Large-Scale Networks

Power laws

Dr Vincent Gramoli | Senior lecturer
School of Information Technologies

THE UNIVERSITY OF SYDNEY

THE UNIVERSITY OF SYDNEY

Introduction

- We talked about information cascade
  - Information cascade can depend on the outcome of few initial decisions
  - A technology can win because it reaches an audience before its competitors

- Let us talk now about a model of uncertain evolution
  - Various quantities, like popularity, have highly skewed distributions
  - How can this be explained?

THE UNIVERSITY OF SYDNEY

Introduction

- Popularity
- Power law
- The preferential attachment model
- Unpredictability
- The long tail
- Search tools and recommendation systems

Popularity

Popularity

- Popularity is a phenomenon characterized by extreme imbalances
  - While most people are known by people in their immediate social circles
  - Very few people achieve global name recognition

- How can we quantify these imbalances?
- Why do they arise?

- It is hard to answer these questions for people popularity
- Let's try to answer these questions for Web page popularity

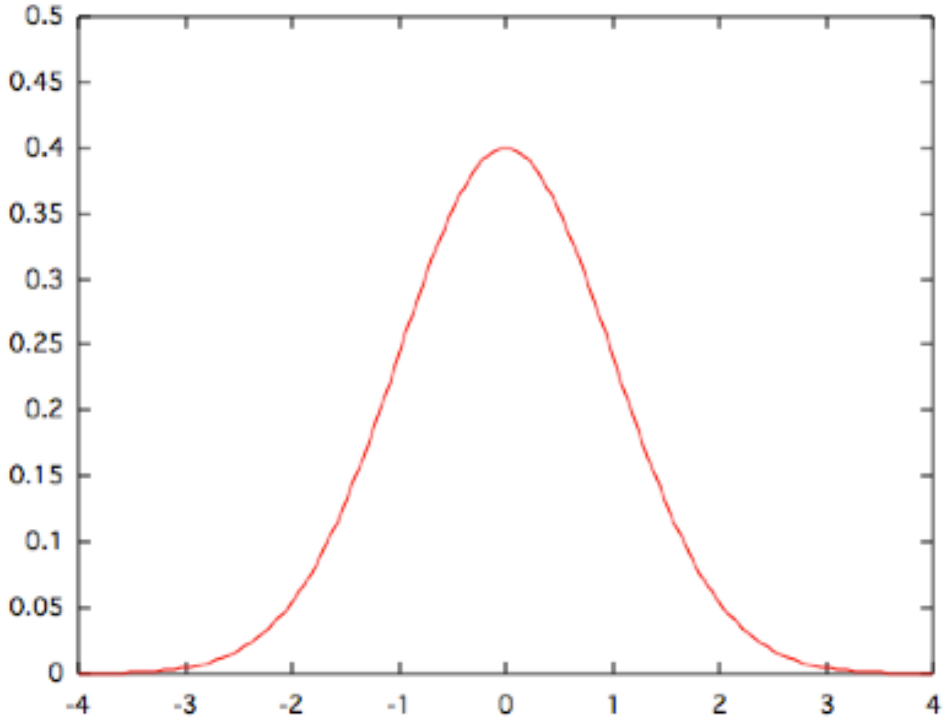
THE UNIVERSITY OF SYDNEY

Popularity

- We refer to the links pointing to a given page as the *in-links* of the page
- Let us define the *popularity of a Web page* as the number of its in-links
- Consider the distribution of the number of in-links
  - As a function of k, what fraction of pages on the Web have k in-links?
  - k translates into popularity: the higher k is, the more popular the page

Popularity

› What could be the distribution?



› The normal distribution is popular (cf. Central Limit Theorem)

- The sum of independent random quantities follows the normal distribution
- k would be normally distributed if pages would connect independently at random

Power Law

THE UNIVERSITY OF SYDNEY

Power law

› The distribution of links on the Web is different from a normal distribution [BKM00]

- The fraction of Web pages that have in-links is approximately $1/k^2$

(More precisely, the exponent on k is generally a number slightly larger than 2)

› How does it differ from the normal distribution?

- $1/k^2$ decreases much more slowly as k increases than in the normal distribution

(there are more pages very high numbers of in-links)

- $1/k^2$ is only 1/1000000 for k=1000, while $1/2^k$ is unimaginably low

› A *power law* is a function that decreases as k increases to some fixed power, such as $1/k^2$ in the present case

⇒ There is an extreme imbalance in the distribution of in-links on Web pages

Power law

- Similar power laws exist:
  - The fraction of telephone numbers that receive k calls per day is $O(1/k^2)$
  - The fraction of books that are bought by k people is $O(1/k^3)$
  - The fraction of scientific papers that receive k citations is $O(1/k^3)$
  - …

- Hence, if someone gives you a table showing the number of monthly downloads for each song at a large online music site, then it is worth testing whether it is approximately a power law $1/k^c$ for some c, and if so to estimate c

THE UNIVERSITY OF SYDNEY

Power law

› How to test that some dataset is power law?

- Given a table showing the number of monthly downloads for each song at a large online music site, how do you test whether it is approximately a power law $1/k^c$ for some c, and if so how do you estimate c?

› Let f(k) be the fraction of items that have value k

› Suppose you want to know whether the equation $f(k) = a/k^c$ approximately holds

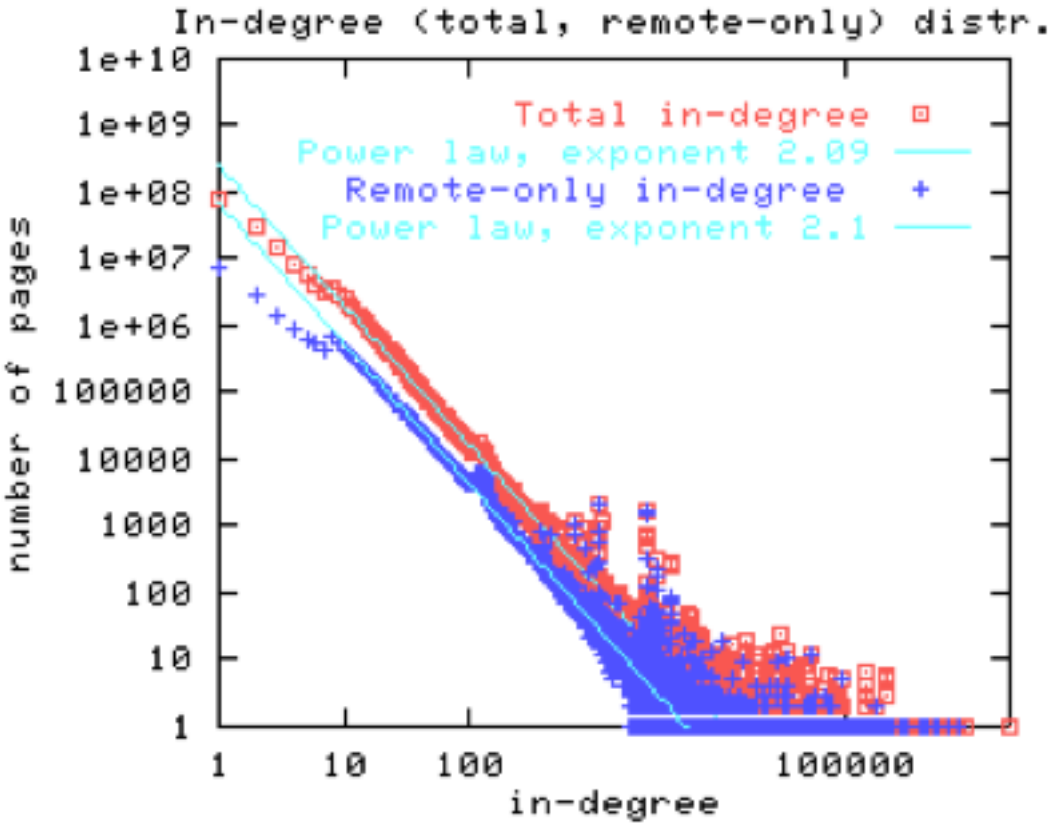
› Note that $f(k) = ak^{-c}$ and if we apply the log to both sides we have:

$$\log f(k) = \log a - c \log k.$$

⇒ If we plot *log(f(k))* as a function of *log(k)* then we should have a straight line with *c* the slope and *log a* the y-intercept

Power law

› A power law distribution, like the Web page in-links, shows up as a straight line on a log-log plot



The Preferential Attachment Model

The preferential attachment model

- Here is a simple model for the creation of links among Web pages
  1. Pages are created in order and named 1, 2, 3, …, N
  2. When page j is created, it produces a link to an earlier Web page by choosing between actions (a) and (b) below according to the following probabilistic rule (controlled by a single number p between 0 and 1):
     a) With probability p, page j chooses a page i uniformly at random from among all earlier pages and creates a link to i
     b) With proba 1-p, page j instead chooses a page l uniformly at random from among all earlier pages and creates a link to the page i points to
     c) This described the creation of a single link from page j; one can repeat this process to obtain multiple, independently generated links from page j

THE UNIVERSITY OF SYDNEY

The preferential attachment model

- If we repeat for many pages, the fraction of pages with k in-links follow a power law distribution $1/k^c$

- Step 2(b) is the key: j copies the behavior of node i instead of linking l

- We could have replaced Step 2(b) by:

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

⇒ The probability that l's popularity increases is proportional to l's popularity

⇒ *Preferential attachment*: links are formed "preferentially" to pages that already have high popularity [BA99]

The preferential attachment model

- Popularity grows at a rate proportional to its current value

  ⇒ Popularity grows exponentially with time

- The populations of cities have been observed to follow a power law dist.
  - The fraction of cities with population k is roughly 1/kc for some constant c [Sim55]
  - If we assume that cities are created at different times and grow at in proportion to its current size simply as a result of people having children, then we have almost precisely the same model

- Note that there are other classes of simple model designed to capture power-law distributions

Unpredictability

THE UNIVERSITY OF SYDNEY

Unpredictability

- Salgankik, Dodds, Watts created a music download site [SDW06]
  - With 48 obscure songs of varying quality written by actual performing groups
  - Visitors were presented with a list of the songs and could listen to them
  - Each visitor would see a “download count” for each song
  - At the end, the visitor was proposed to download the song that she liked

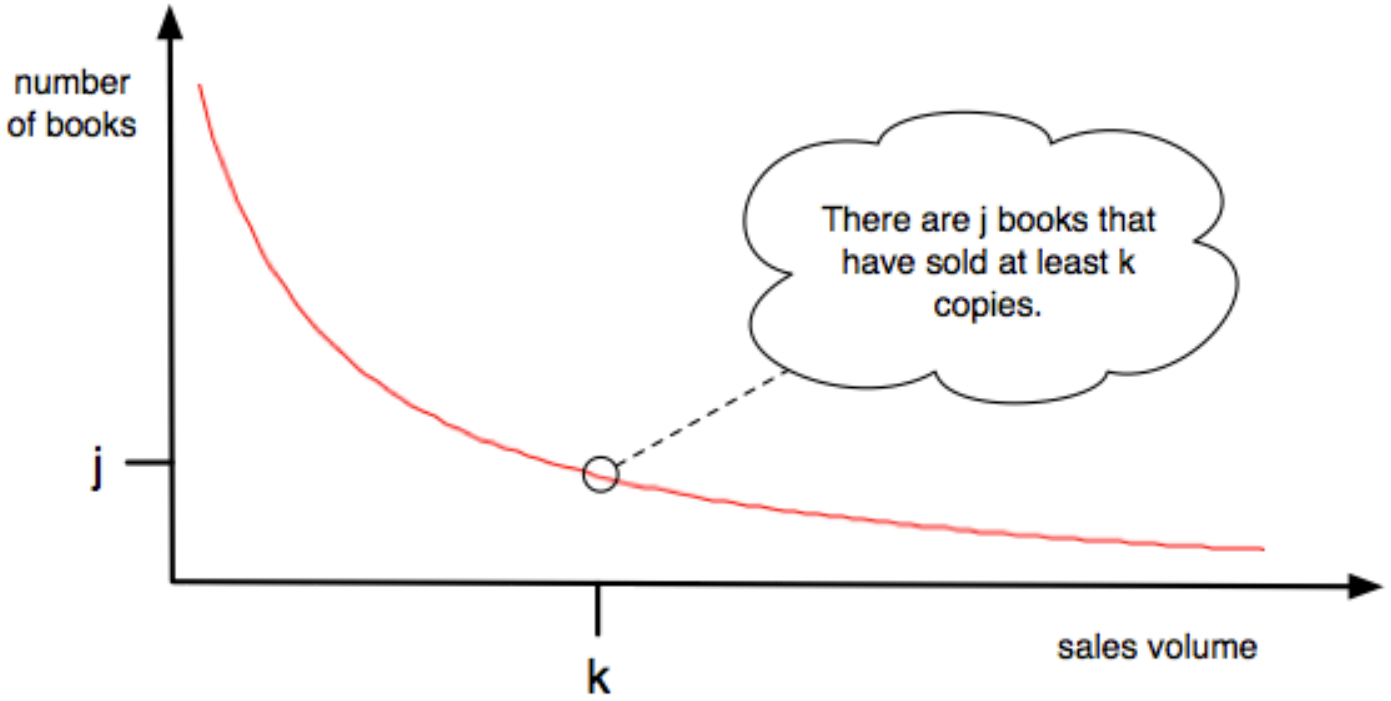
- Upon arrival, visitors were redirected to one of the 8 copies of the site
  - These copies were initially the same with download count set to 0
  - These copies evolved differently as visitors arrived

- The market share of the different songs varied considerably across the different parallel copies
- Even though the best (resp. worst) song was never at the bottom (resp. top)

The Long Tail

The long tail

- Are most sales generated by:
  1. Few items that are enormously popular or
  2. Many items that are each individually less popular?



As a function of k, what fraction of items have popularity exactly k?

THE UNIVERSITY OF SYDNEY

The long tail

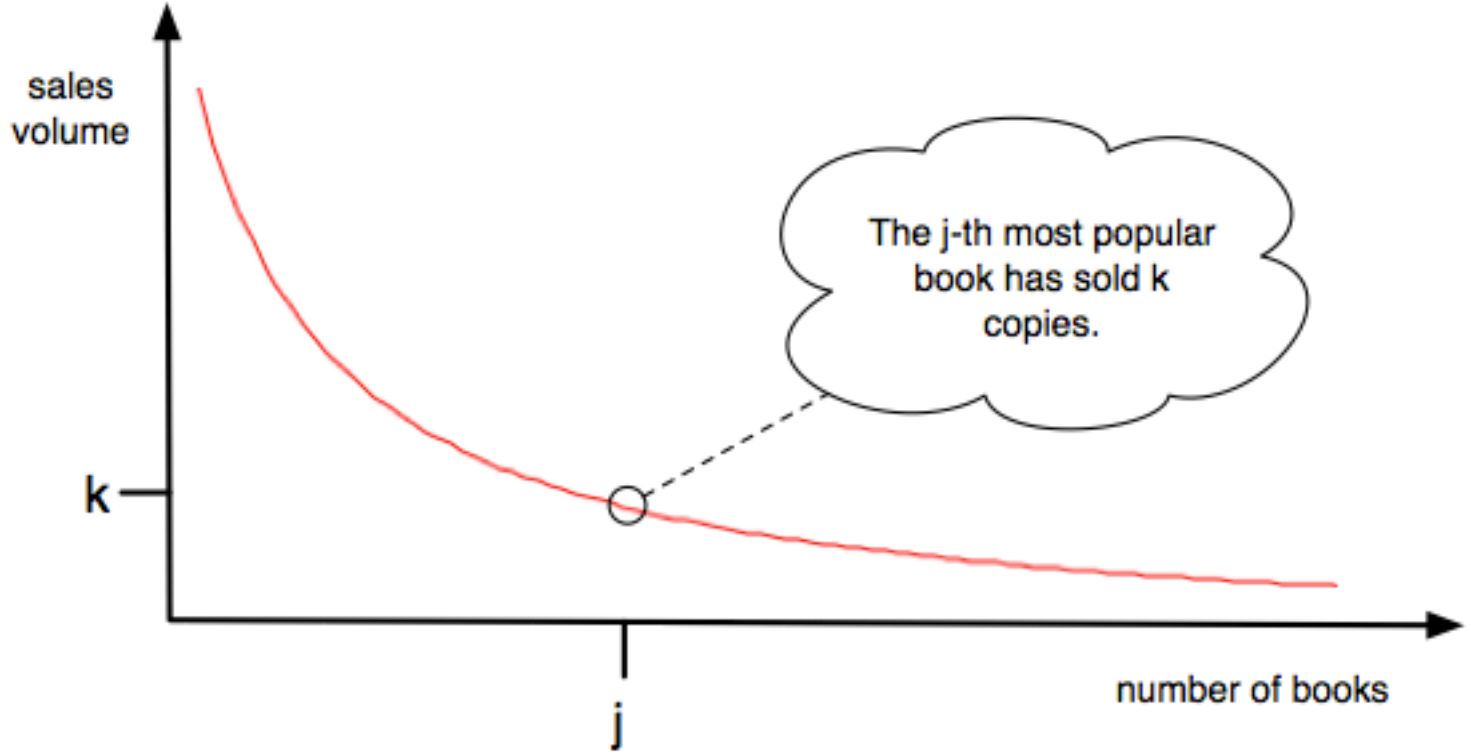
- Chris Anderson argued that Internet-based distribution and other factors were driving the media and entertainment industries toward (2) [And04]

- Amazon and Netflix carry huge inventories (w/o physical store constraints) making it feasible to sell an astronomical diversity of products even when very few of them generate much volume on their own.

- We are now viewing things out the opposite end of the telescope

The long tail

› Let us swap the two axes to get a *Zipf plot* in which x-axis is rank rather than popularity [Zipf49]



As a function of k, what number of items have popularity at least k?

Search Tools and Recommendation Systems

THE UNIVERSITY OF SYDNEY

The long tail

- Copying links to popular pages increases popularity and imbalances
- People using Google search engine will be biased when finding pages to copy from, hence strengthening imbalances
- However, queries are user-specific and users may be led towards pages they would never find by simply browsing
- Finally, in order to make money, a giant inventory must help its users find non popular items as well:
  - We can see the Amazon, Netflix recommendation systems made for this purpose
  - They propose items based on user similarities rather than item popularity

THE UNIVERSITY OF SYDNEY

References

- [BKM00] Broder, Kumar, Maghoul, Raghavan, Rajagopalan, Stata, Tomkins, Wiener. Graph Structure in the Web. Proc. of International World Wide Web Conference, p. 107-117, 2000.
- [BA99] Barabasi, Albert. Emergence of Scaling in Random Networks. Science, 286:509-512, 1999.
- [Sim55] Simon. On a class of skewed distribution function. Biometrica 42:425-440, 1955.
- [SDW06] Experimental study of inequality and unpredictability in an artificial cultural market. Science, 311:854-856, 2006.
- [And04] Chris Anderson. The long tail. Wired, October 2004.
- [Zipf49] G. K. Zipf. Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology. Addison Wesley. 1949.