Social and Technological Network Analysis

Lecture 3: Structure of the Web and Power Laws

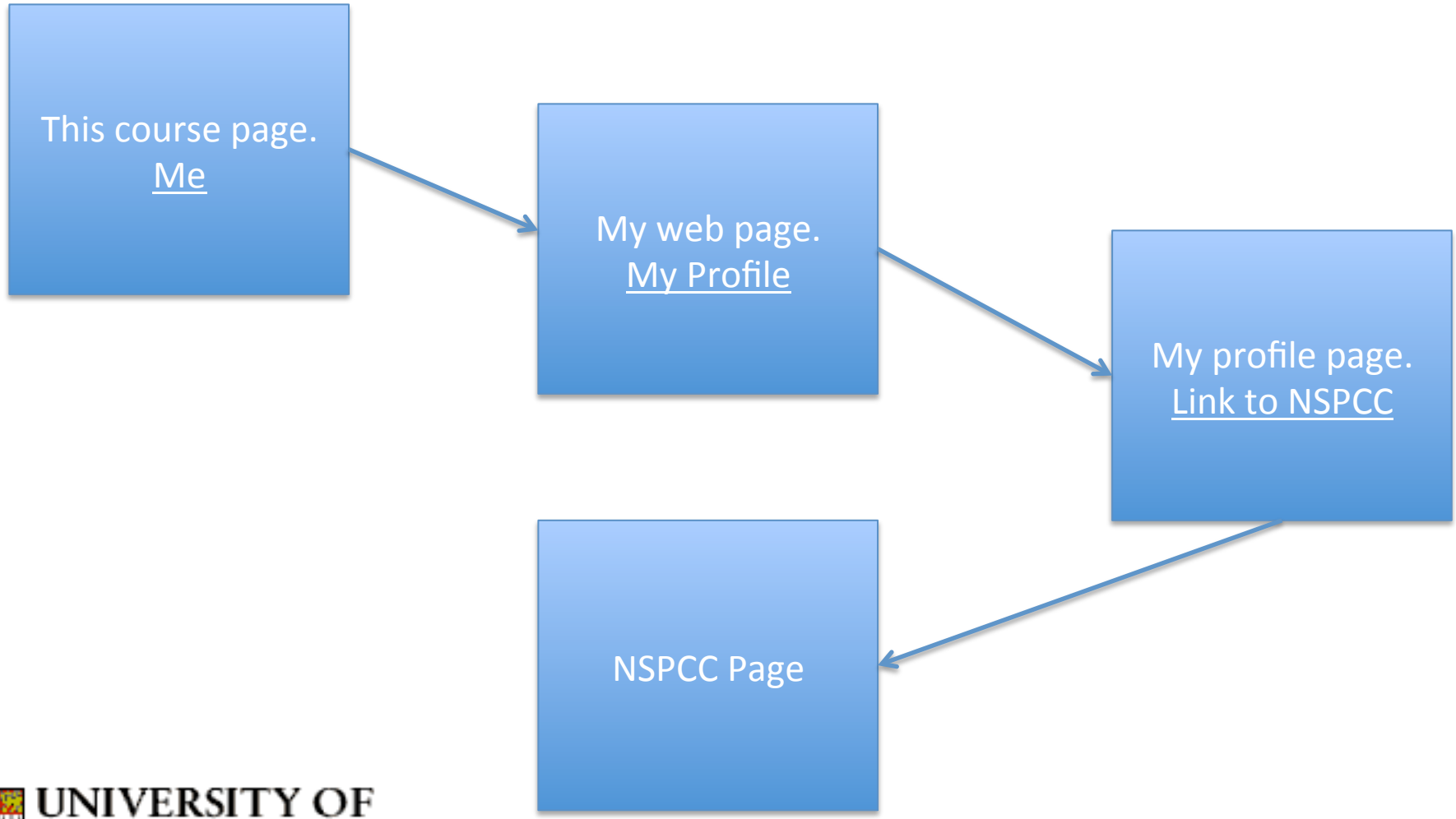
Dr. Cecilia Mascolo

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.

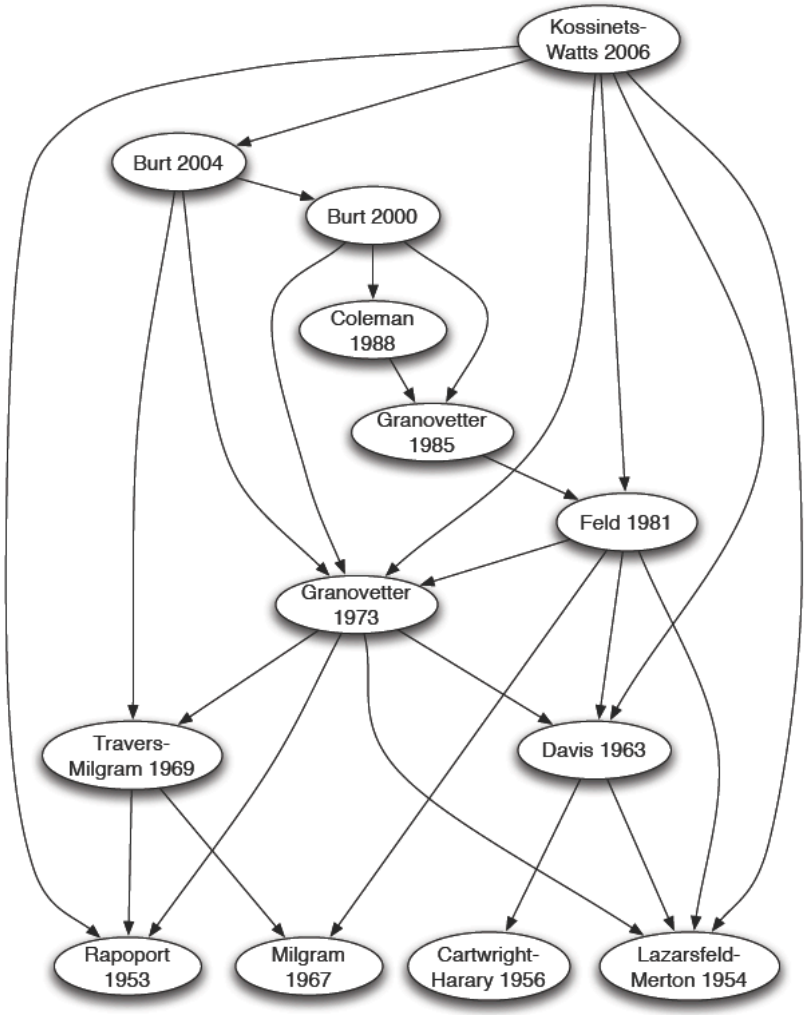
The Web is a Graph…



UNIVERSITY OF CAMBRIDGE

Precursor of hypertexts

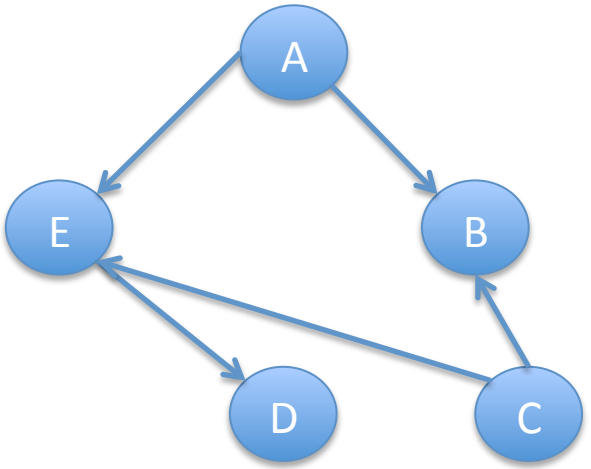
- Citation networks of books and articles.
- Difference: links point only backwards in time



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.



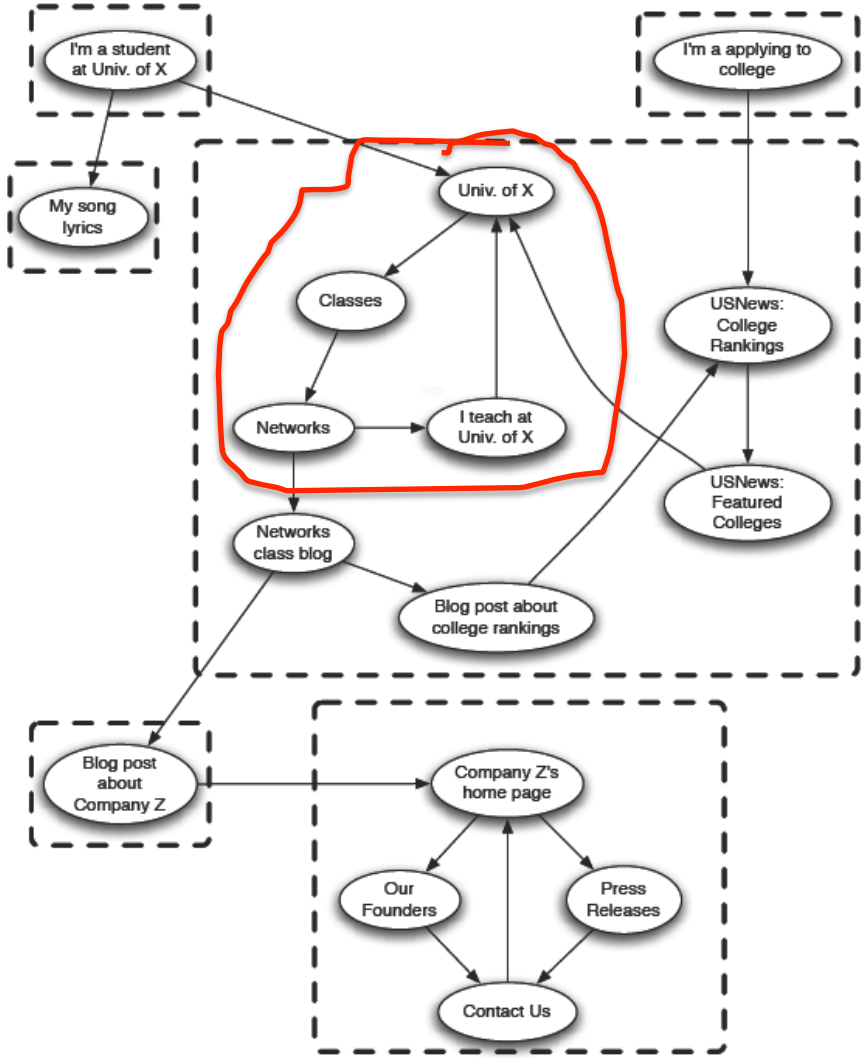
Strongly Connected Component

- 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.

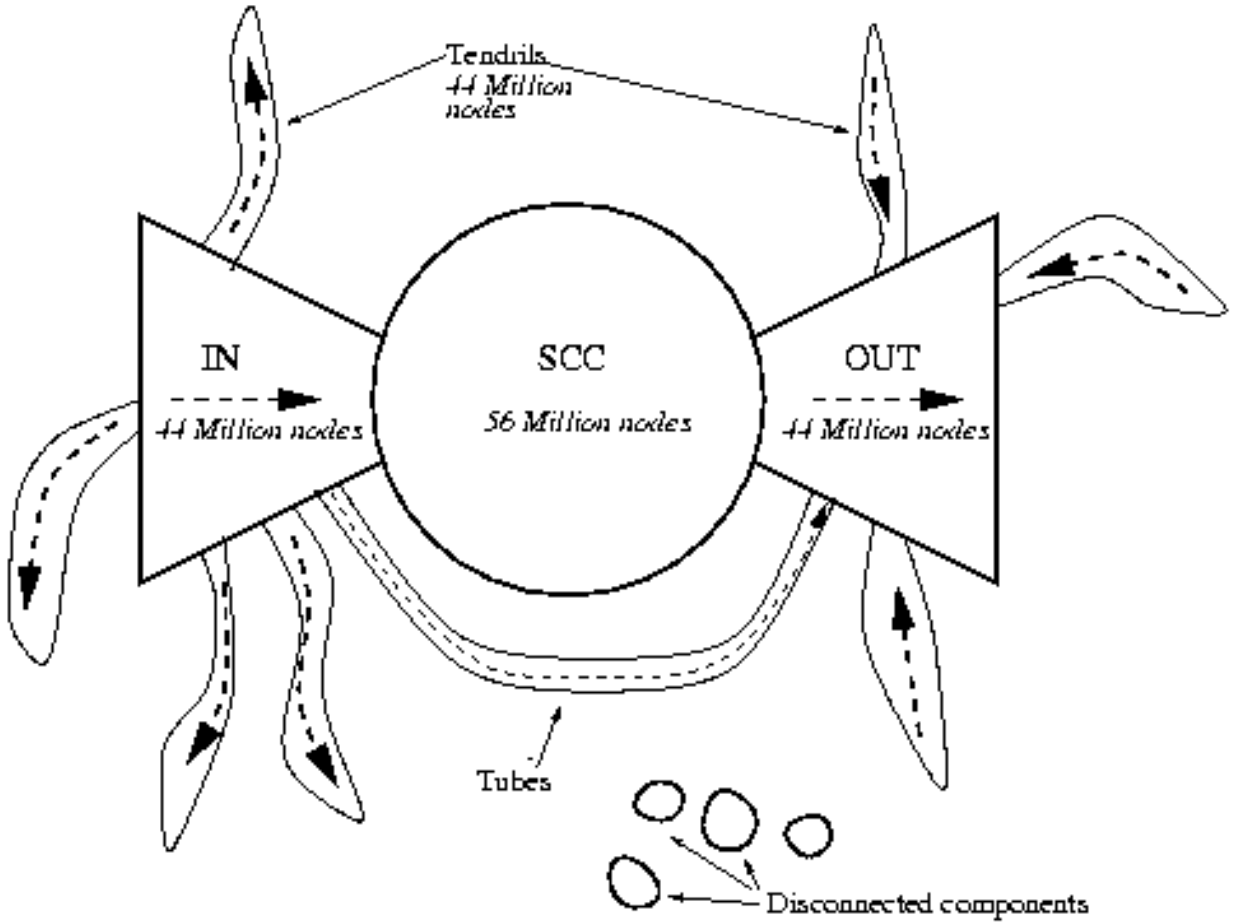
SCC example



UNIVERSITY OF CAMBRIDGE

The Web

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



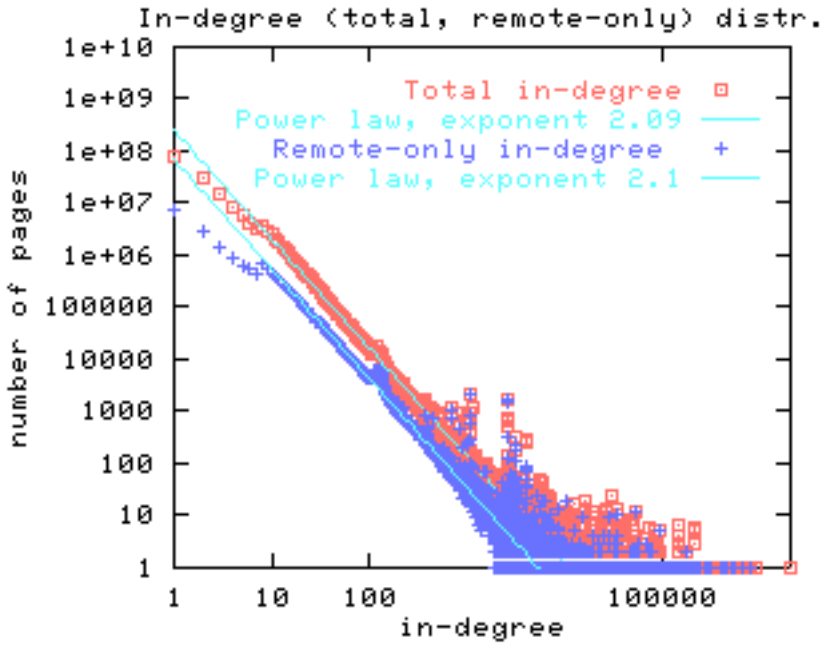
UNIVERSITY OF CAMBRIDGE

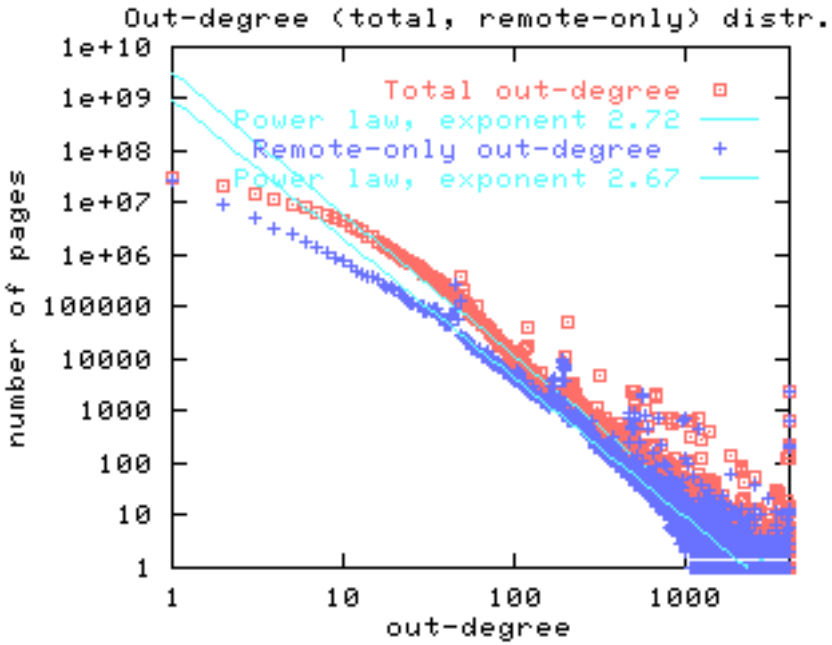
Popularity of Web Pages

- 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 n pages with k in-links decreases exponentially in k
  - Is this true for the Web?

Degree distribution for the Web

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



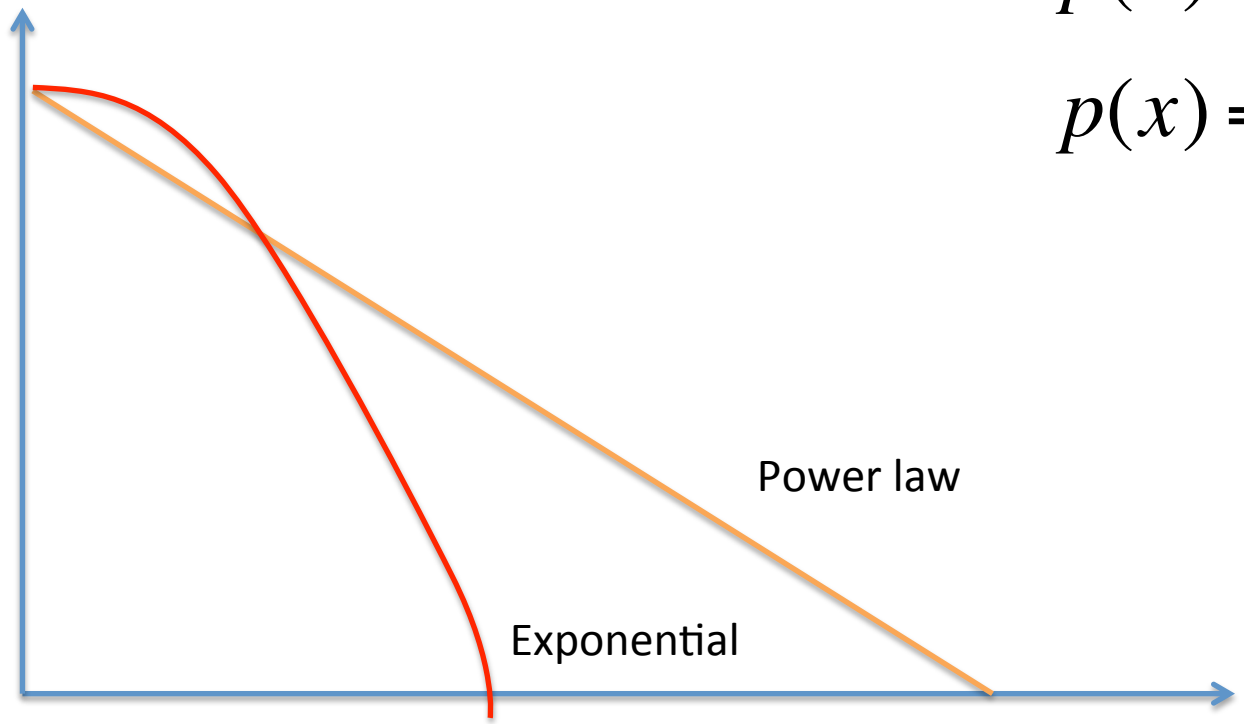


UNIVERSITY OF CAMBRIDGE

Power Law vs Exponential

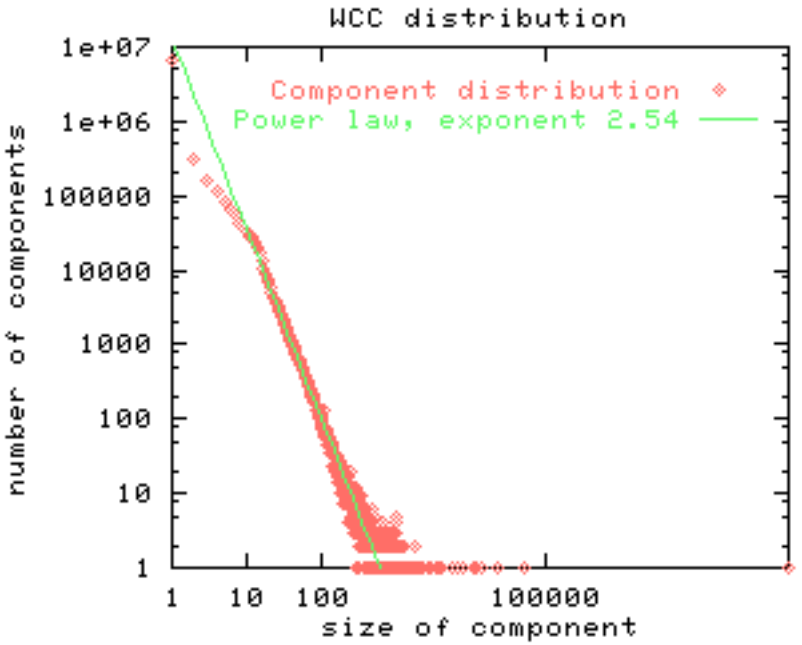
$$p(x) = x^{-\alpha}$$

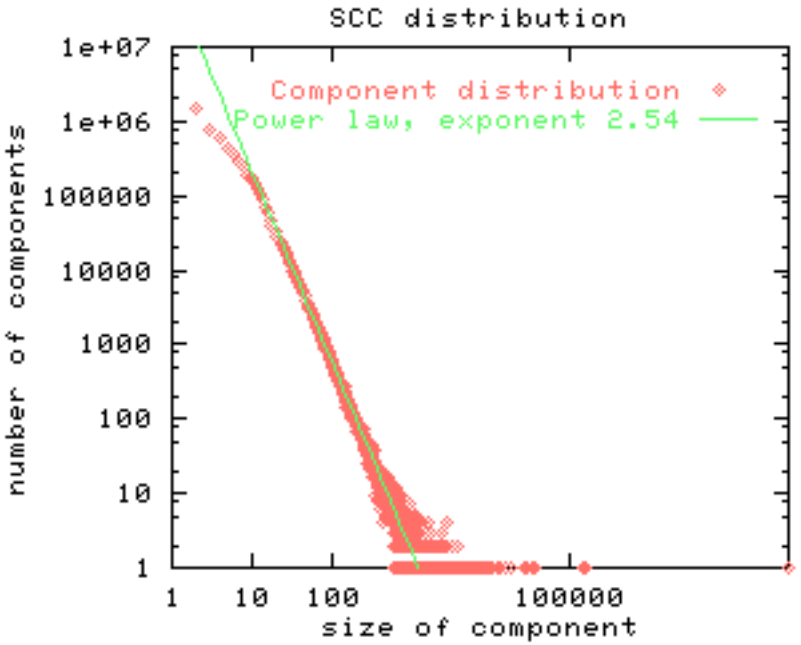
$$p(x) = e^{-\lambda x}$$



UNIVERSITY OF CAMBRIDGE

Distribution of WCC and SCC

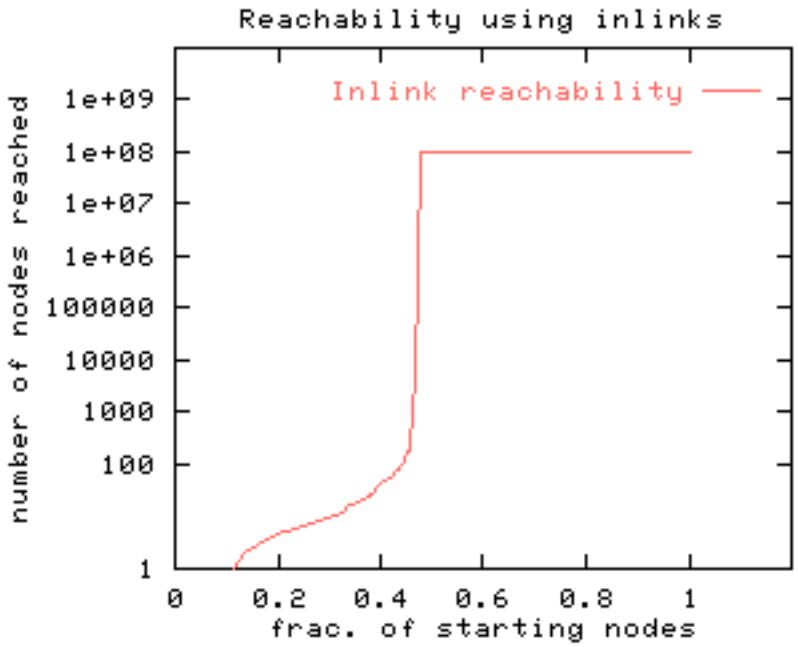


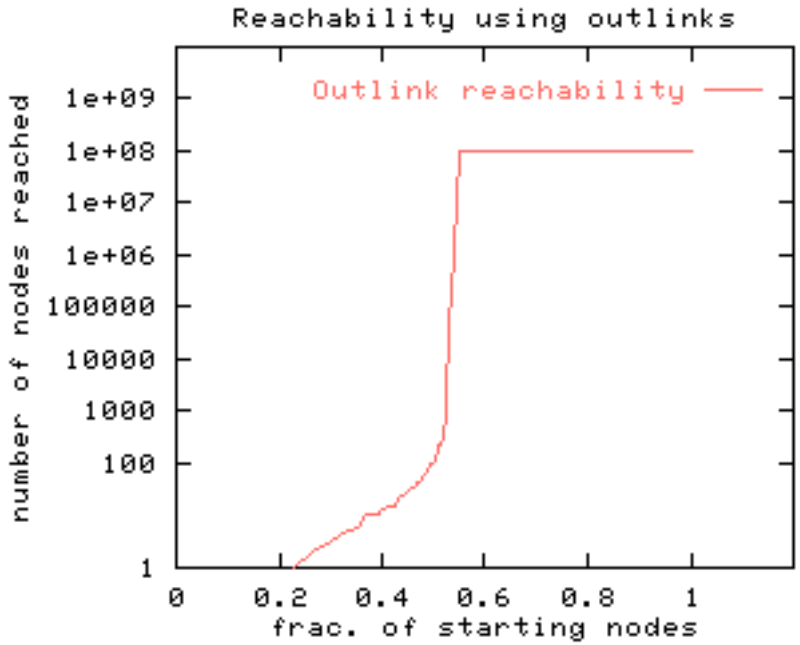


UNIVERSITY OF CAMBRIDGE

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

UNIVERSITY OF CAMBRIDGE

Power Laws aka Scale Free Networks

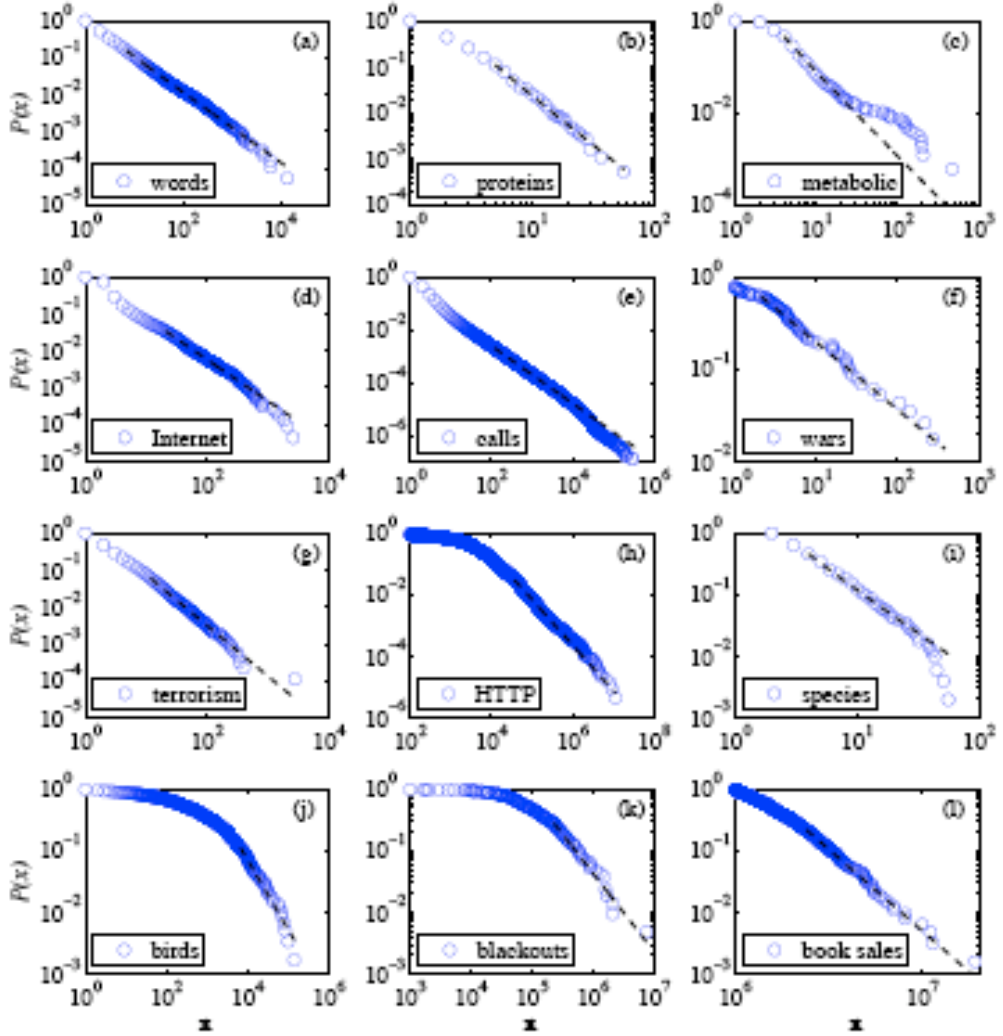
- We have seen that the degree distribution followed a straight line in log-log

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

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

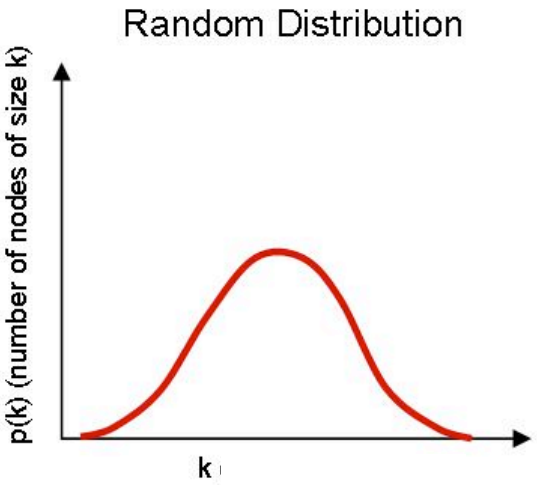
- α defines the slope of the curve
- α is typically between 2 and 3.

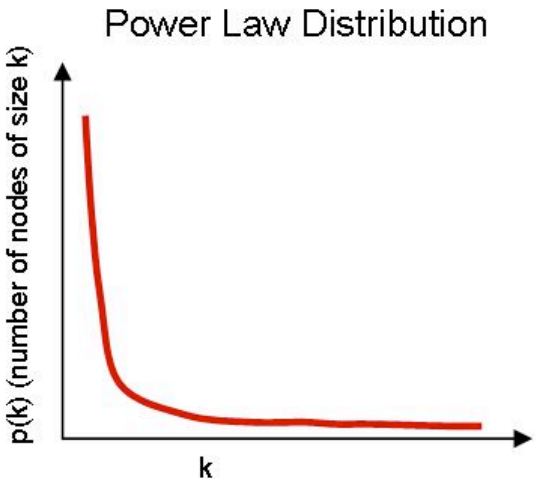
Power Laws in various domains



UNIVERSITY OF CAMBRIDGE

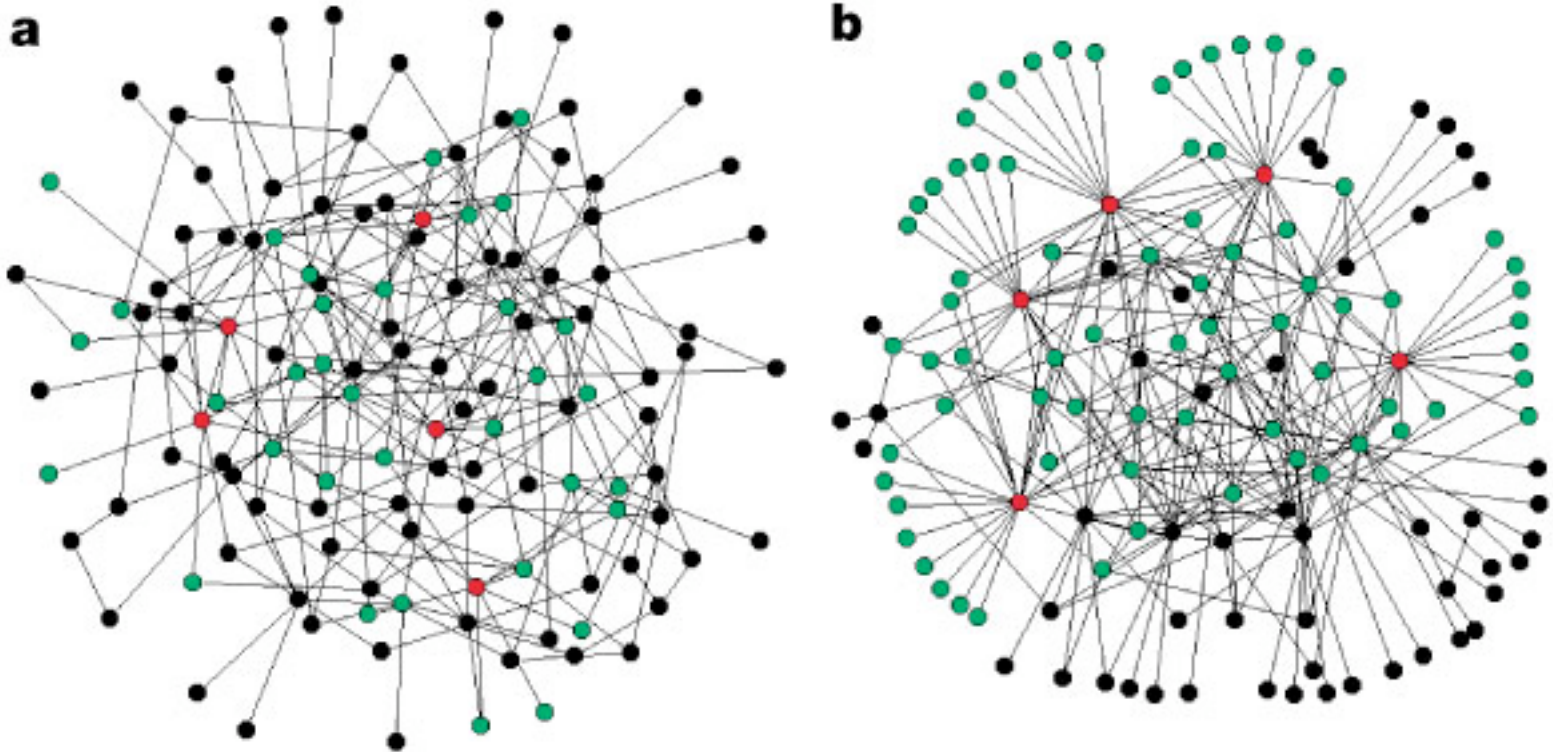
What does it mean?





UNIVERSITY OF CAMBRIDGE

Random vs Power Law Networks



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

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 $1/k^c$

UNIVERSITY OF CAMBRIDGE

Let's get formal…

- Number of in-links Xj(t) at time t.
- Initially (creation of j at time j) Xj(j)=0
- What's the probability that j gains a link at step t+1?
  - Node t+1 links with prob p to an earlier node: so probability of linking to j is 1/t
  - Node t+1 links with prob 1-p to a node proportionally to its links. Total number of links is t of which Xj(t) point to j: Xj(t)/t so:

$$\frac{p}{t} + \frac{(1-p)X_j(t)}{t}$$

UNIVERSITY OF CAMBRIDGE

And more formal…

- If we use a continuous function $x_j(t)$

- Initial condition $x_j(j)=0$
- Growth condition:

$$\frac{dx_j}{dt} = \frac{p}{t} + \frac{(1-p)x_j}{t}$$

- Let's use q=1-p
- Divide both sides by $p+qx_j$

$$\frac{dx_j}{dt} = \frac{p}{t} + \frac{qx_j}{t}$$

$$\frac{1}{p+qx_j}\frac{dx_j}{dt} = \frac{1}{t}$$

- Integrating

$$\int \frac{1}{p+qx_j}\frac{dx_j}{dt}dt = \int \frac{1}{t}dt$$

$$\ln(p+qx_j) = q\ln t + c$$

- $A=e^c$

$$p+qx_j = At^q$$

$$x_j(t) = \frac{1}{q}(At^q - p)$$

- Initial xj(j)=0

$$0 = x_j(j) = \frac{1}{q}(Aj^q - p)$$

...

- hence: $A = p / j^q$

$$x_j(t) = \frac{1}{q}\left(\frac{p}{j^q}t^q - p\right) = \frac{p}{q}\left[\left(\frac{t}{j}\right)^q - 1\right]$$

- For a given k, and a given t what fraction of nodes have at least k in-links at time t? x_j approximates j's in-links number so:
- For a given k and a given t what fraction of all functions xj satisfy xj(t)>=k?

...

$$x_j(t) = \frac{1}{q}\left(\frac{p}{j^q}t^q - p\right) = \frac{p}{q}\left[\left(\frac{t}{j}\right)^q - 1\right] \geq k$$

$$j \leq t\left[\frac{q}{p}k + 1\right]^{-1/q}$$

- Out of x1, x2…xt at time t, the fraction of values j that satisfy this is

$$\frac{1}{t}t\left[\frac{q}{p}k + 1\right]^{-1/q} = \left[\frac{q}{p}k + 1\right]^{-1/q}$$

Proportional to k

...

- Fraction of xj which are at least k (F(k)) is proportional to $k^{-1/q}$
- Fraction of nodes with exactly k in-links (fk)
  - Take the derivative (approximation) –dF/dk

$$\frac{1}{q}\frac{q}{p}[\frac{q}{p}k+1]^{-1-1/q}$$

- So the fraction of nodes with k in-links is proportional to $k^{-(1+1/q)}$ ie, a powerlaw with exponent

$$1+\frac{1}{q}=1+\frac{1}{1-p}$$

Preferential Attachment

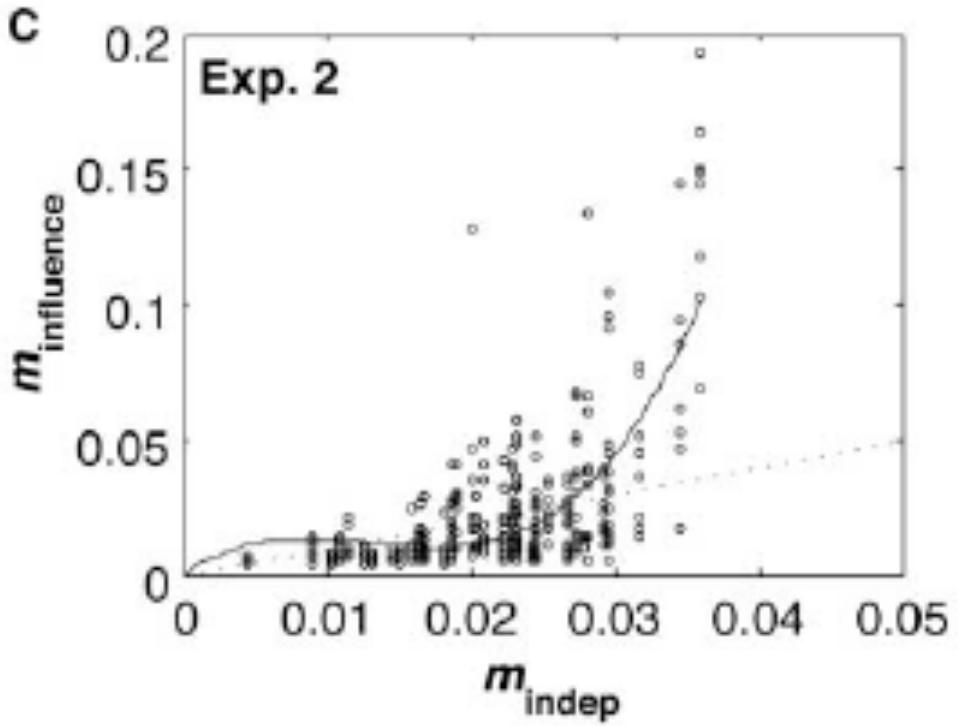
- What have we shown?
- There is a “copying” behaviour happening in these networks where node seem to emulate other nodes.
- This is shown true for selection of books, songs, web pages, movies etc.

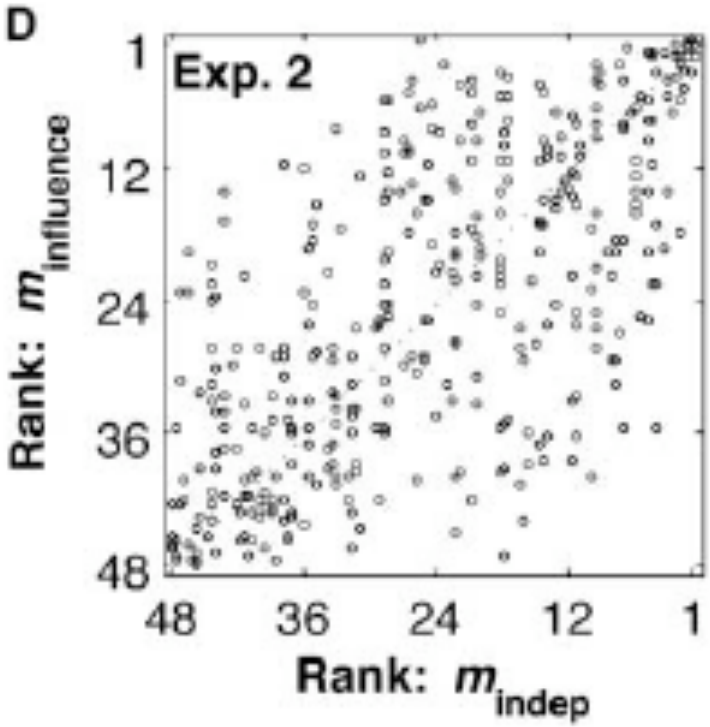
How predictable is the rich-get-richer process?

- 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…

Unpredictability [Salganik et al 06]

- 48 songs, 14,000 participants, 8 servers

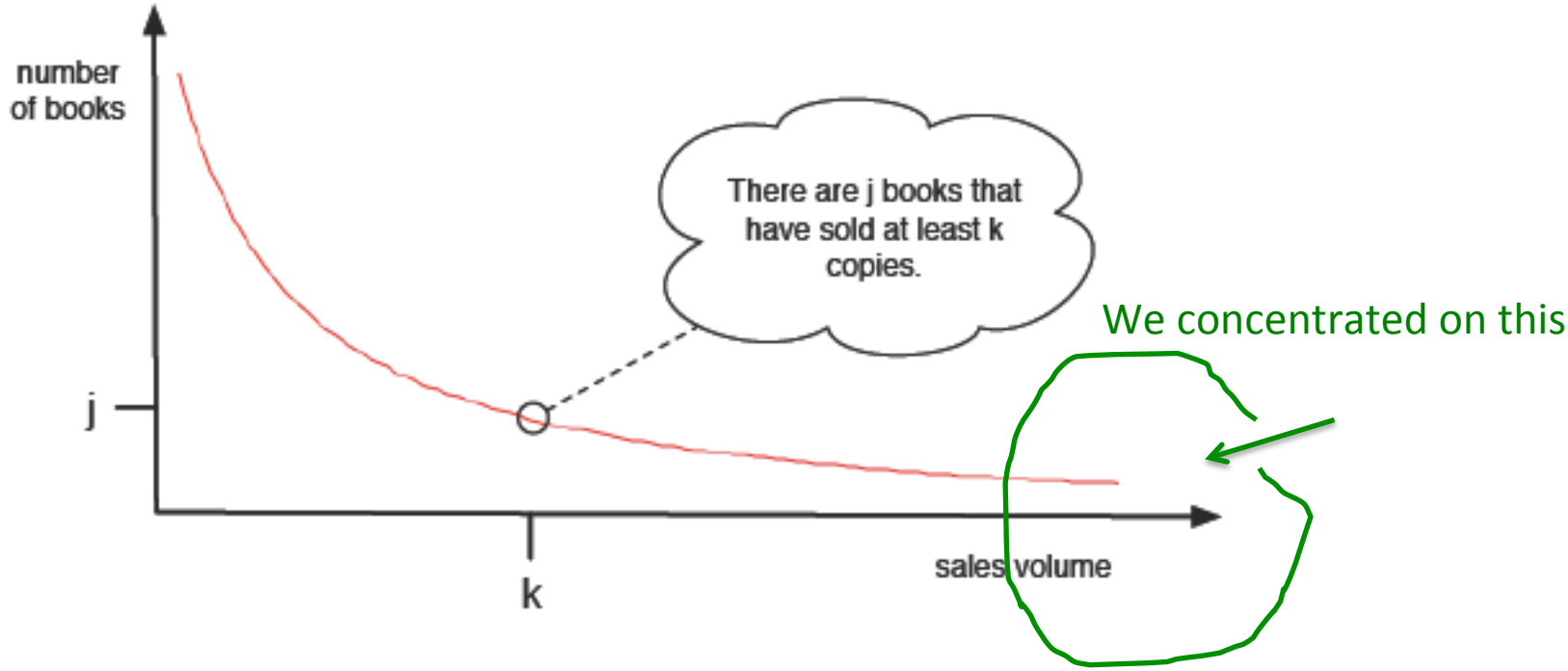




UNIVERSITY OF CAMBRIDGE

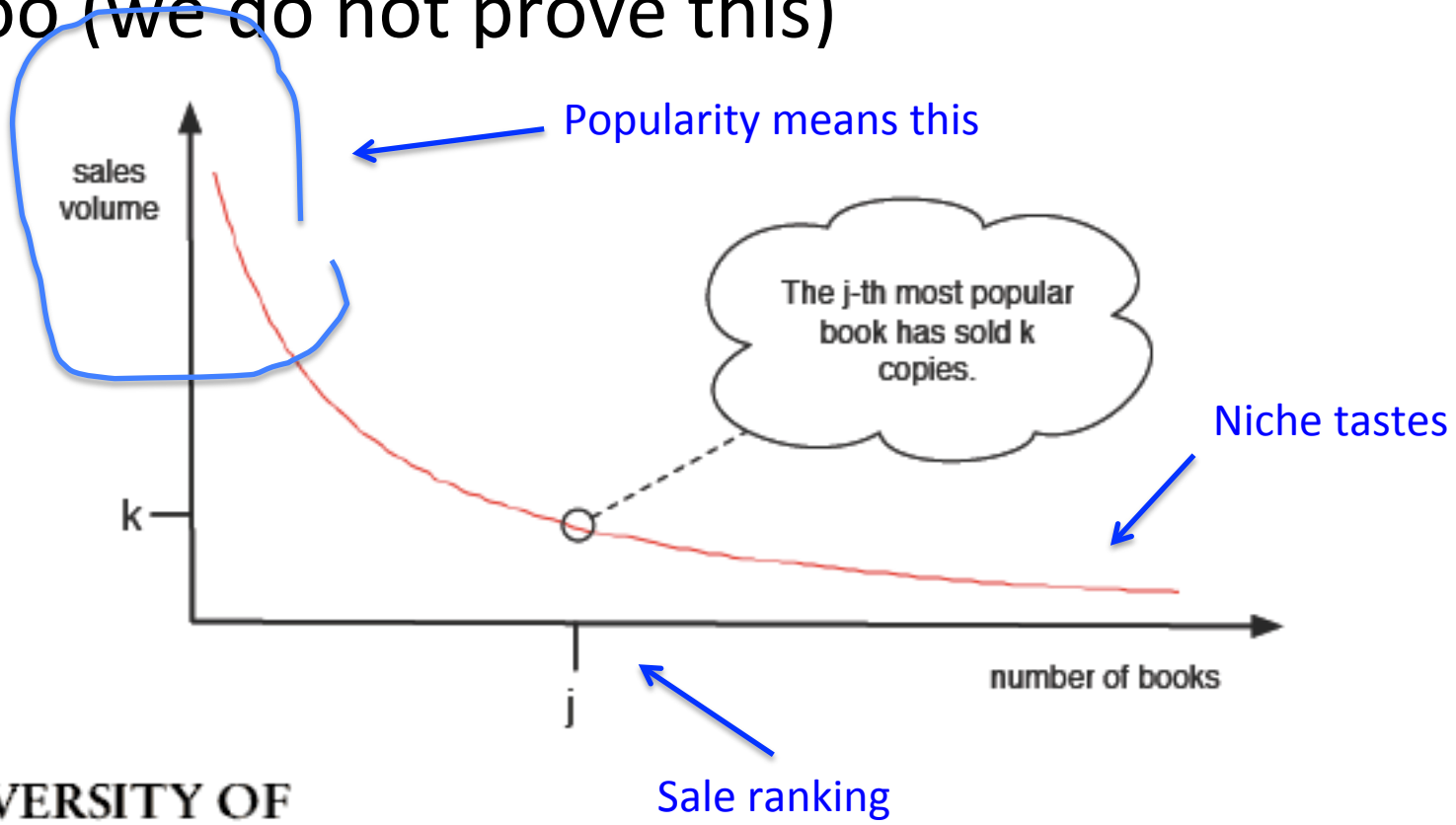
View of the curve

- The way we have seen the curve so far…



Let's transform the function

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



References

- Chapter 13 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.