

# Popularity dynamics on the Web and Wikipedia

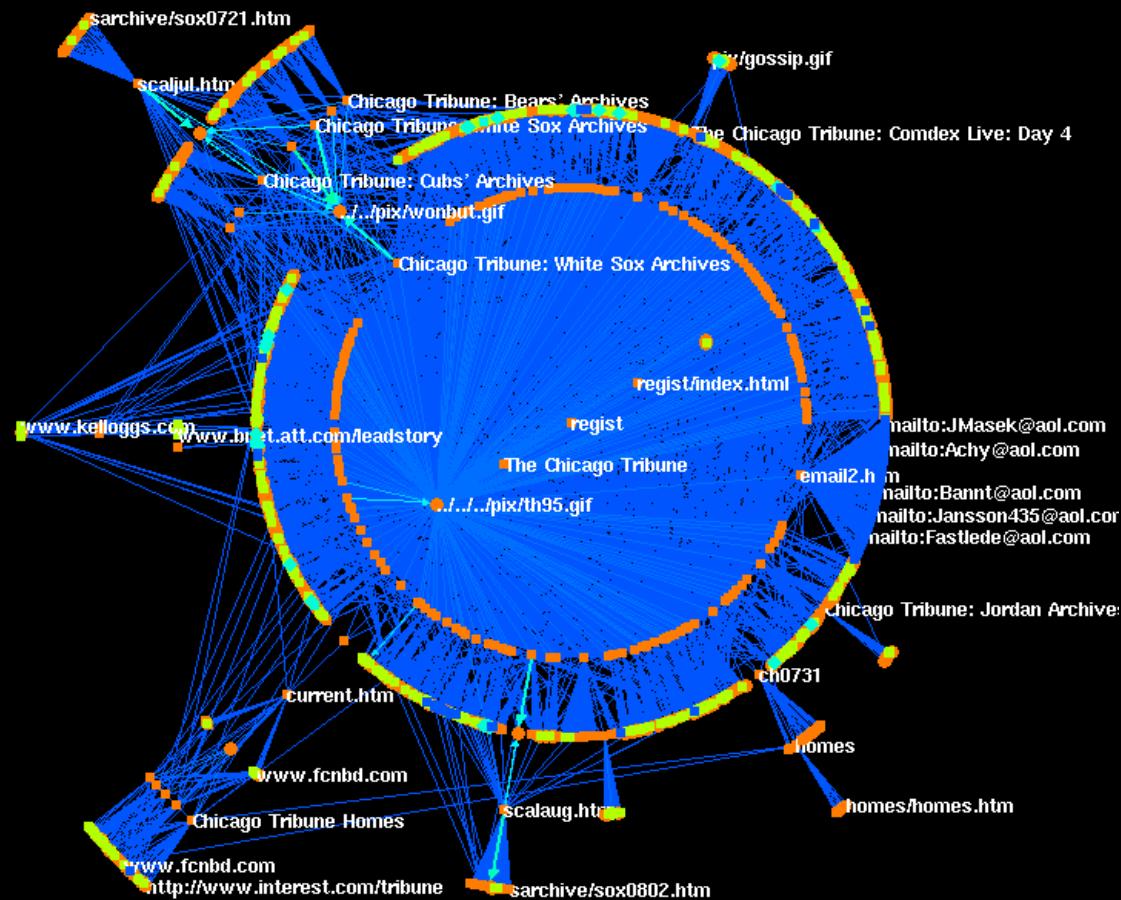
Santo Fortunato



# The World-Wide-Web

Virtual network to find and share information  
(T. Berners-Lee, 1991)

- web pages
- hyperlinks



CRAWLS

# Wikipedia

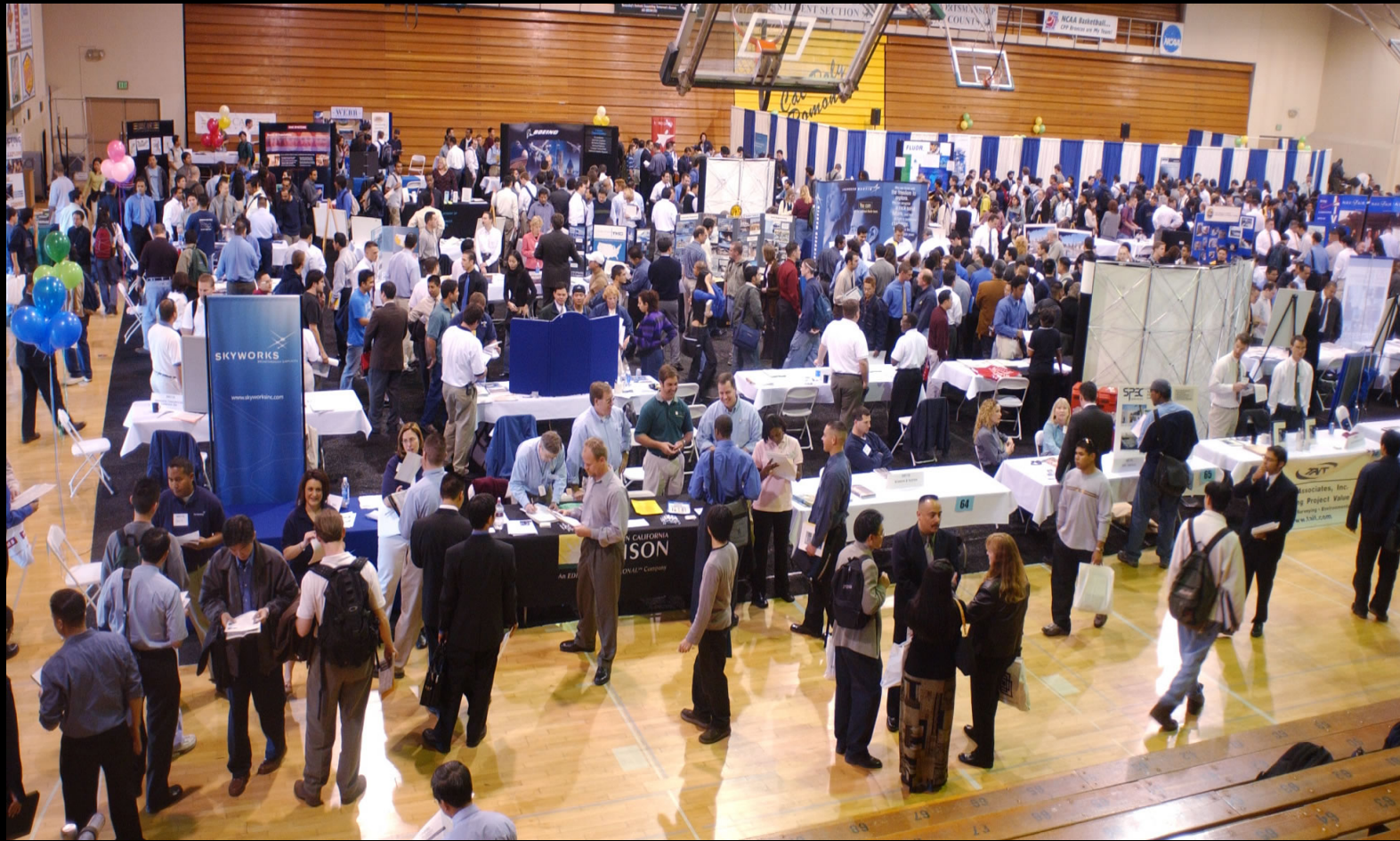
Free collaborative online encyclopedia  
(J. Wales & L. Sanger, 2001)



The image shows the Wikipedia homepage with the following content:

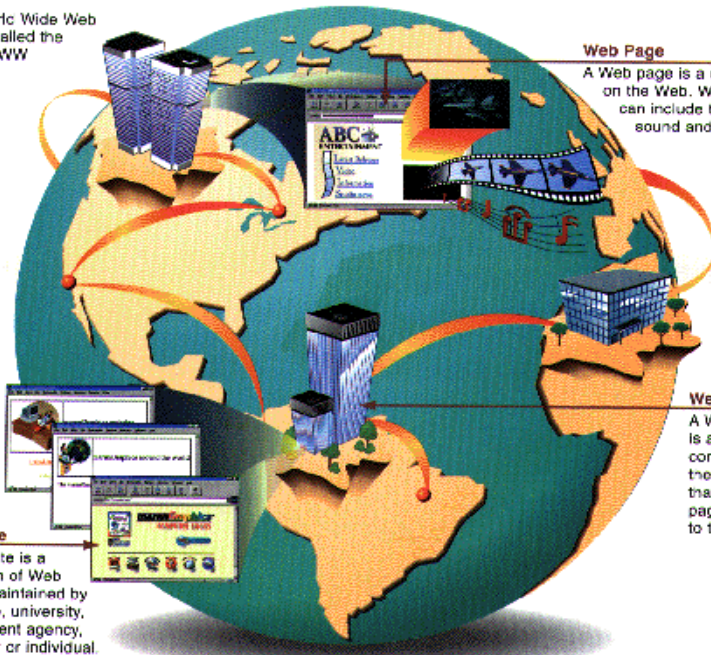
- WIKIPEDIA** logo at the top center.
- Language links arranged around a central globe logo:
  - English**: *The Free Encyclopedia*, 1 551 000+ articles
  - Deutsch**: *Die freie Enzyklopädie*, 517 000+ Artikel
  - Français**: *L'encyclopédie libre*, 415 000+ articles
  - Polski**: *Wolna encyklopedia*, 328 000+ haset
  - Nederlands**: *De vrije encyclopedie*, 251 000+ artikelen
  - Português**: *A enciclopédia livre*, 211 000+ artigos
  - Español**: *La enciclopedia libre*, 183 000+ artículos
  - Svenska**: *Den fria encyklopedin*, 200 000+ artiklar
  - 日本語**: *フリー百科事典*, 305 000+ 記事
  - Italiano**: *L'enciclopedia libera*, 227 000+ voci
- Search bar at the bottom with the text: search • suche • rechercher • szukaj • 検索 • zoeken • ricerca • busca • sök • buscar • поиск • 搜索
- Search bar containing the text:  English

# Websites: a daily fight for visibility!



The World Wide Web is part of the Internet. The Web consists of a huge collection of documents stored on computers around the world.

The World Wide Web is also called the Web, WWW or W3.



#### Web Page

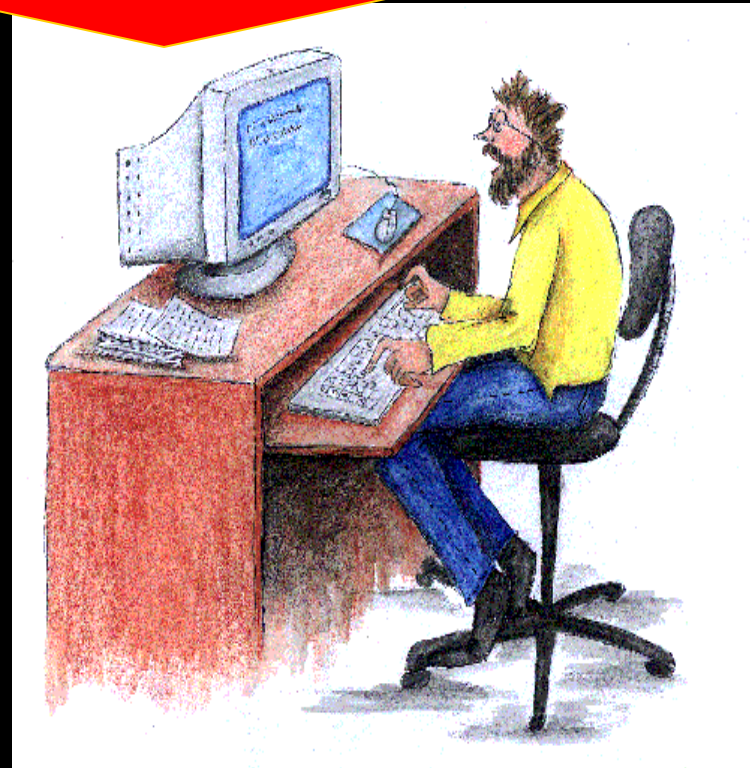
A Web page is a document on the Web. Web pages can include text, pictures, sound and video.

#### Web Server

A Web server is a computer connected to the Internet that makes Web pages available to the world.

#### Web Site

A Web site is a collection of Web pages maintained by a college, university, government agency, company or individual.



Website popularity is the product of the interplay between users and the Web!

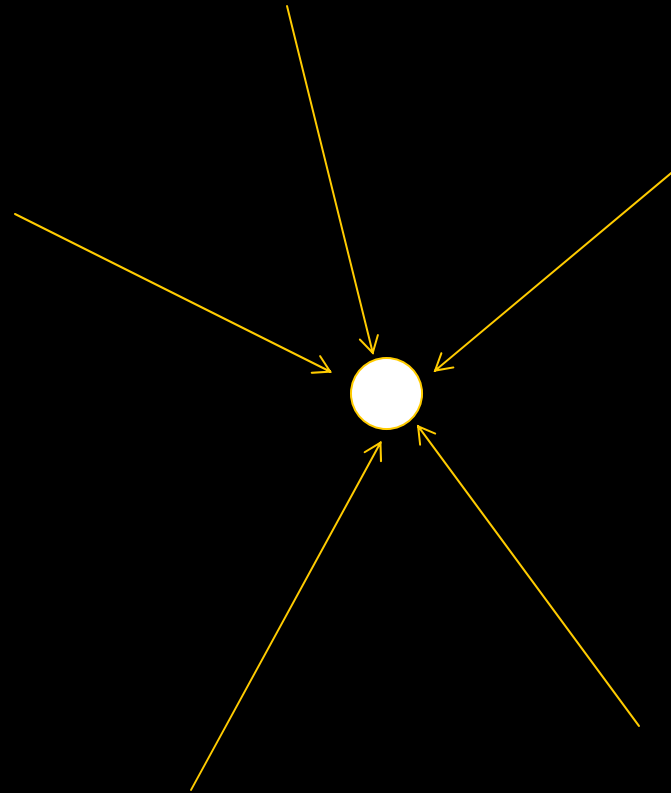
# The Web graph

Web pages → vertices/nodes

Hyperlinks → edges/links

Huge directed graph ( $10^{11}$  nodes?)

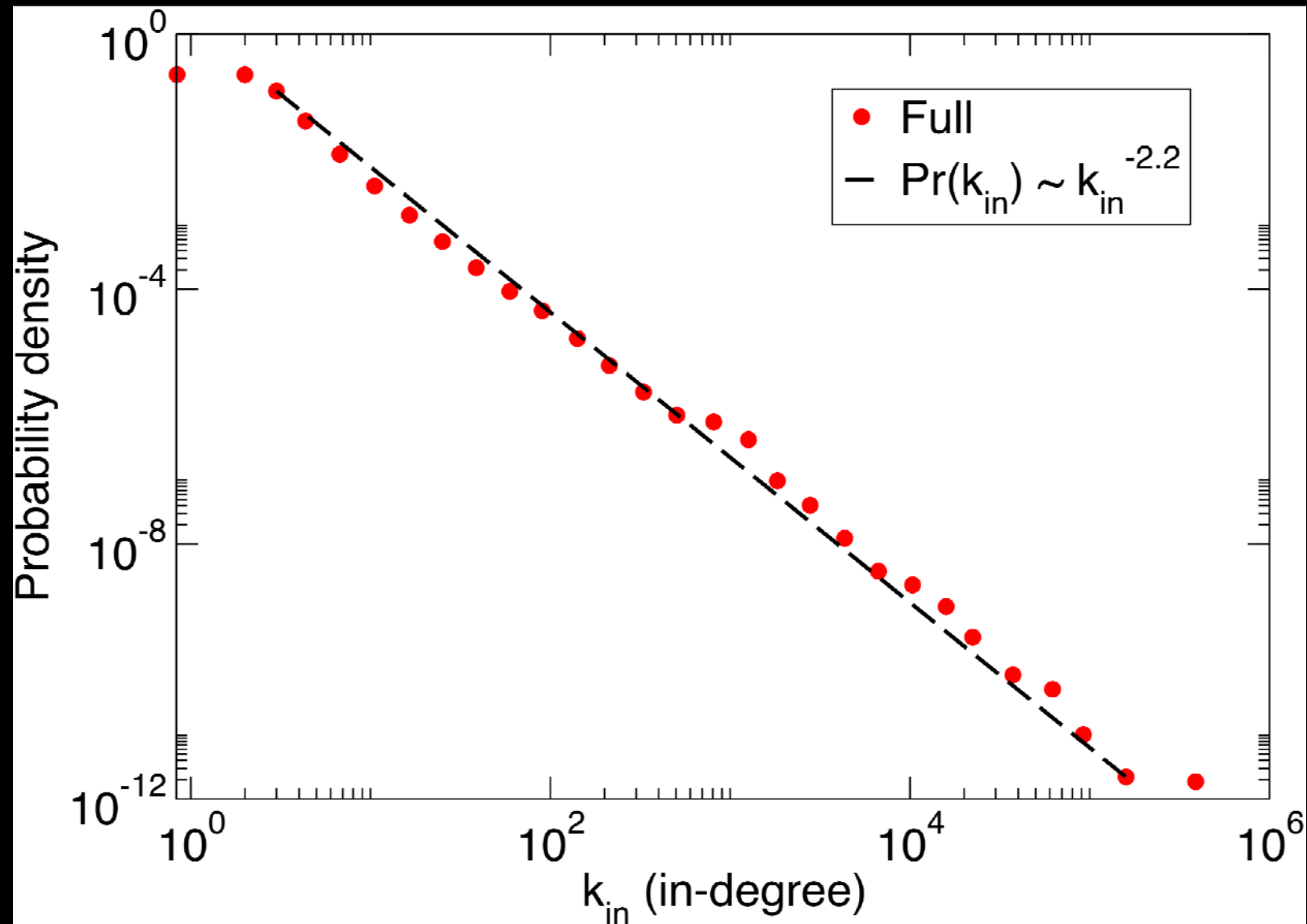
# Link popularity on the Web



Indegree: number of incoming hyperlinks to a page

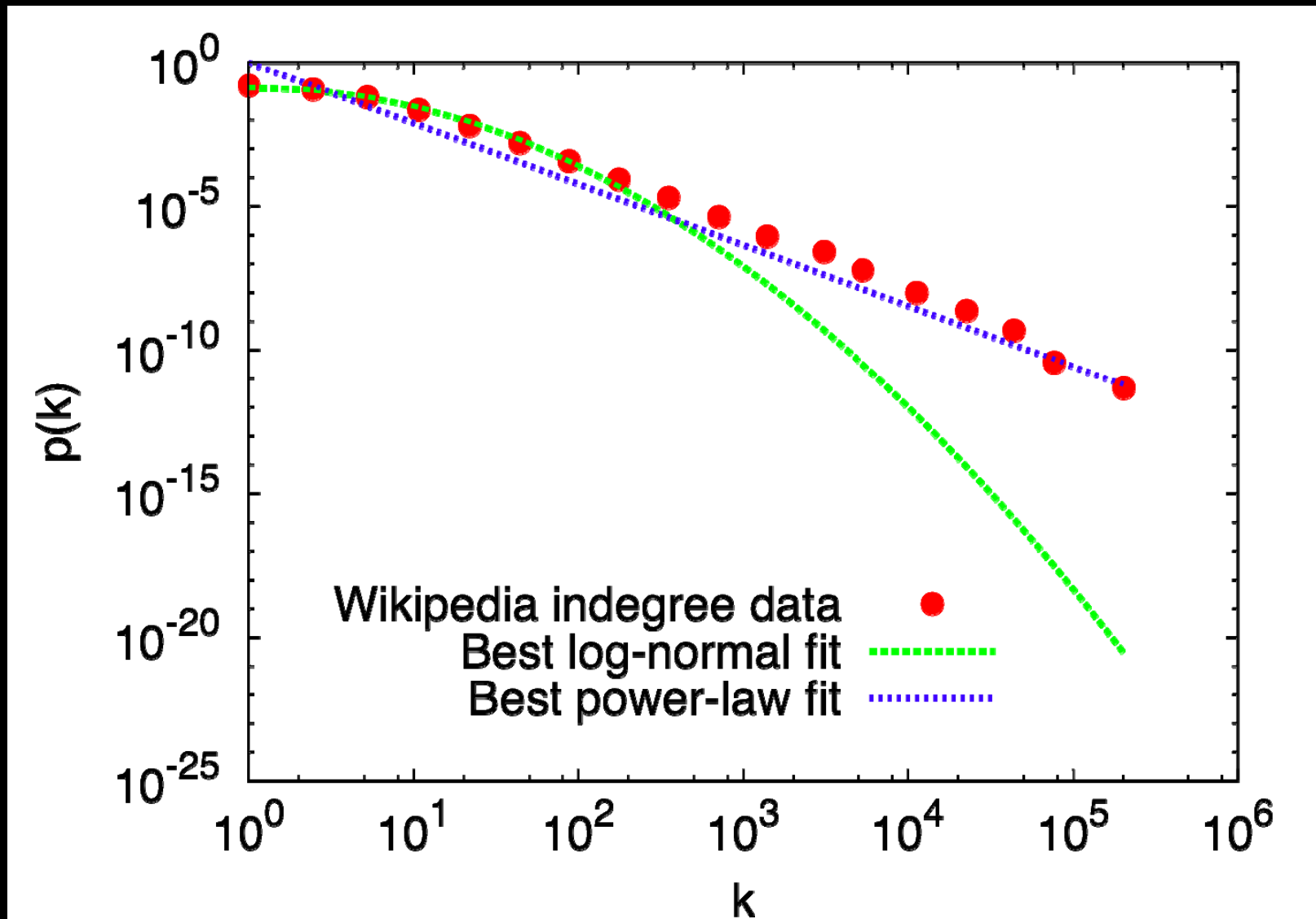
Natural popularity measure: the more inlinks, the more popular the page

# Link popularity on the Web

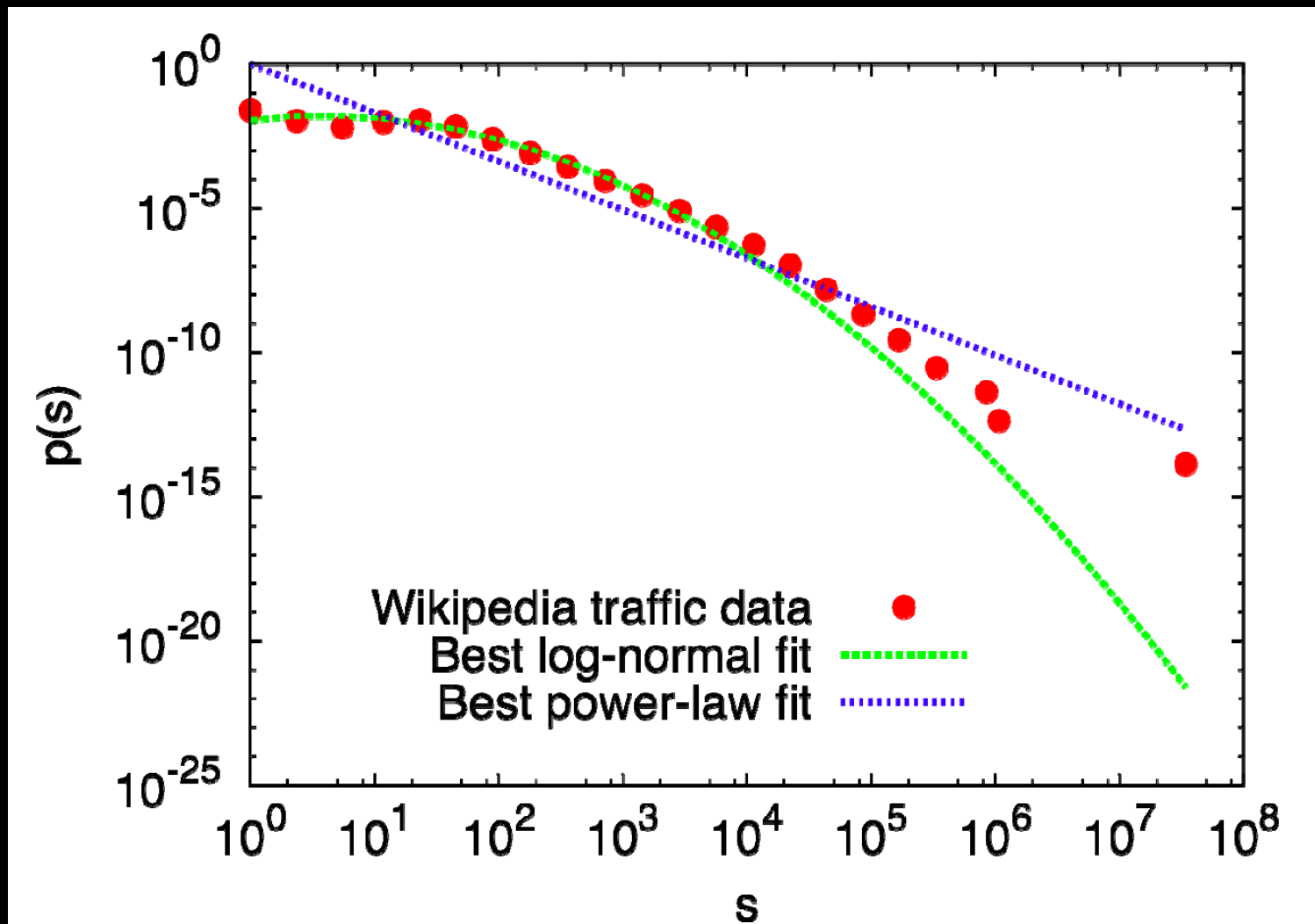




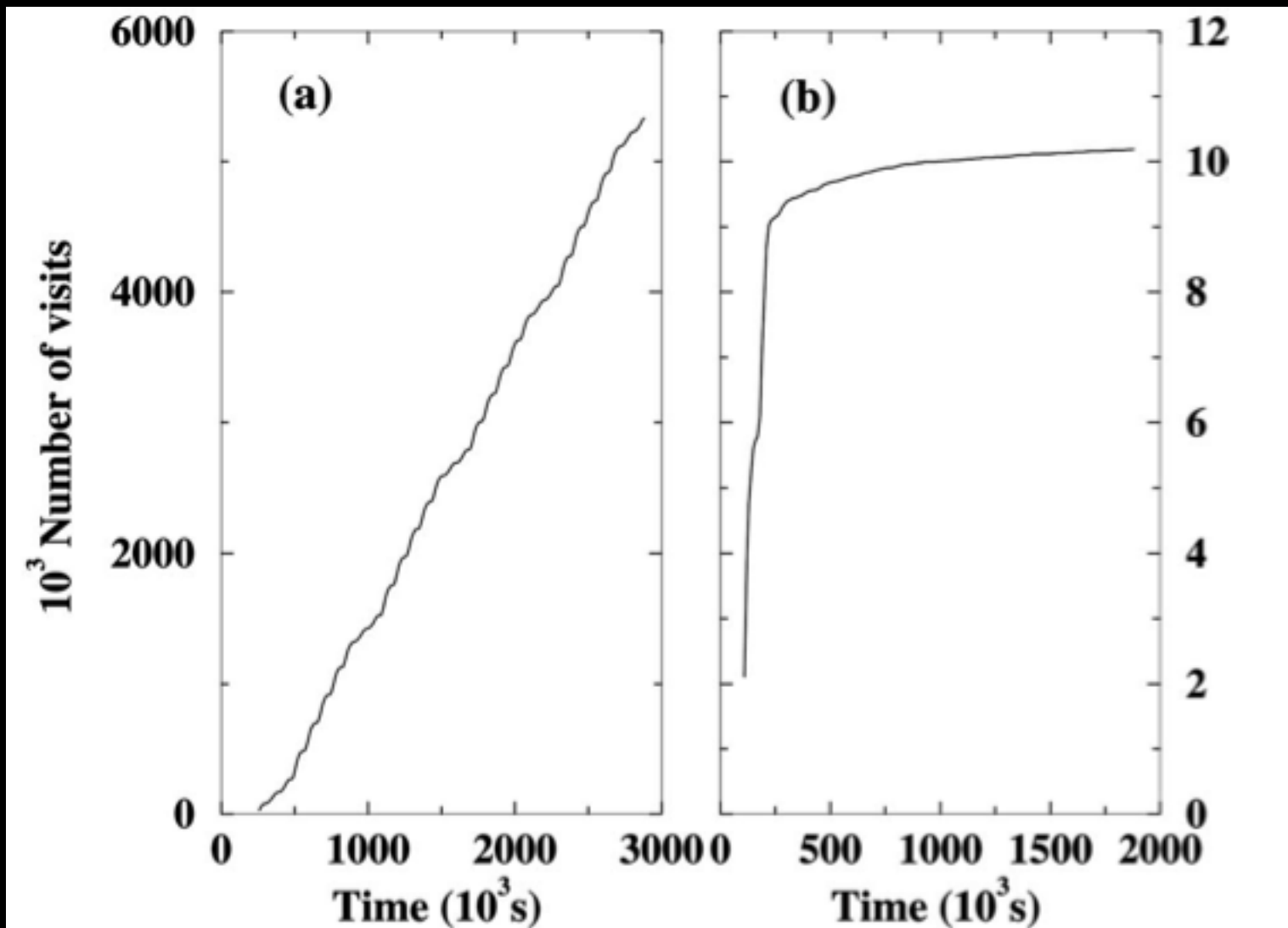
# Link popularity on Wikipedia



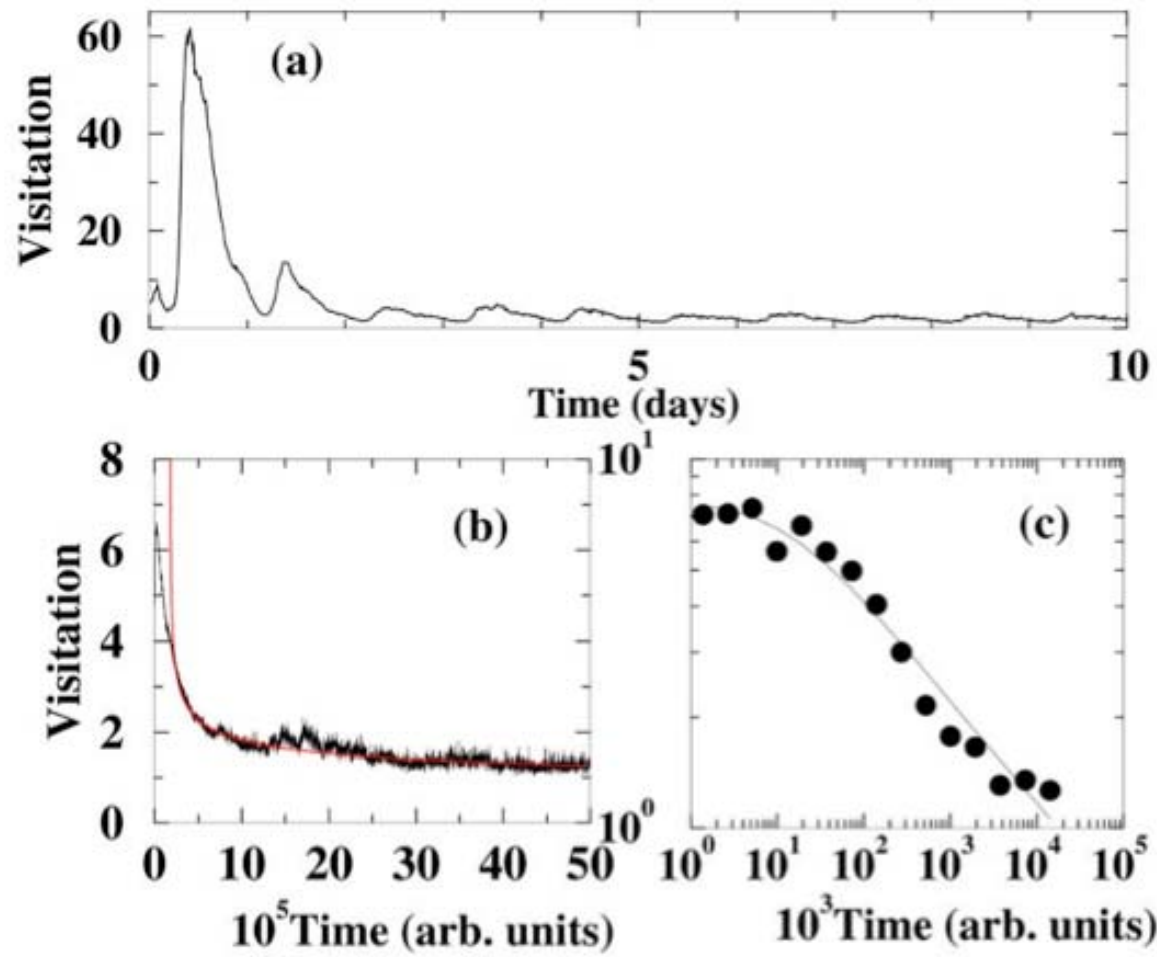
# Another popularity measure: user traffic



# Some background: news popularity



# Some background: news popularity



# Summary of news popularity

- News popularity is, almost by definition, short-lived
- Access to news items significantly decays after 36 hours from posting

# Our data

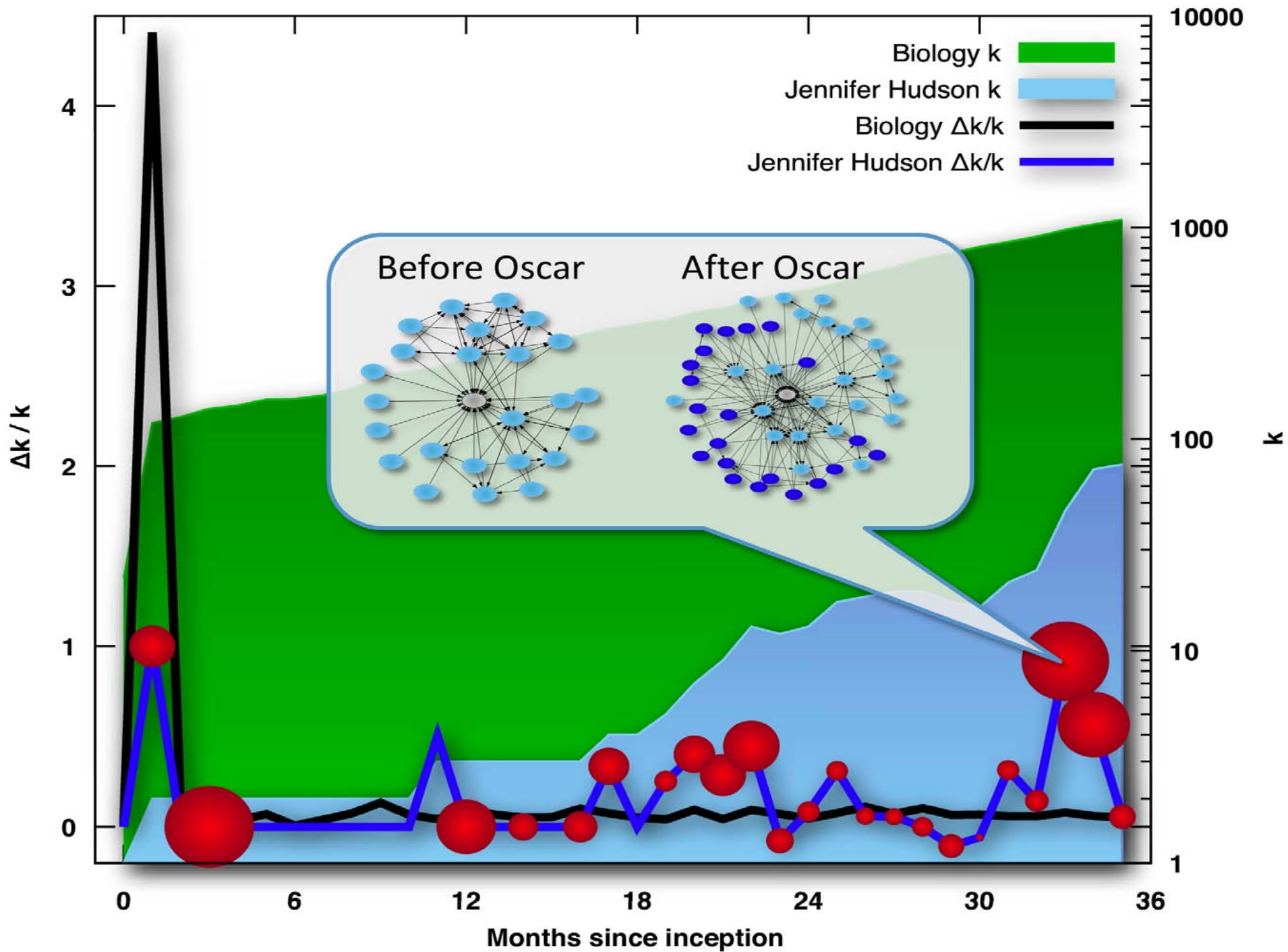
- Wikipedia history (in several languages) from 2001 till March 2007
- Web traffic data from users of Indiana University
- Yearly sequence of crawls of the Chilean Web

# The analysis

*J. Ratkiewicz, F. Menczer, S. Fortunato, A. Flammini,  
A. Vespignani, submitted to HyperText 2010*

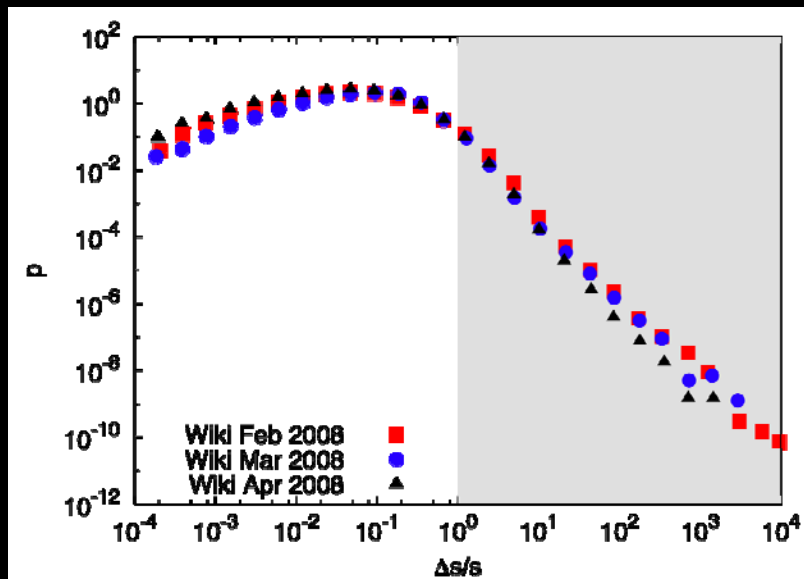
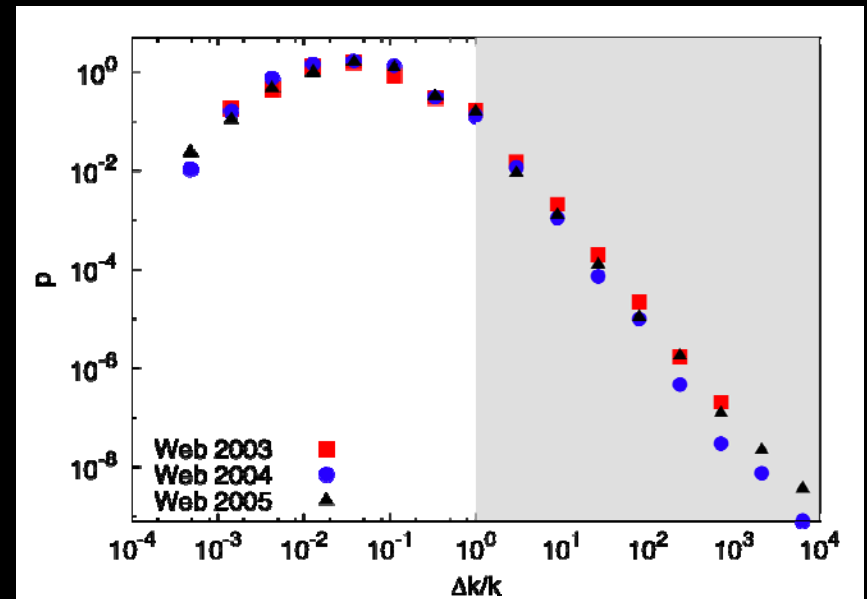
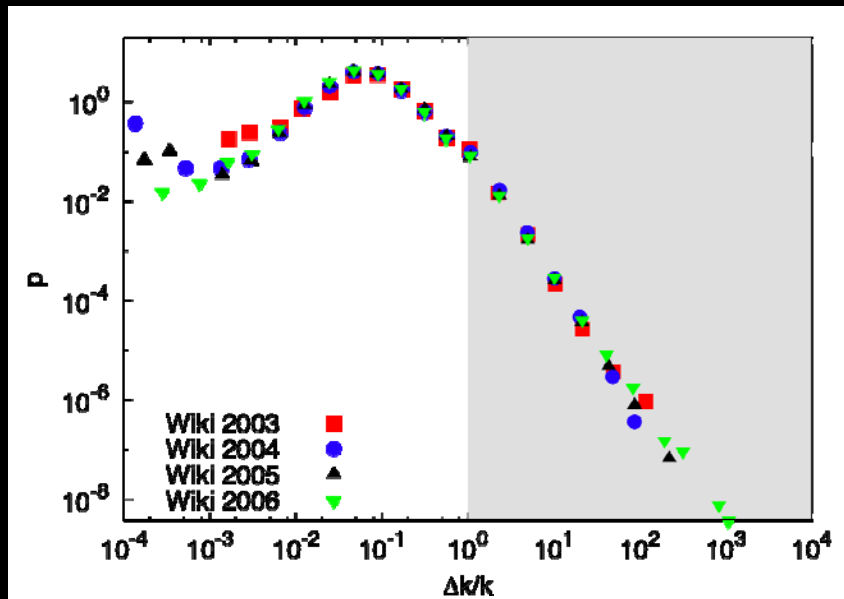
Popularity is measured in terms of indegree and user traffic

Popularity dynamics is studied via the *relative increment*  $\Delta x/x$  of the chosen popularity measure  $x$



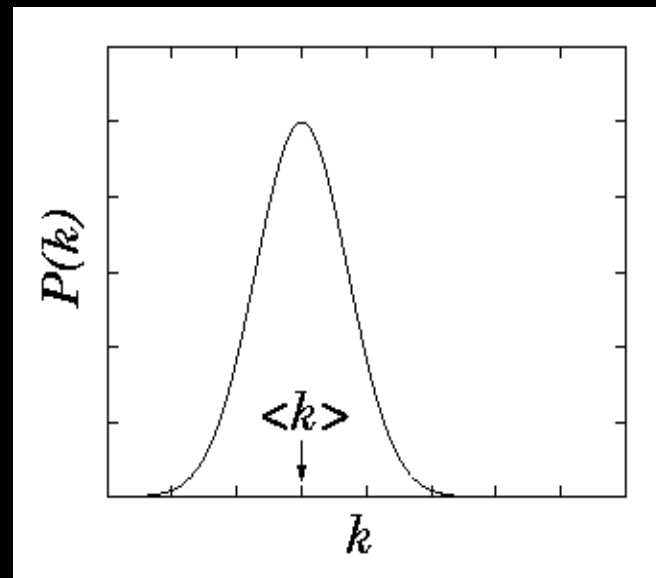


# Distributions: fat tails!



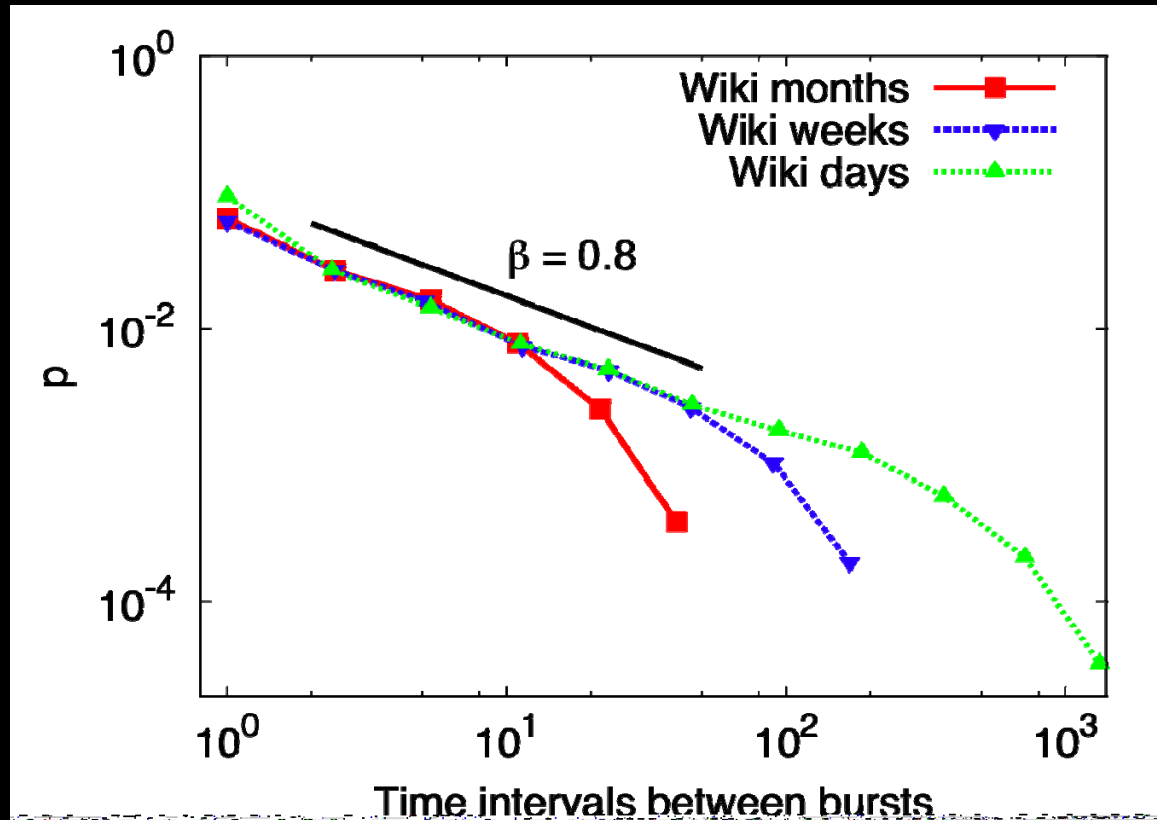
# Inter-event time distributions

Independent events: events that have the same probability to occur regardless of other events



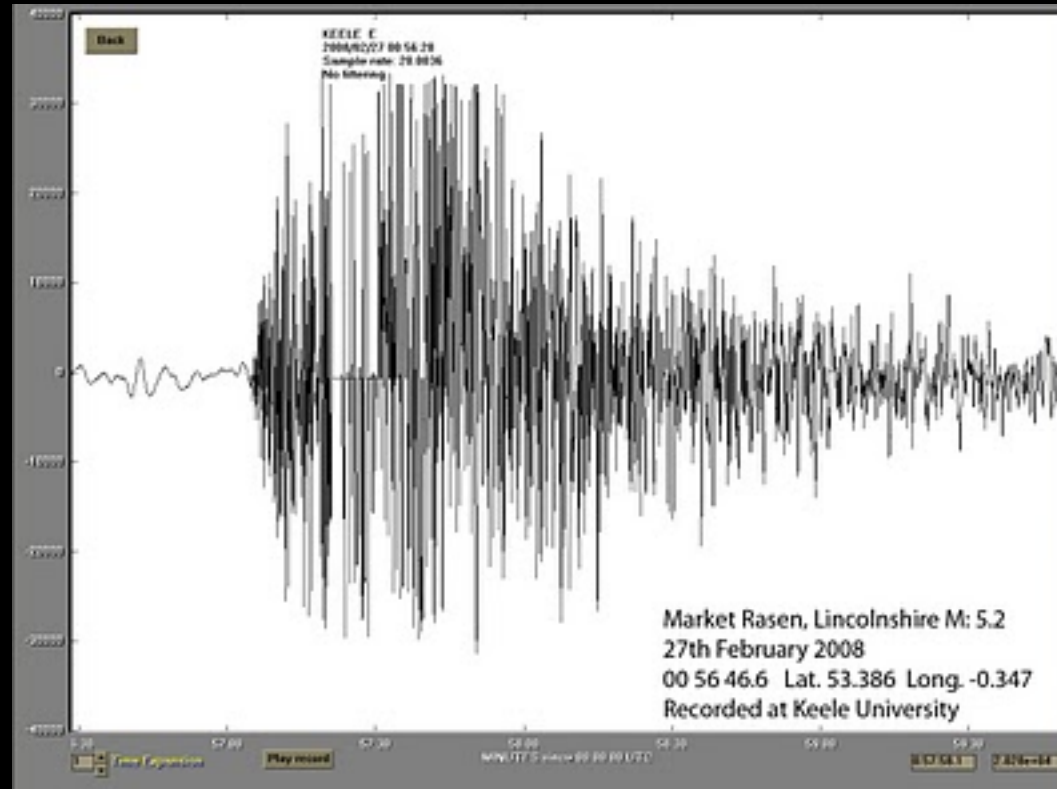
Poissonian behavior!

# Inter-event time distributions



Non-Poissonian behavior (power law):  
events take place also after long times!

# Popularity bursts like earthquakes!



Omori's law (1894): the frequency of earthquake aftershocks decreases as the reciprocal of the time elapsed since the main shock!

The same happens for online popularity bursts!

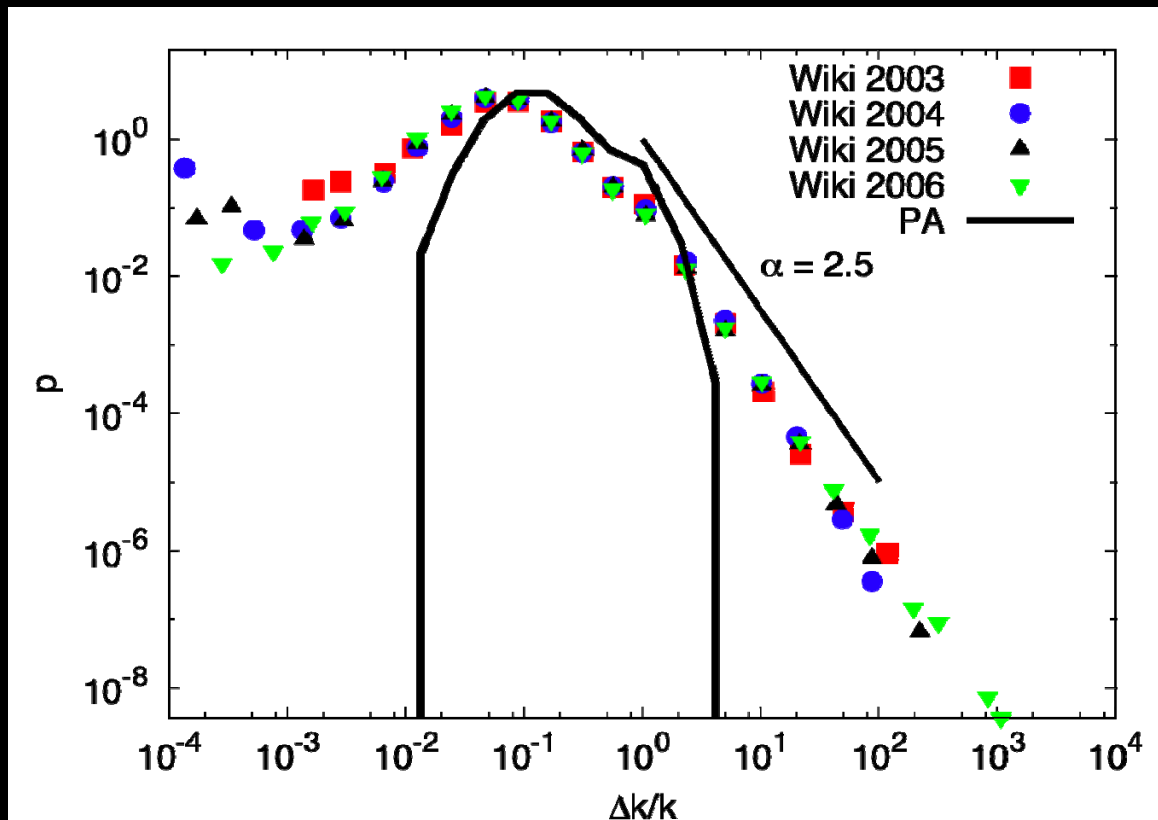
# Cumulative advantage (preferential attachment) ?

D. de Solla Price, *A general theory of bibliometric and other cumulative advantage processes*, J. Amer. Soc. Inform. Sci. 27, 292 (1976)

H. A. Simon, *On a class of skew distribution functions*, Biometrika 42, 425 (1955)

Principle: a page receives a number of inlinks/clicks proportional to the current number of inlinks/clicks

# Preferential attachment unable to explain fat tail!



# The ranking model



Absolute importance of items is often not perceived: ranking is easier!

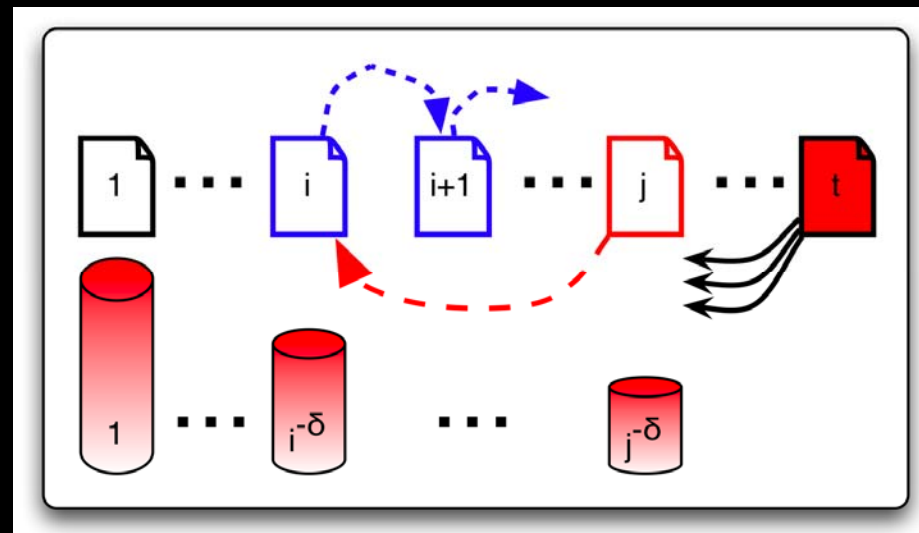
Probability that page  $j$  receives an inlink/click depends on rank of  $j$ :

$$p(i \rightarrow j) \sim R_j^{-\delta}$$

S. F., A. Flammini, F. Menczer,  
PRL 96, 218791 (2006)

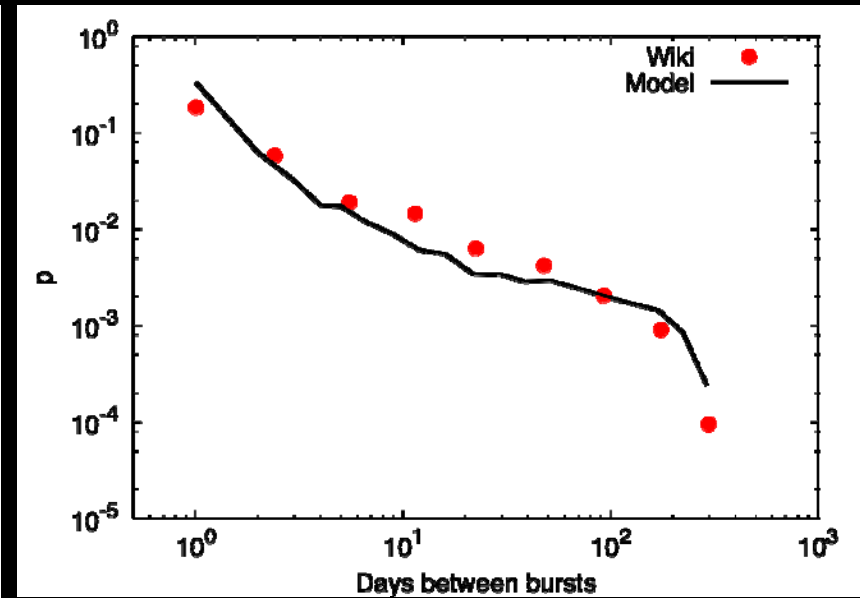
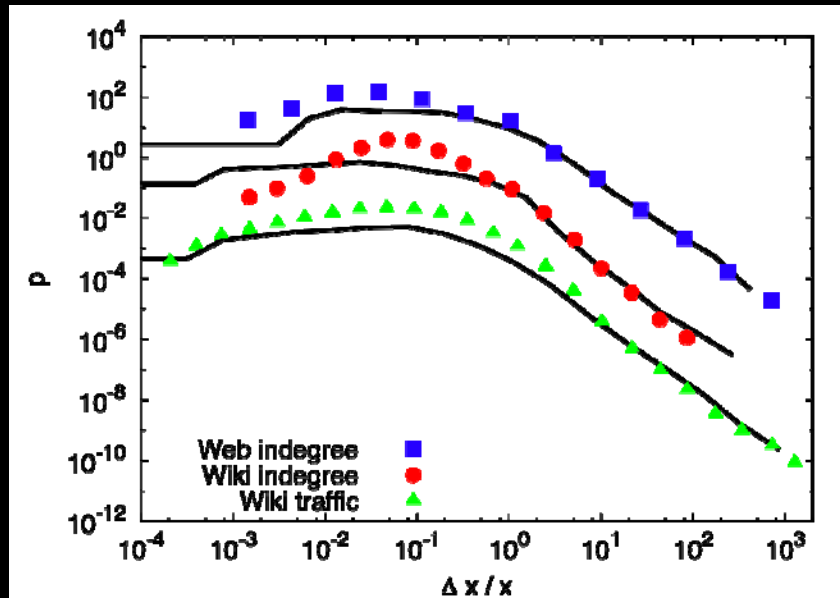
# Rank-shift model

In the rank-shift model, nodes are occasionally re-ranked: at any iteration a randomly chosen node is assigned a new rank, taken at random among all higher ranks





# Rank-shift model describes well the data



# Outlook

- ✓ Popularity dynamics of Websites and Wikipedia pages is "bursty"
- ✓ The size of the bursts is very heterogeneous and their frequency decreases as a power law, just like in earthquakes!
- ✓ Standard cumulative advantage cannot explain the data
- ✓ By modelling bursts as sudden endogeneous/exogeneous variations in the importance of a page we are able to reproduce the data

# Future?

- ✓ Detailed study of burst dynamics
- ✓ Burst prediction?
- ✓ Relationship with seismic phenomena!