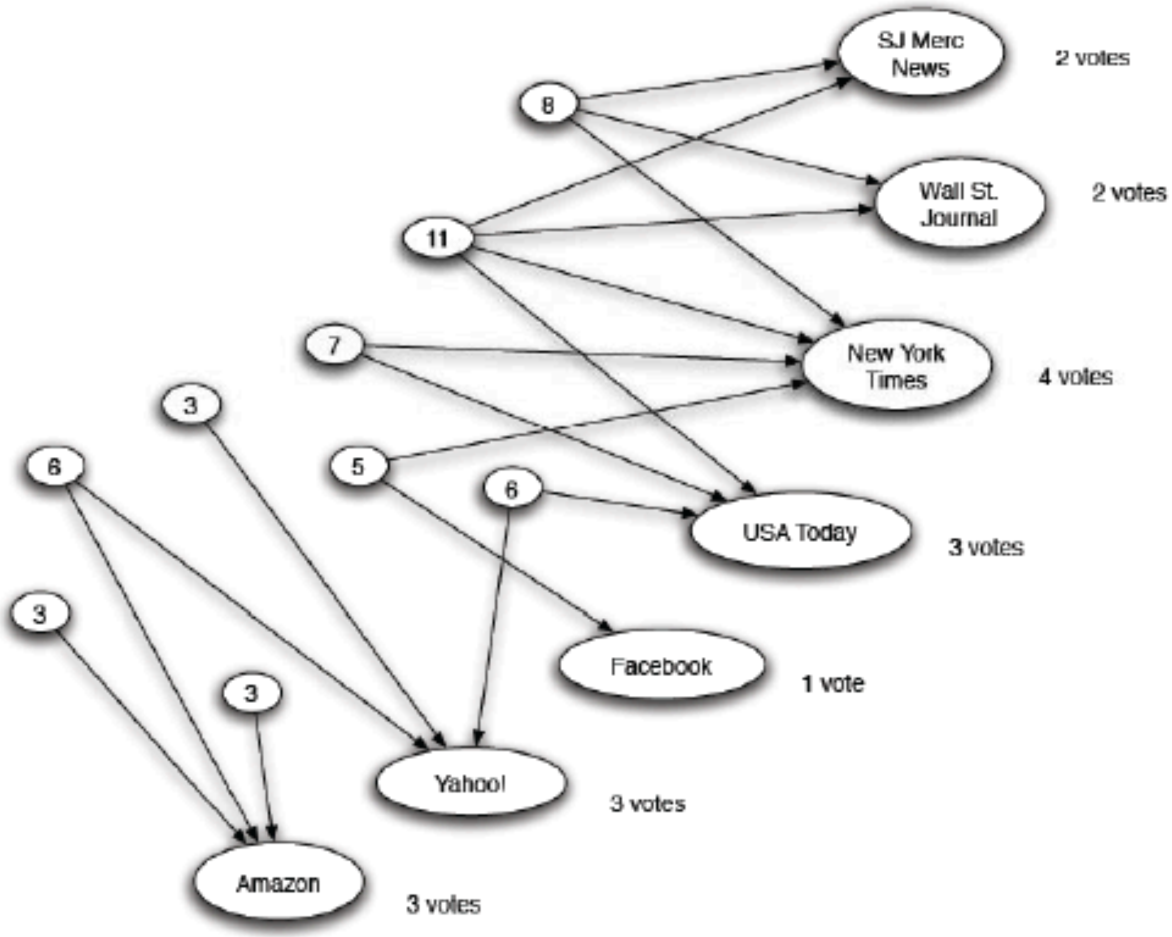
Example: Query “newspaper” Authorities

- 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

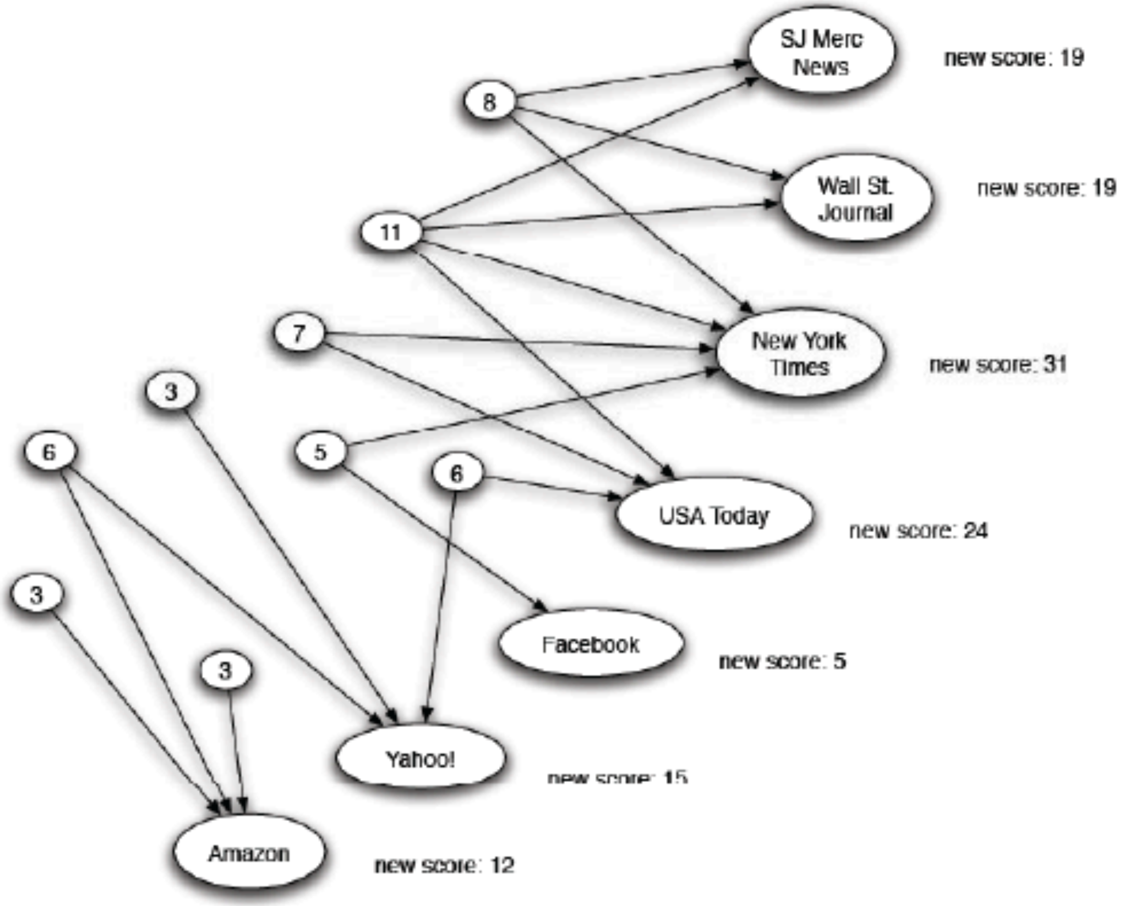


UNIVERSITY OF CAMBRIDGE

Principle of Repeated Improvement

- And we are now reweighting the authorities

- When do we stop?



UNIVERSITY OF CAMBRIDGE

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

UNIVERSITY OF CAMBRIDGE

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

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

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

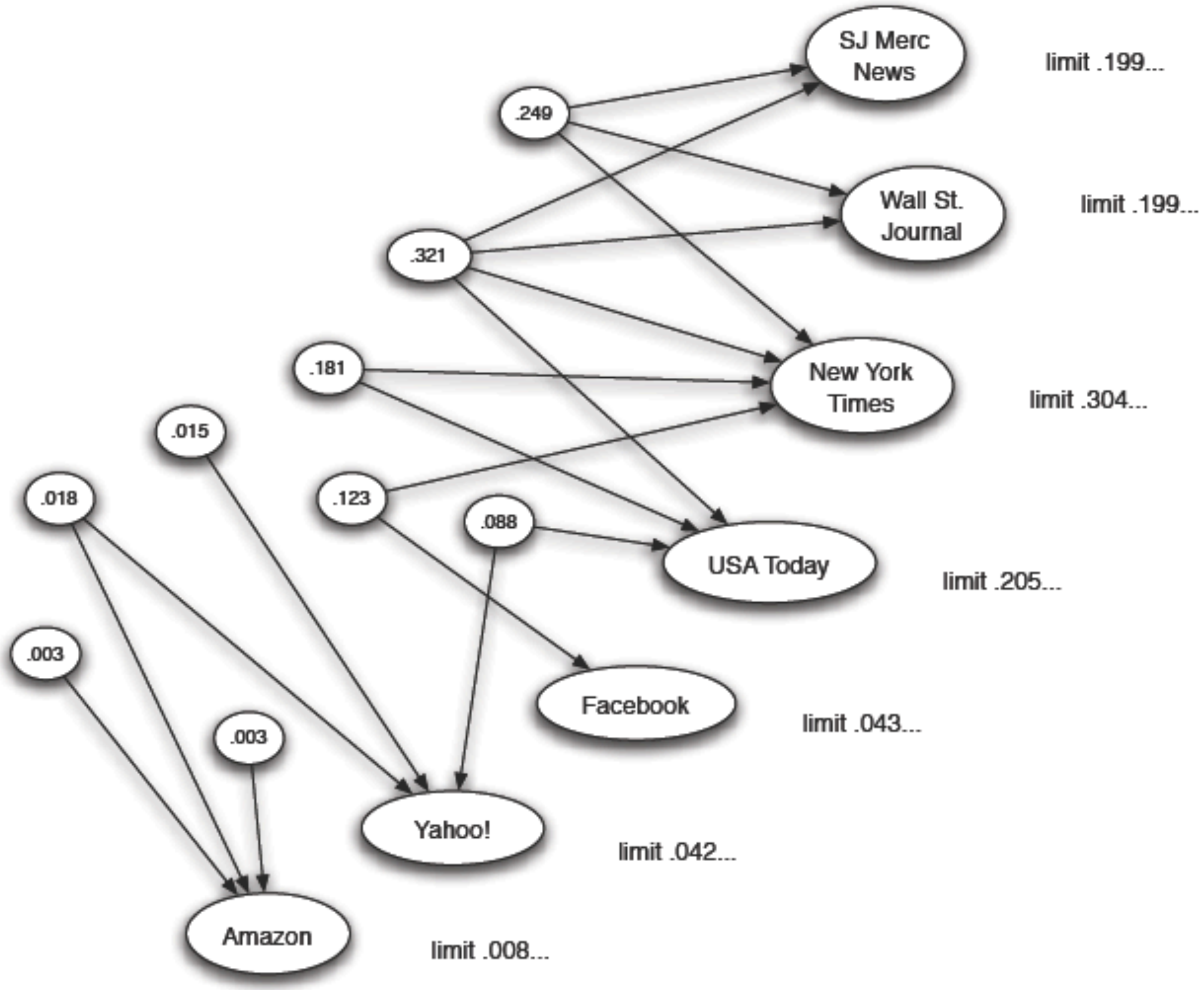
- Normalize

$$\sum a_i = 1$$

$$\sum h_j = 1$$

UNIVERSITY OF CAMBRIDGE

The process converges



UNIVERSITY OF CAMBRIDGE

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

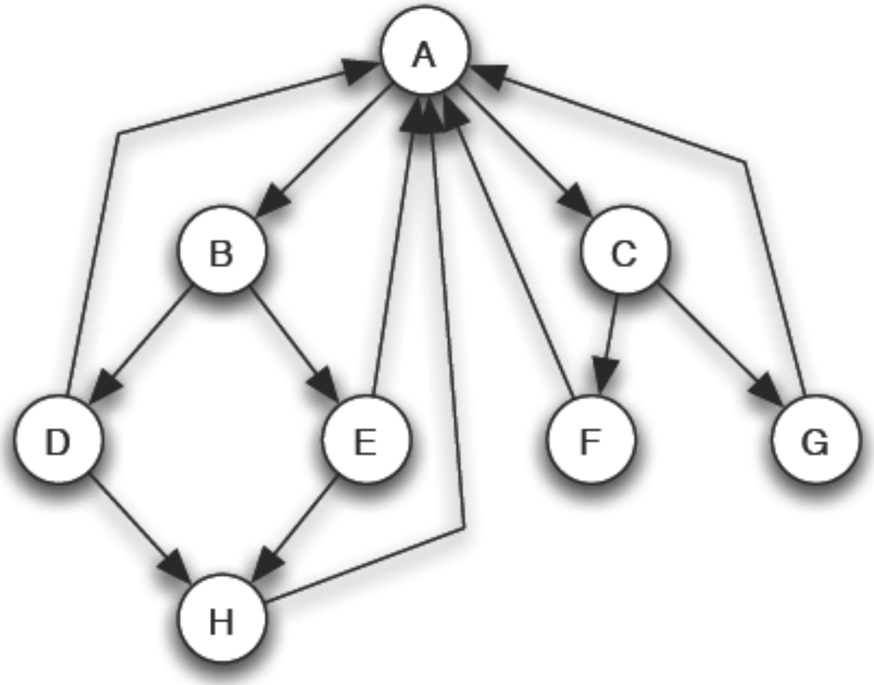
- Pages pass endorsements on outgoing links as fractions which depend on out-degree
- Initial PageRank value of each node in a network of n nodes: 1/n.
- Choose a number of steps k.
- **[Basic] Update rule**: each page divides its pagerank equally over the outgoing links and passes an equal share to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



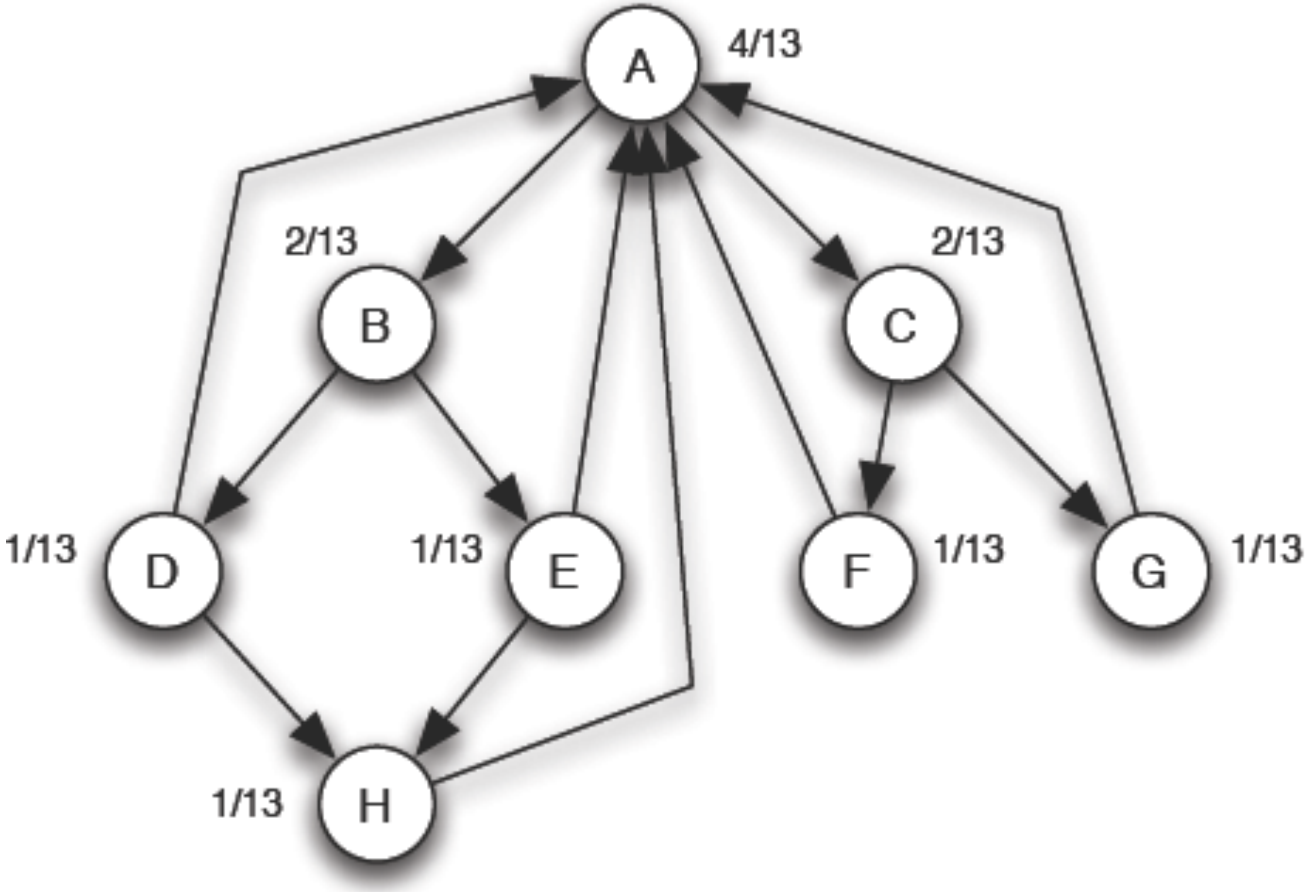
UNIVERSITY OF CAMBRIDGE

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.

UNIVERSITY OF CAMBRIDGE

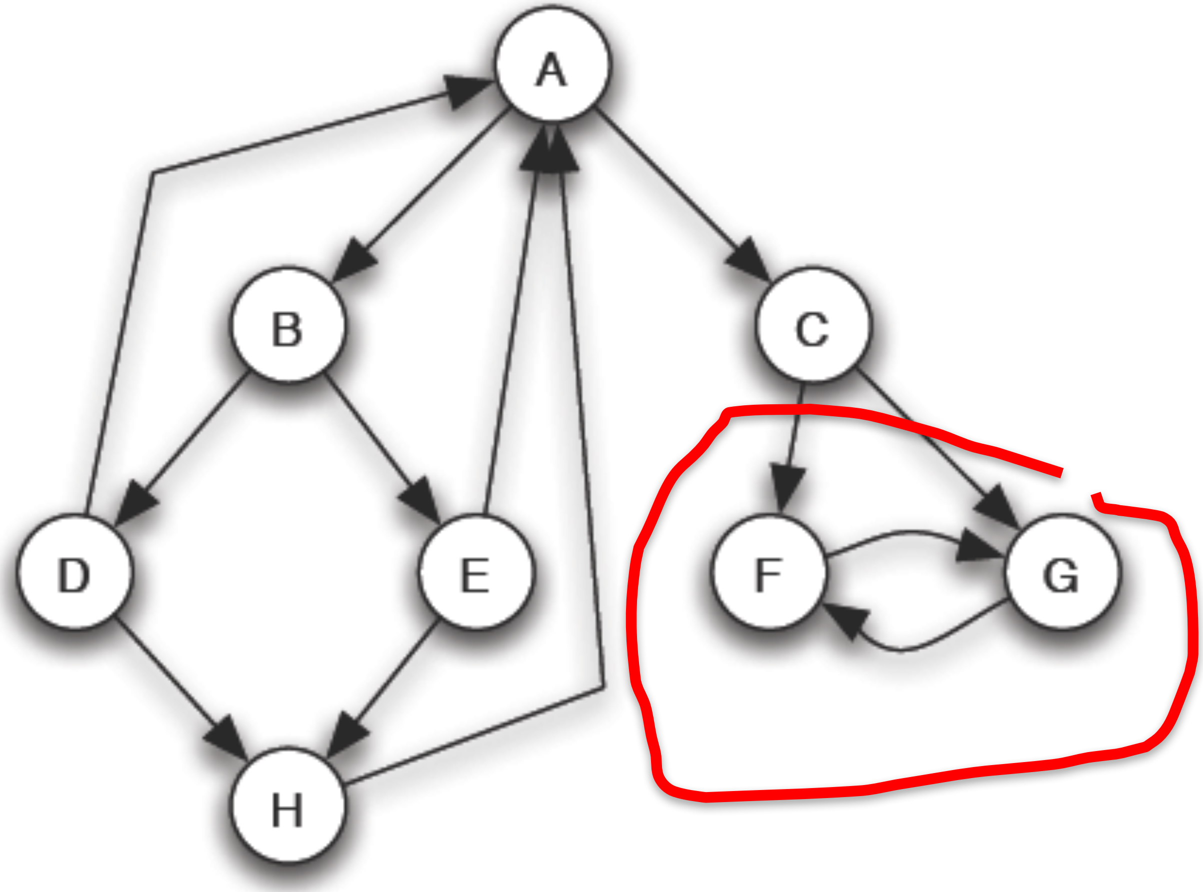
Example of Equilibrium



UNIVERSITY OF CAMBRIDGE

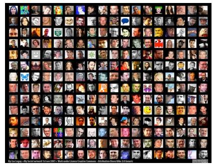
Problems with the basic PageRank
Dead ends

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



UNIVERSITY OF CAMBRIDGE

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

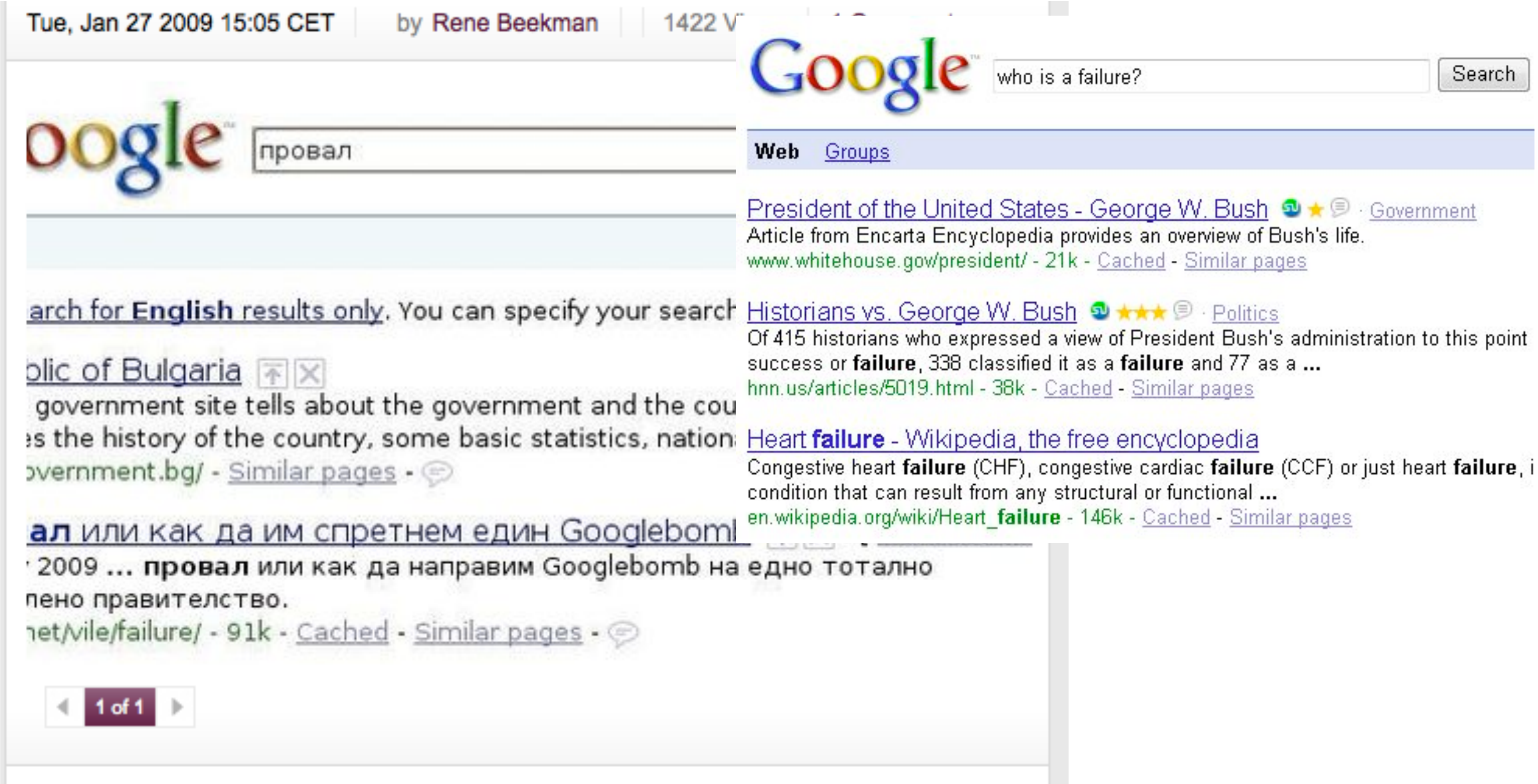
UNIVERSITY OF CAMBRIDGE

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.

UNIVERSITY OF CAMBRIDGE

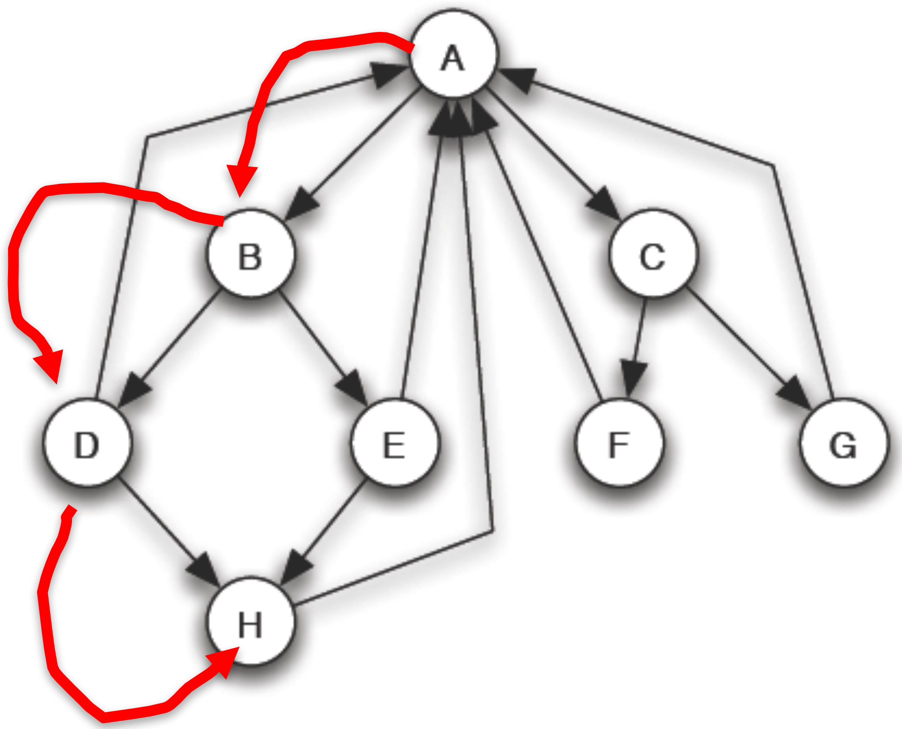
Examples of Google Bombs



UNIVERSITY OF CAMBRIDGE

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

- Chapter 13, 14 and 18

- 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.

UNIVERSITY OF CAMBRIDGE

Barabasi’s book has a good chapter on scale free networks too!