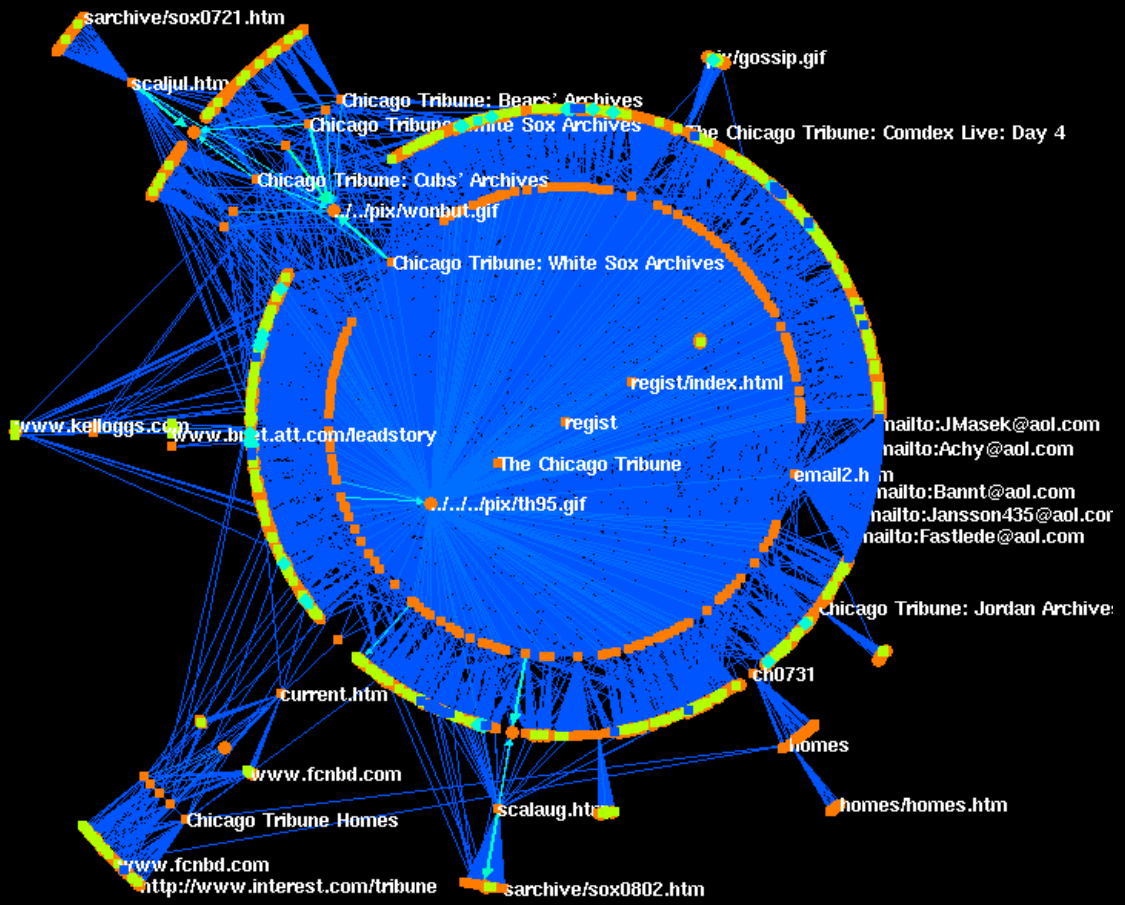
Popularity dynamics on the Web and Wikipedia

Santo Fortunato

The World-Wide-Web

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

- web pages
- hyperlinks

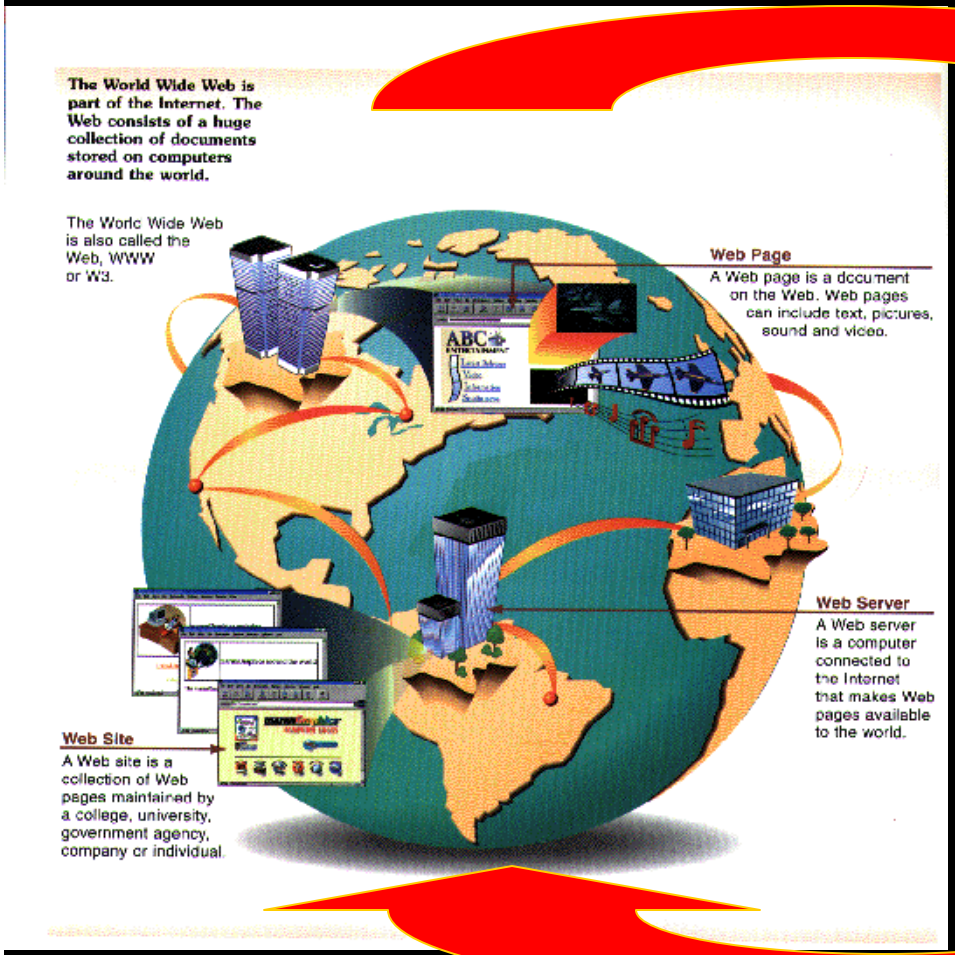


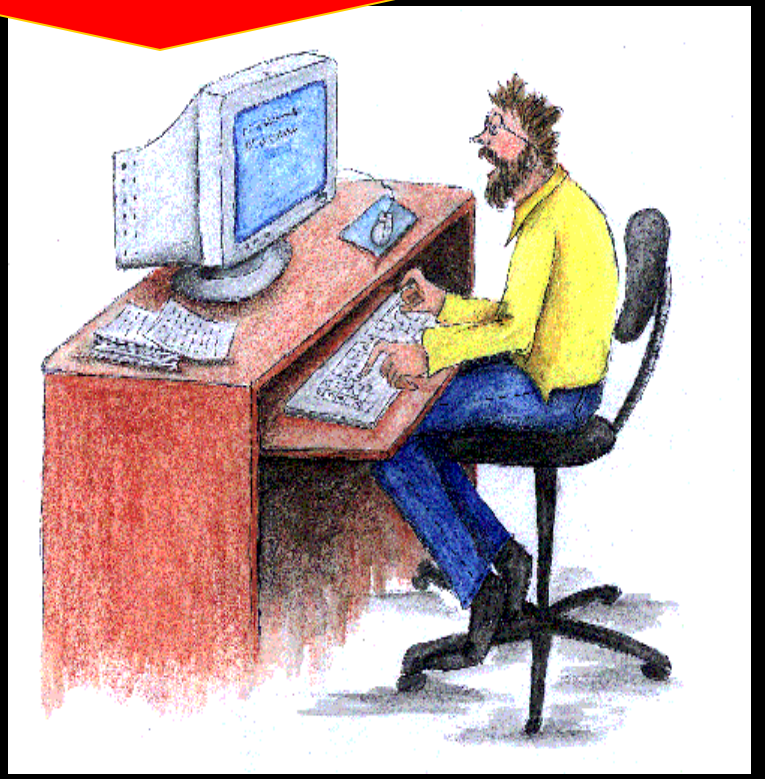
CRAWLS

Wikipedia

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

Websites: a daily fight for visibility!





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

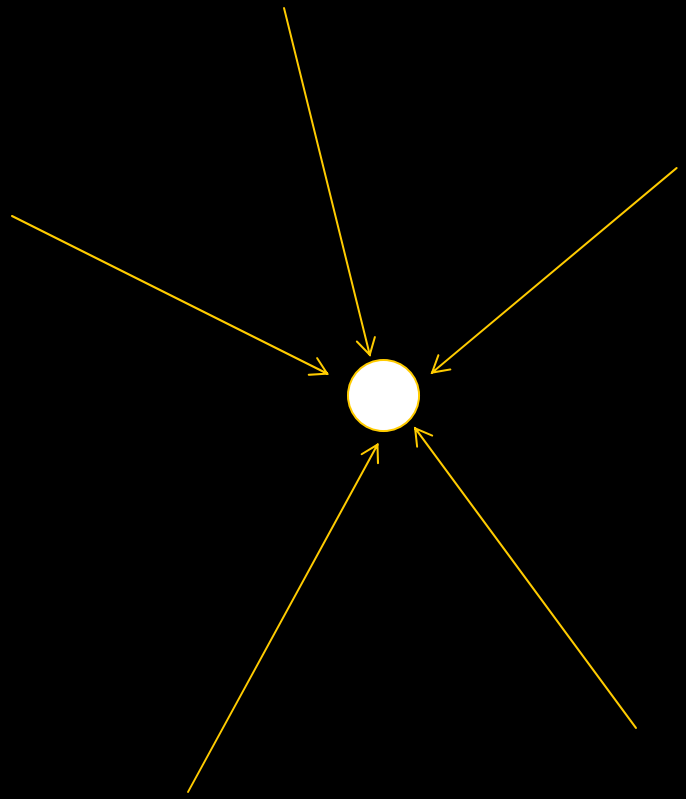
The Web graph

Web pages ➔ vertices/nodes

Hyperlinks ➔ edges/links

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

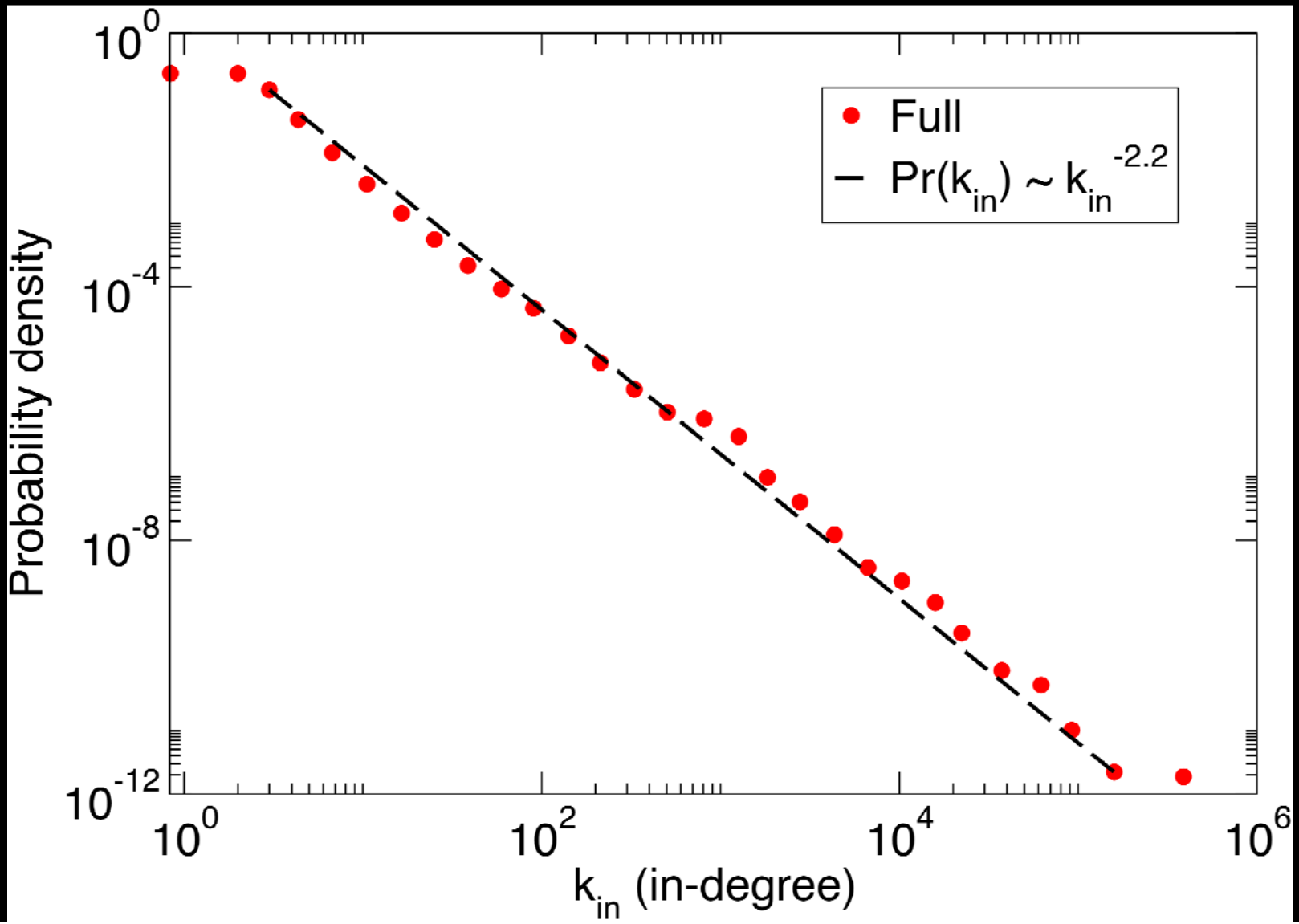
Link popularity on the Web



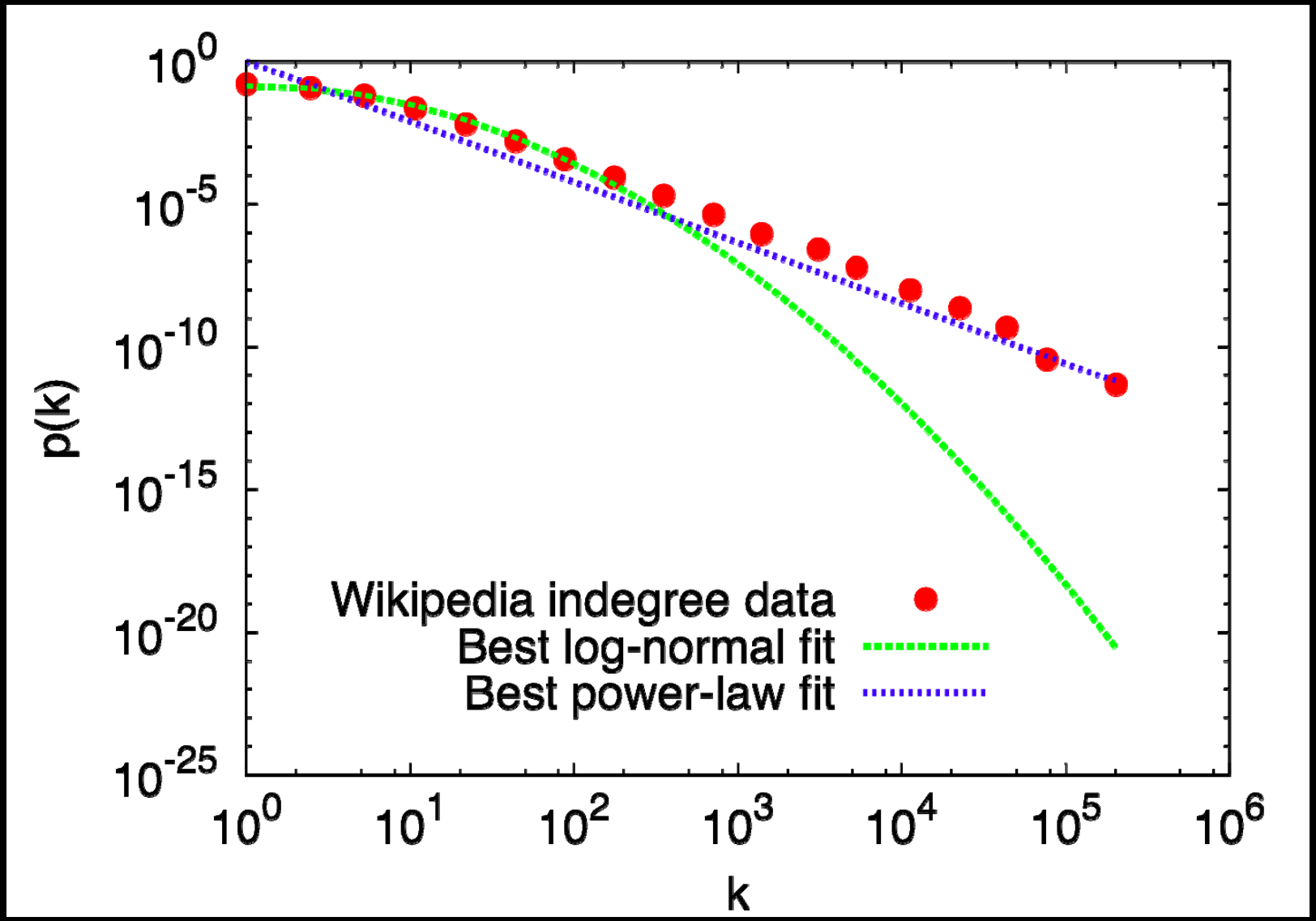
Indegree: number of incoming hyperlinks to a page

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

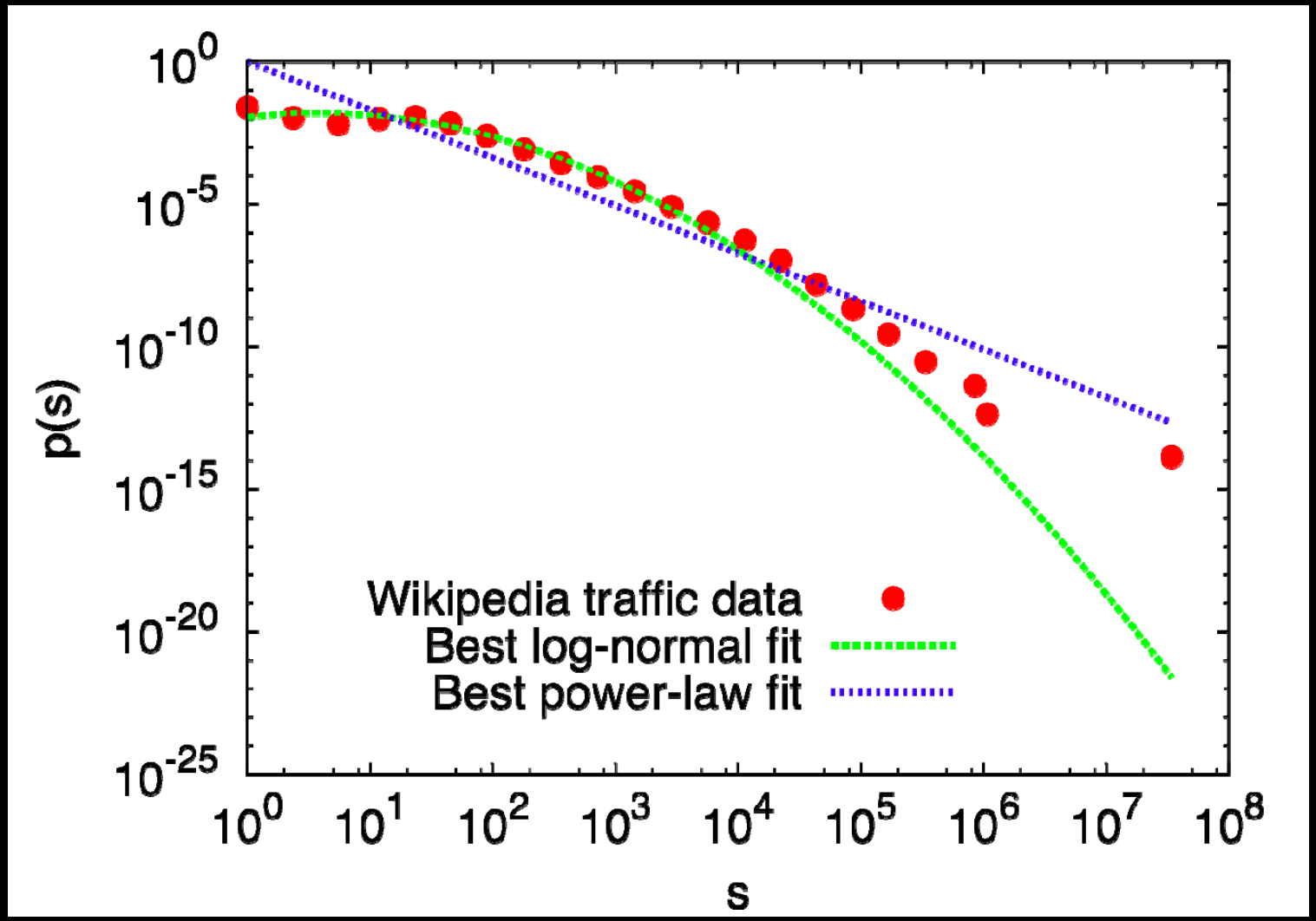
Link popularity on the Web



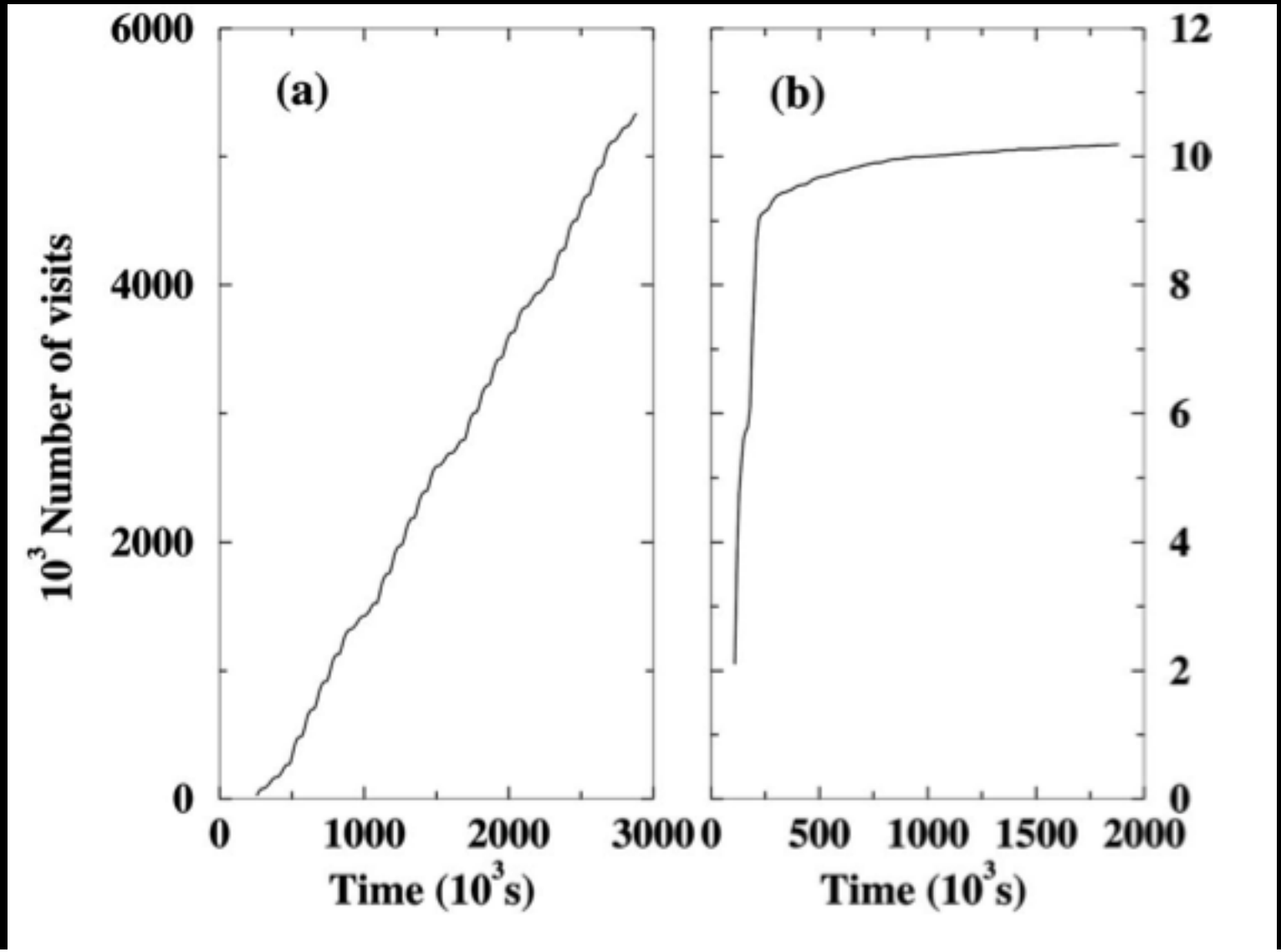
Link popularity on Wikipedia



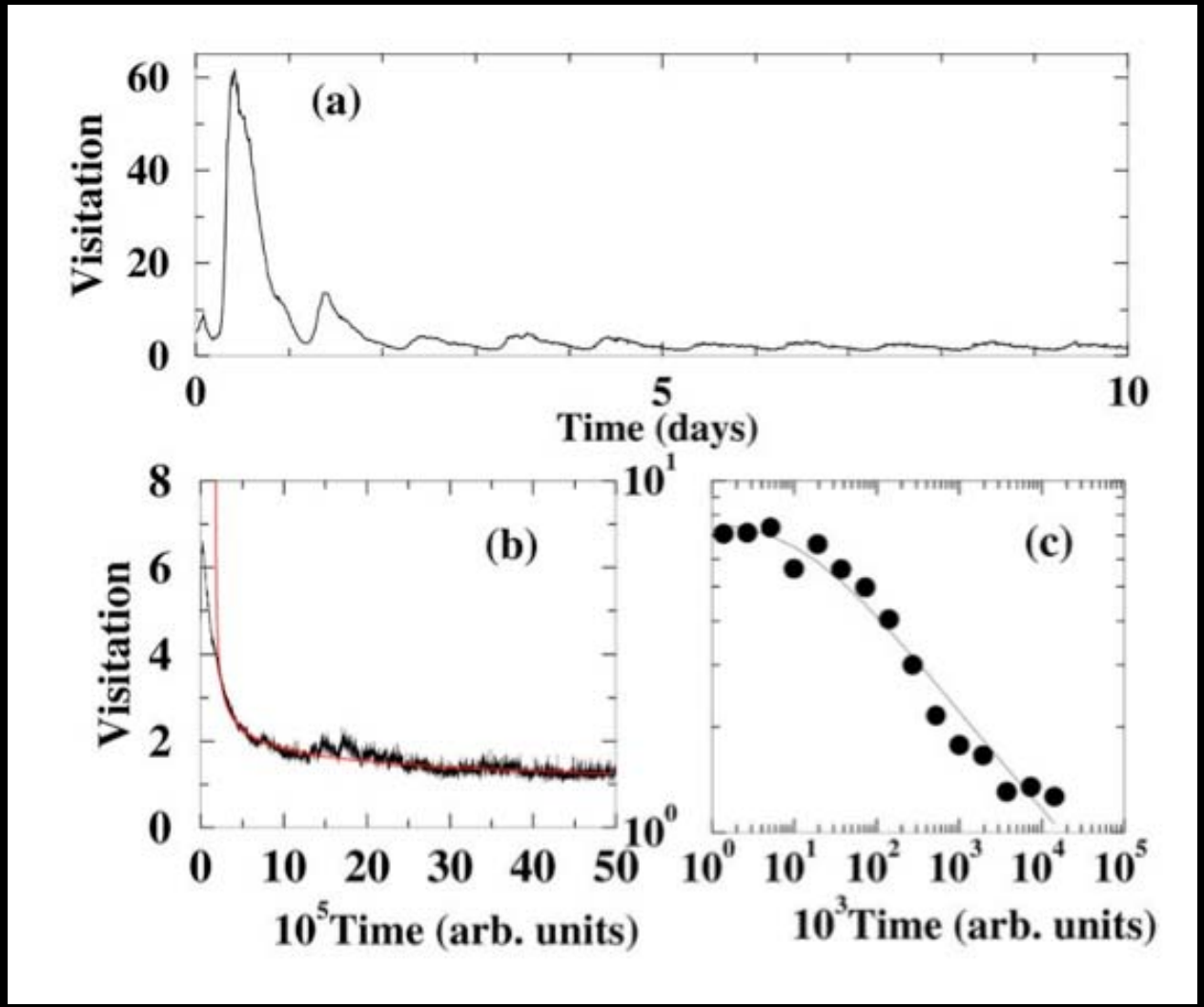
Another popularity measure: user traffic



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

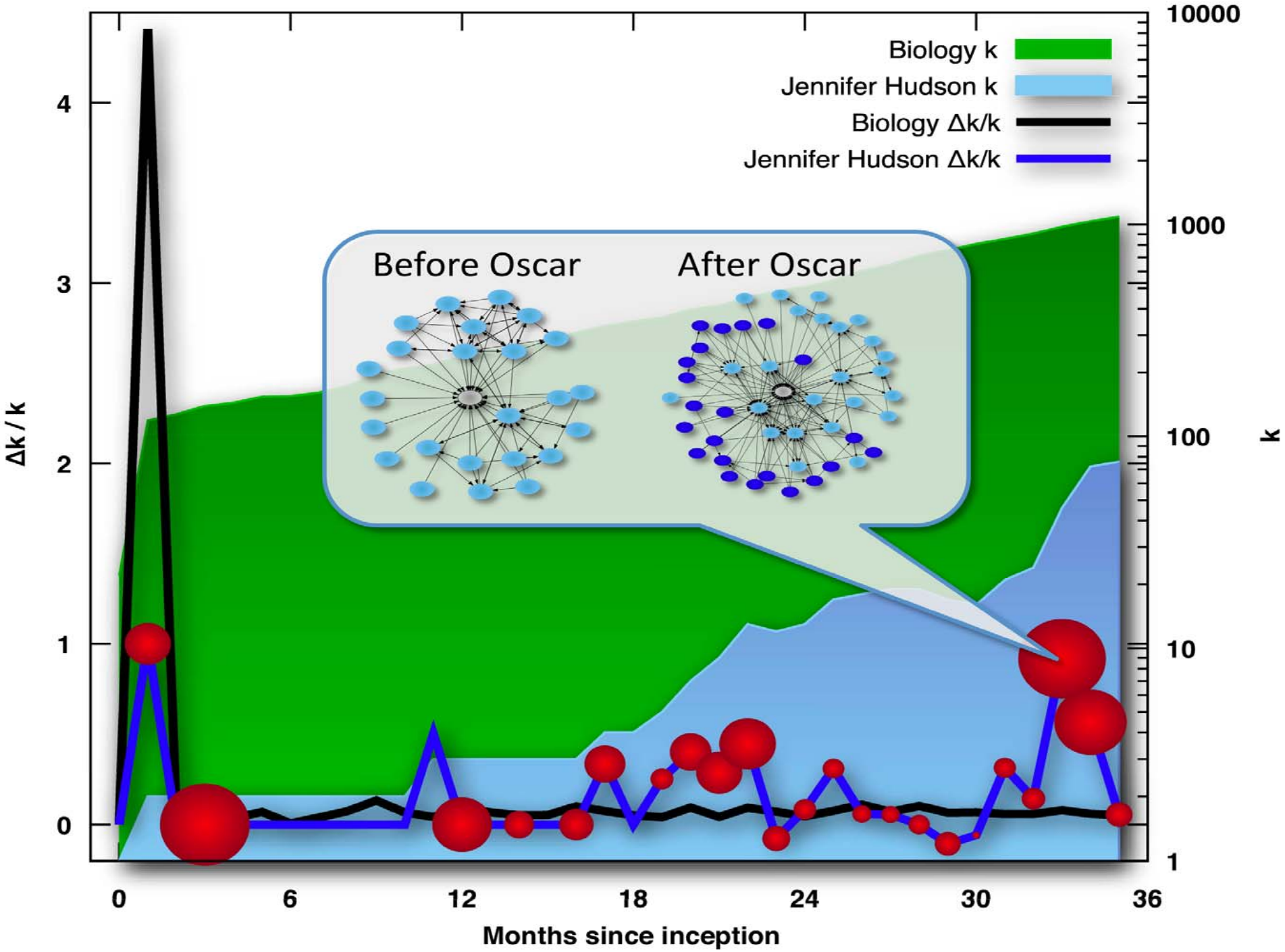
- Wikipedia history (in several languages) from 2001 till March 2007
- Web traffic data from users of Indiana University
- Yearly sequence of crawls of the Chilean Web

The analysis

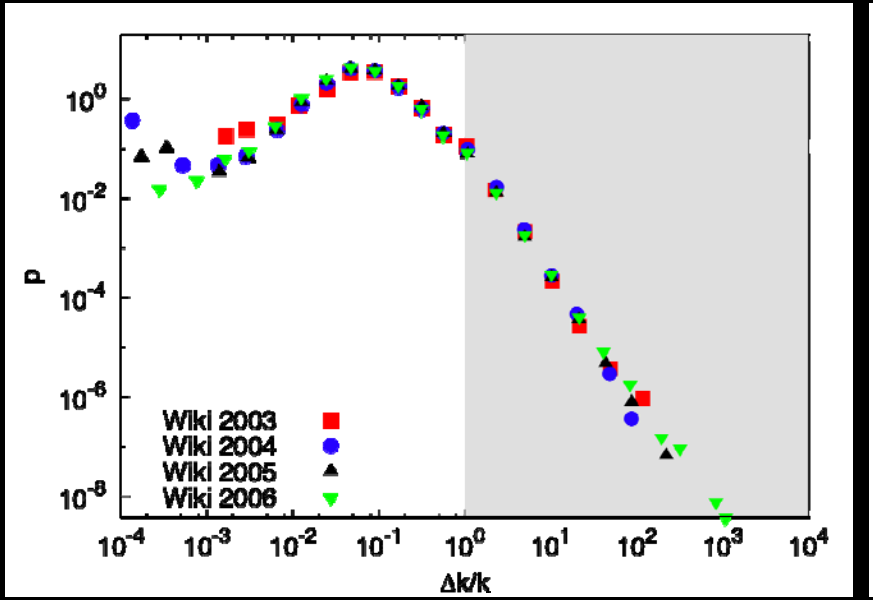
J. Ratkiewicz, F. Menczer, S. Fortunato, A. Flammini, A. Vespignani, submitted to HyperText 2010

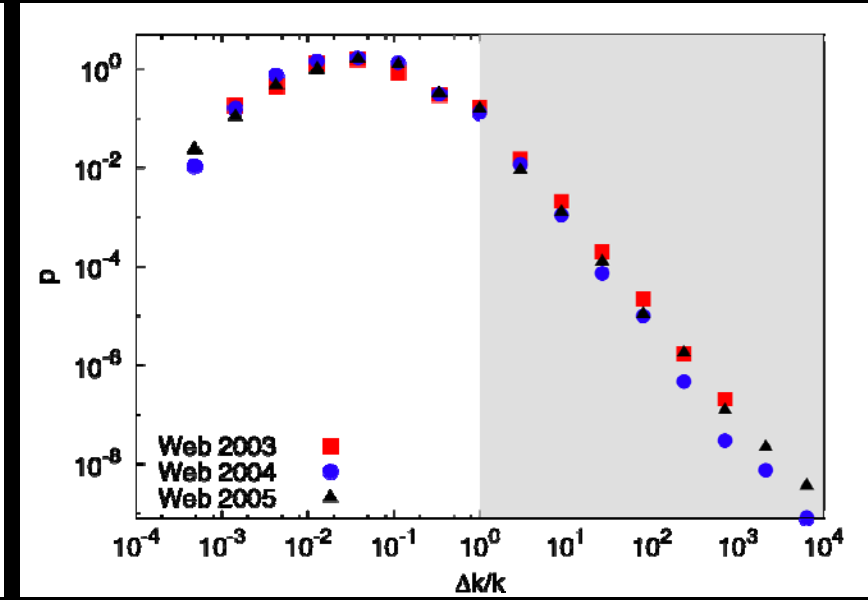
Popularity is measured in terms of indegree and user traffic

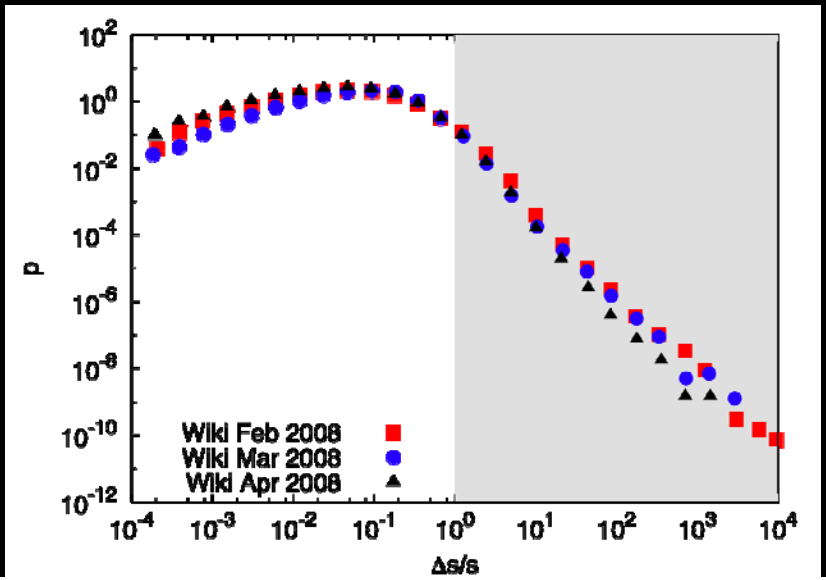
Popularity dynamics is studied via the *relative increment* $\Delta x/x$ of the chosen popularity measure x


Biology k
Jennifer Hudson k
Biology Δk/k
Jennifer Hudson Δk/k
Before Oscar
After Oscar
Δk / k
k
Months since inception
0
6
12
18
24
30
36
1
10
100
1000
10000
2
3
4

Distributions: fat tails!

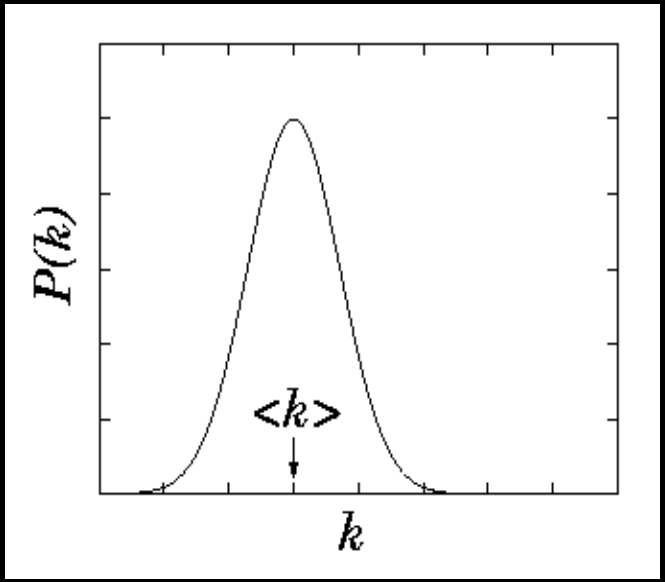






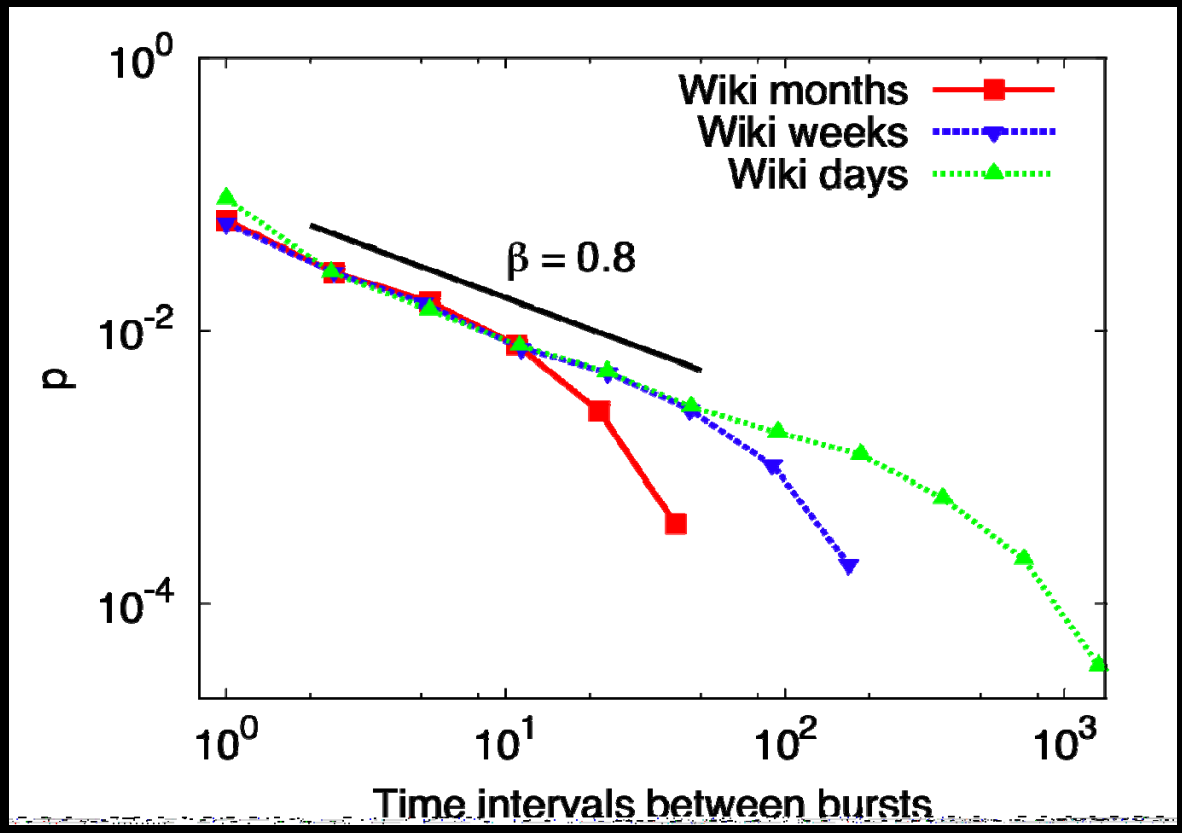
Inter-event time distributions

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



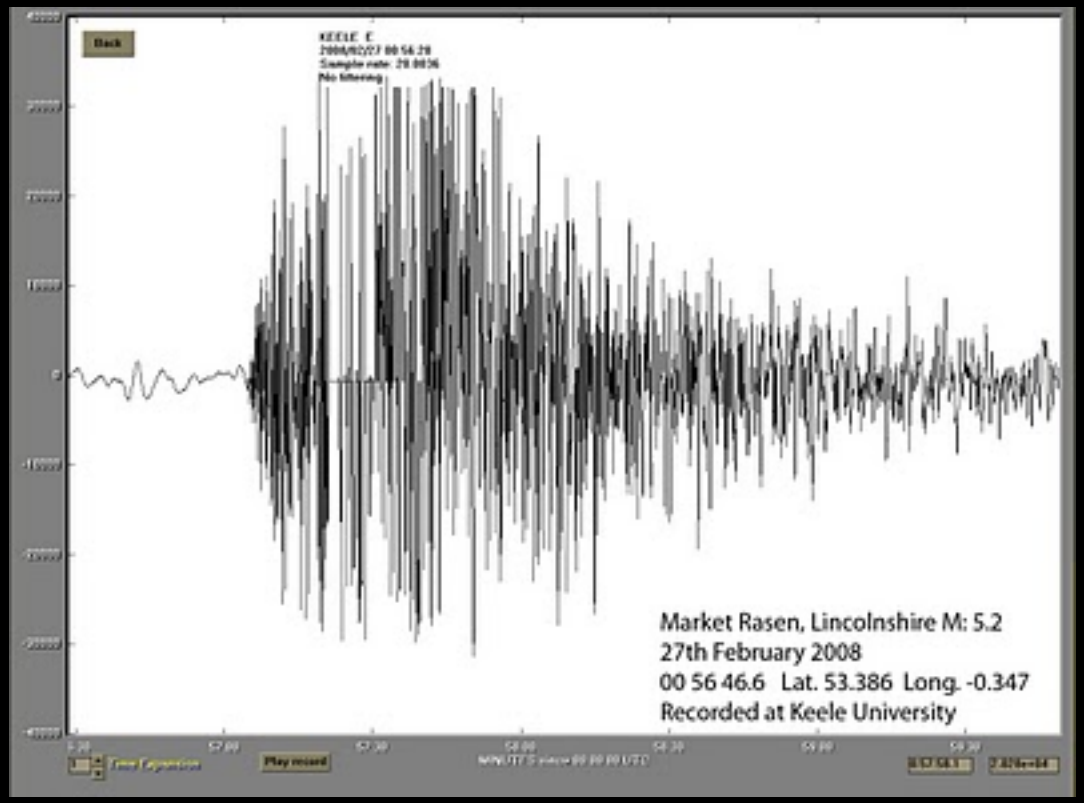
Poissonian behavior!

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!

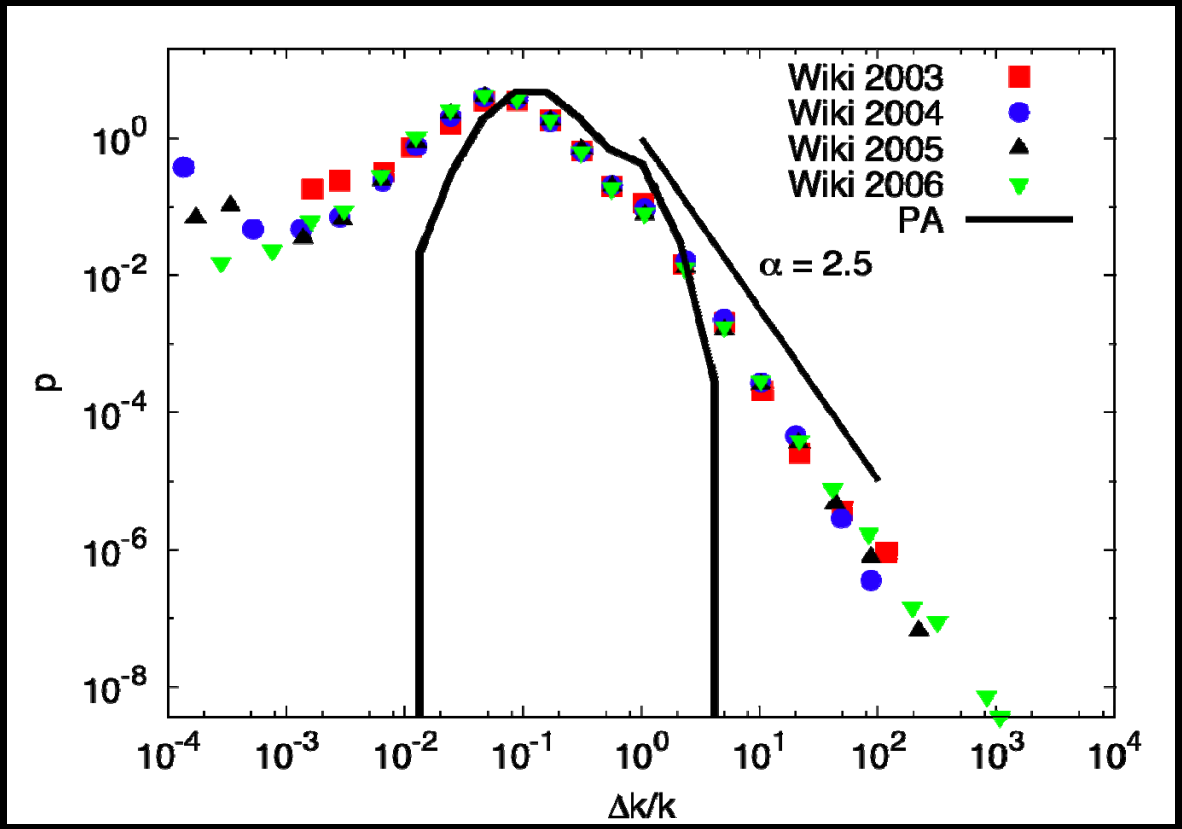
Cumulative advantage (preferential attachment) ?

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!

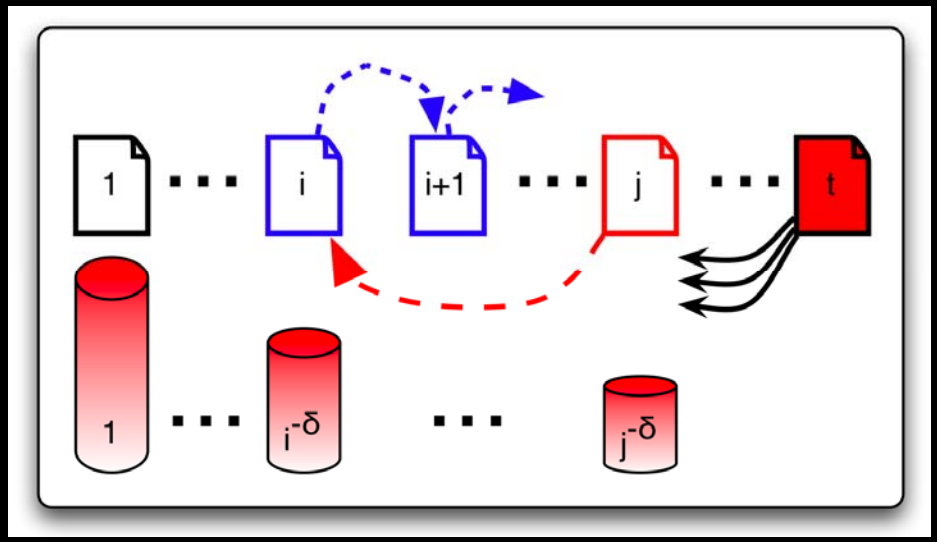
Probability that page j receives an inlink/click depends on rank of j:

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

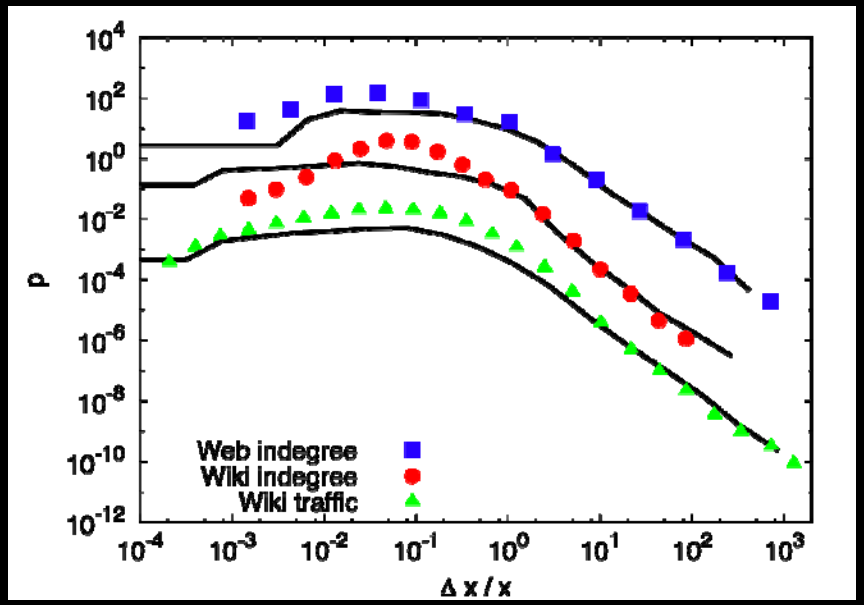
S. F., A. Flammini, F. Menczer, PRL 96, 218791 (2006)

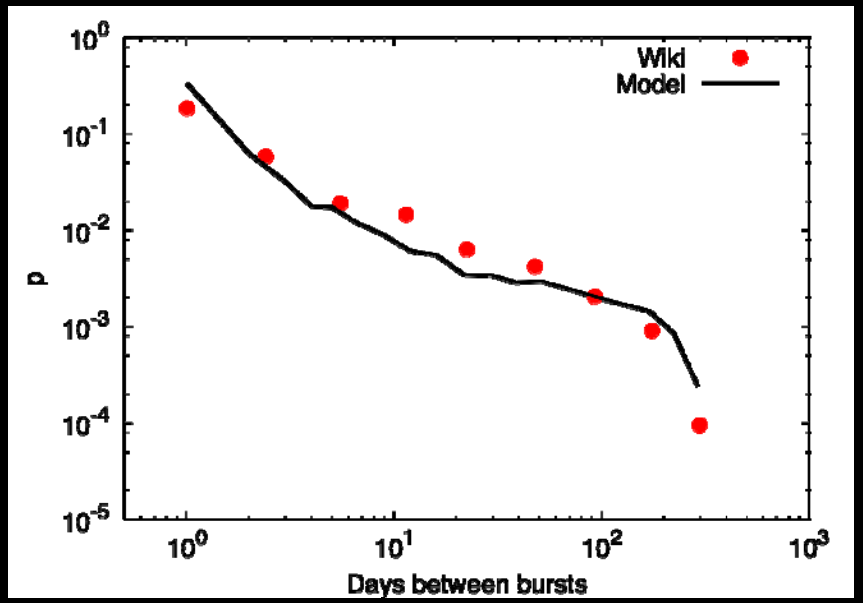
Rank-shift model

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!