Social Network Analysis

Course Outline

- Graph Theory and Social Networks
- Visualizing Social Networks
- **Network Dynamics**
- Information Networks and the World Wide Web
- Game Theory
- Applications of SNA in various domains

Power Laws and Rich-Get-Richer Phenomena

Popularity as a Network Phenomenon

- Popularity is a phenomenon characterised by extreme imbalances:
  - Almost everyone goes through life known only to people in their immediate social circles, a few people achieve wider visibility, and a very, very few attain global name recognition.
  - The same can be said of books, movies, or almost anything that commands an *audience*.
- How to quantify these imbalances in popularity?
- Are they intrinsic to the idea of popularity?

Popularity as a Network Phenomenon

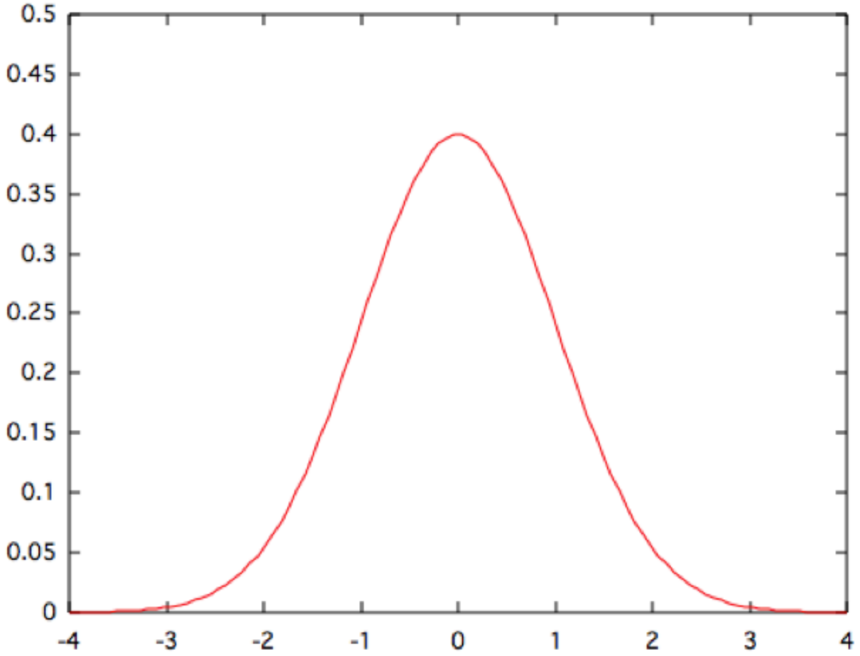
- The number of in-links to a web page is a measure of the page's popularity.

As a function of k,
what fraction of pages on the Web have k in-links?

- Larger values of k indicate greater popularity, so this is precisely the question of how popularity is distributed over the set of Web pages.

- Any educated guesses?

Density of Values in a Normal Distribution



A Normal Distribution is characterised by a “mean” and a “standard deviation” 0 and 1 in the above, respectively.

Popularity as a Network Phenomenon

- The probability of observing a value that exceeds the mean by more than c times the standard deviation decreases exponentially in c.
- A Normal distribution is a natural guess because it occurs in many real-life situations.
- Why does it occur in so many places?
- Central Limit Theorem

  Take any sequence of small independent quantities, then in the limit, their sum (or average) will be distributed according to the normal distribution.

Popularity as a Network Phenomenon

- Central Limit Theorem

  Any quantity that can be viewed as the sum of many small independent random effects will be well-approximated by the normal distribution.

- For example,
  - if one performs repeated measurements of a fixed physical quantity, and if the variations in the measurements across trials are the cumulative result of many independent sources of error in each trial, then the distribution of measured values should be approximately normal.

- How does this apply to web pages?

Popularity as a Network Phenomenon

- How does this apply to web pages?

- If we model the link structure of the Web,
  - assuming that each page decides independently at random whether to link to any other given page,
  - then the number of in-links to a given page is the sum of many independent random quantities
  - (i.e. the presence or absence of a link from each other page),
  - and hence we'd expect it to be normally distributed.

- So, our hypothesis…

The number of pages with k in-links should decrease exponentially in k, as k grows large.

Power Laws

- But when people actually measured the distribution of links on the web, they found something very different.

  The fraction of pages with k in-links turned out to be approximately proportional to $1/k^2$!

- Why is this so different from the normal distribution?

- *$1/k^2$* decreases much more slowly as k increases, so pages with a very large number of in-links are much more common than we would expect with a normal distribution.
  - ○ 1/(10 x 10) = 0.01
  - ○ 2(-10)   = 0.0009

Power Laws

- Popularity seems to exhibit extreme imbalances, with very large values likely to arise
  - On the web, there are a large number of very popular pages
  - The fraction of telephone numbers that receive k calls per day is roughly proportional to $1/k^2$
  - the fraction of books that are bought by k people is roughly proportional to $1/k^3$
  - the fraction of scientific papers that receive k citations in total is roughly proportional to $1/k^3$

- As the normal distribution is widespread in a family of settings in the natural sciences, power laws seem to dominate in cases where the quantity being measured can be viewed as a type of popularity.

Power Laws

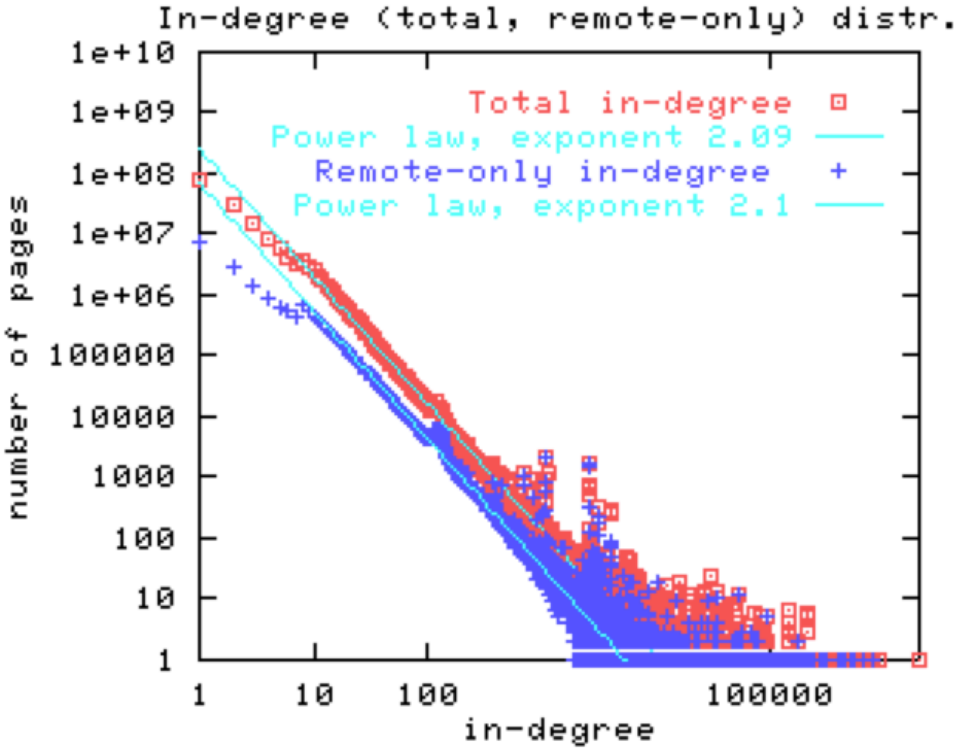
- So, if someone gives you a table showing the number of monthly downloads for each song at a large on-line music site that they're hosting,
  - test whether it's approximately a power law $1/k^c$ for some c, and if so, to estimate the exponent c.

- Let f(k) be the fraction of items that have value k, and suppose you want to know whether the equation $f(k) = a/k^c$, approximately holds

$$f(k) = a/k^c$$

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

Power Laws



A power law distribution (such as this one for the number of Web page in-links, from Broder et al. [2]) shows up as a straight line on a log-log plot.

Power Laws

- if power laws are so widespread, we need a simple explanation
- ..just as the Central Limit Theorem gave us a very basic reason to expect the normal distribution, we'd like something comparable for power laws.
- Just as normal distributions arise from many independent random decisions averaging out, …
- ..we will find that power laws arise from the feedback introduced by correlated decisions across a population.

Rich-get-richer Models

- An open and very interesting research question is to provide a fully satisfactory model of power laws

- We will assume that people have a tendency to copy the decisions of people who act before them.

- (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 according to the following probabilistic rule (which is 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 this page i.

Rich-get-richer Models

- (2) When page j is created, it produces a link to an earlier Web page according to the following probabilistic rule (which is 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 this page i.

  (b) With probability 1 – p, page j instead chooses a page i uniformly at random from among all earlier pages, and creates a link to the page that i points to.

  (c) This describes the creation of a single link from page j; one can repeat this process to create multiple independently generated links from page j.

- Part (2b) of this process is the key: author of page j copies the decision made by the author of page i.

Rich-get-richer Models

- The main result about this model is that if we run it for many pages, the fraction of pages with k in-links will be distributed approximately according to a power law $1/k^c$.

- The value of the exponent c depends on the choice of p [3].

- As p gets smaller, so that copying becomes more frequent, the exponent c gets smaller as well, making one more likely to see extremely popular pages.

- The copying mechanism in (2b) is really an implementation of the following "rich-get-richer" dynamics:
  - When you copy the decision of a random earlier page, the probability that you end up linking to some page y is directly proportional to the total number of pages that currently link to y.

Rich-get-richer Models

- So, we can rewrite step 2(b) as:

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

- This is called a "rich-getting-richer" rule because…

- .. the probability that page y experiences an increase in popularity is directly proportional to y's current popularity.

- This phenomenon is also known as preferential attachment[4], in the sense that links are formed "preferentially" to pages that already have high popularity.

Rich-get-richer Models

- So, the copying model provides as operational story for why popularity should exhibit such rich-get-richer dynamics.
- The more well-known someone is, the more likely you are to hear their name come up in conversation, and hence the more likely you are to end up knowing about them as well.
- A page's popularity grows at a rate proportional to its current value, and hence exponentially with time.
- A page that gets a small lead over others will therefore tend to extend this lead; the rich-get-richer nature of copying actually amplifies the effects of large values, making them even larger.
- So, after all, maybe power laws are not as surprising as they may have initially appeared!

Rich-get-richer Models

- Rich-get-richer models suggest a basis for power laws in a wide array of settings, including some that have nothing at all to do with human decision-making.
- The populations of cities have been observed to follow a power law distribution: the fraction of cities with population k is roughly $1/k^c$ for some constant c [5].
- The number of copies of a gene in a genome approximately follows a power-law distribution [6].
- These are still simple models designed just to approximate what's going on; there are other classes of simple models designed to capture power-law behaviour that we have not discussed here.

Rich-get-richer Models

- What all these simple models suggest is that when one sees a power law in data, the possible reasons why it's there can often be more important than the simple fact that it's there.

Rich-get-richer Models

- Once any one of these items (book, song) is well-established, the rich-get-richer dynamics of popularity are likely to push it even higher…

- …but getting this rich-get-richer process ignited in the first place seems like a precarious process, full of potential accidents and near-misses.

- The dynamics of popularity suggest that random effects early in the process should play a role – suppose we could roll time back 15 years, and then run history forward again, would the Harry Potter books again sell hundreds of millions of copies, or would they languish in obscurity while some other works of children's fiction achieved major success?

Rich-get-richer Models

- More generally, if history were to be replayed multiple times, it seems likely that there would be a power-law distribution of popularity each of these times, but it’s far from clear that the most popular items would always be the same.
- Salgankik, Dodds, and Watts performed an experiment [7].
- They created a music download site, populated with 48 obscure songs of varying quality written by actual performing groups.
- Visitors to the site were presented with a list of the songs and given the opportunity to listen to them.

Rich-get-richer Models

- Each visitor was also shown a table listing the current “download count” for each song — the number of times it had been downloaded from the site thus far.
- At the end of a session, the visitor was given the opportunity to download copies of the songs that he or she liked.
- What the visitors did not know is that they were actually being assigned at random to one of eight “parallel” copies of the site.
- The parallel copies started out identically, with the same songs and with each song having a download count of zero.
- *What do you think happened?*

Rich-get-richer Models

- Each parallel copy then evolved differently as users arrived.

- This experiment provided a way to observe what happens to the popularities of 48 songs when you get to run history forward eight different times.

- The “market share” of the different songs varied considerably across the different parallel copies.

- Although the best songs never ended up at the bottom and the worst songs never ended up at the top.

- Salganik et al. also used this approach to show that, overall, feedback produced greater inequality in outcomes.

Rich-get-richer Models

- They assigned some users to a ninth version of the site in which no feedback about download counts was provided at all.
- There was no direct opportunity for users to contribute to rich-get-richer dynamics.
- There was significantly less variation in the market share of different songs.
- Clear implications for popularity in less controlled environments
- The future success of a book, movie, celebrity, or Web site is strongly influenced by these types of feedback effects, and hence may to some extent be inherently unpredictable.

The Long Tail

- The distribution of popularity can have important business consequences.
- Consider a media company selling books and music.
- Are most sales being generated by:
  a small set of items that are enormously popular, or by
  a much larger population of items that are each individually less popular?
- In the former case, the company is basing its success on selling "hits", In the latter case, the company is basing its success on a multitude of "niche products".

The Long Tail

- In a 2004 article, “The Long Tail” [8], Chris Anderson argued that
  - Internet-based distribution and other factors were
  - driving the media and entertainment industries toward a world
  - in which the latter alternative would be dominant, with a “long tail” of obscure products driving the bulk of audience interest.

- This tension between hits and niche products makes for a compelling organizing framework.

- This is the fundamental models for companies like Amazon or Netflix,
  - where the ability to carry huge inventories,
  - makes it feasible to sell an astronomical diversity of products
  - even when very few of them generate much volume on their own.

- And ultimately, quantifying the importance of the Long Tail comes down to an analysis of power laws…

So, where are we?

- Initially, we asked about Popularity.
- We started from a baseline in which we expected to see Gaussian distributions and tight concentration around the average.
- We observed that the number of highly popular items was much higher than this baseline would suggest.
- The observation of the power laws brought us to the "rich-getting-richer" models.
- The distribution of popular items can have important business consequences, and that brought us here

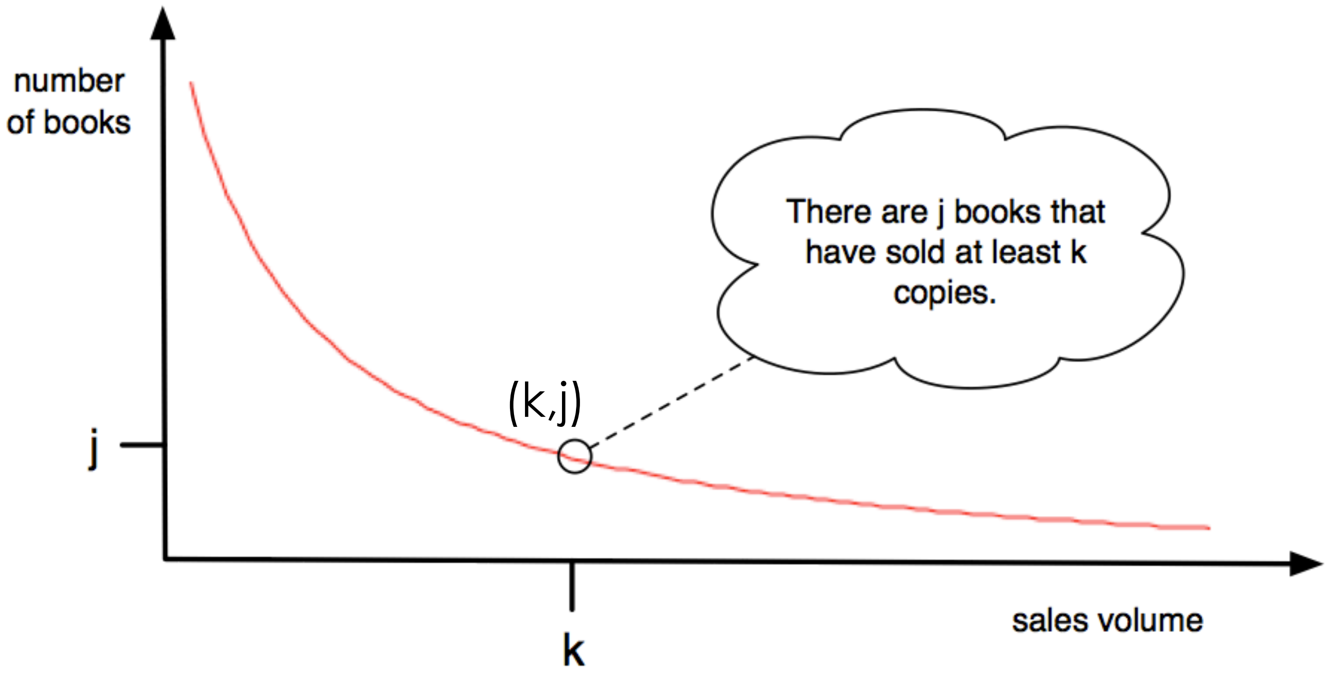
So, where are we?

- Now, we are looking at "long tails" and asking the opposite of the question that we started with.
- A sort of stereotype of the media business in which only blockbusters matter,…
- We're observing that the total sales volume of unpopular items, taken together, is really very significant.
- The observation of the power laws brought us to the "rich-getting-richer" models.

The Long Tail

- We were asking:
  - As a function of k, what fraction of items have popularity exactly k?
- Let us instead ask:
  - As a function of k what number of items have popularity at least k?
- If the original function was a power-law, then this new one is too.

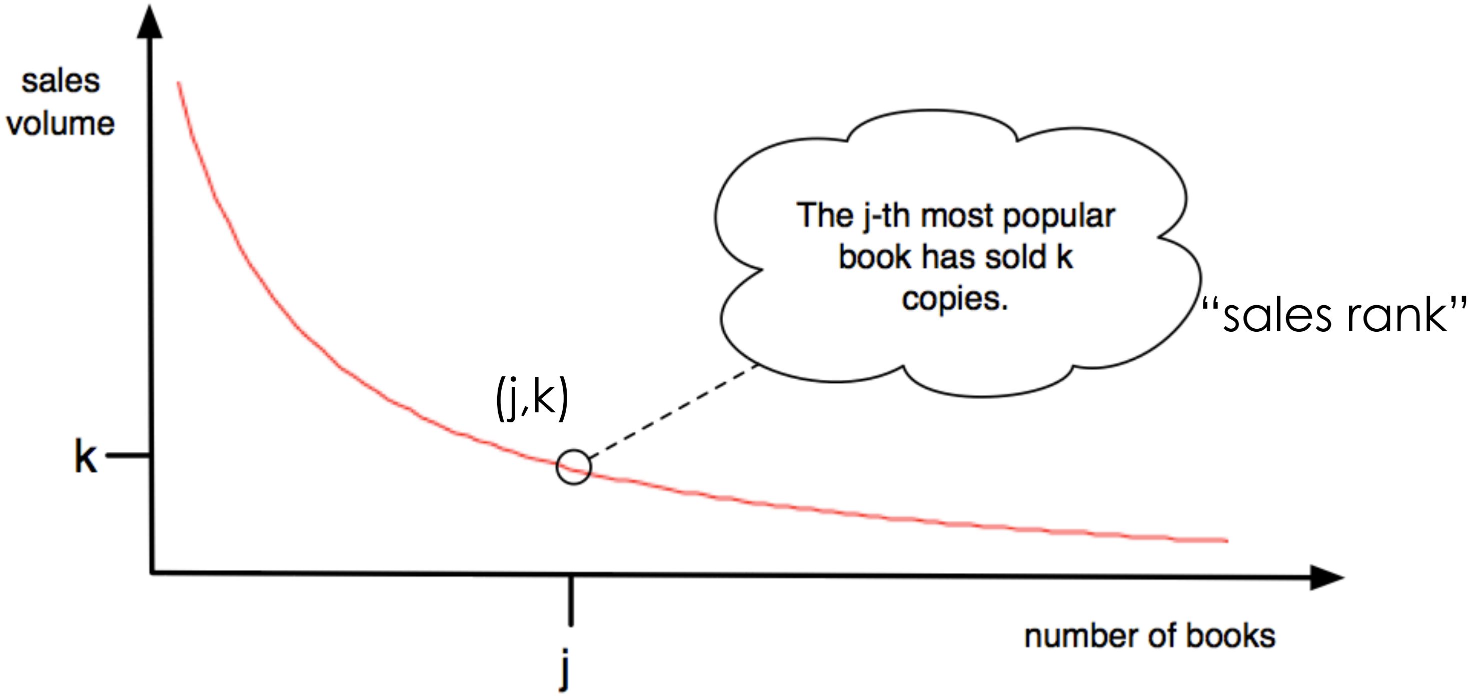
The Long Tail



The distribution of popularity: how many items have sold at least k copies?

As we follow the x-axis of the curve to the right, we're essentially asking, "As you look at larger and larger sales volumes, how few books do you find?"

The Long Tail



The distribution of popularity: how many copies of the jth most popular item have been sold?

"As you look at less and less popular items, what sales volumes do you see?"

The Long Tail

- Now, we can easily discuss trends in sales volume,

- Essentially, the area under the curve from some point j outward is the total volume of sales generated by all items of sales-rank j and higher;

- So a concrete version of the hits-vs.-niche question, for a particular set of products, is

  whether there is significantly more area under the left part of this curve (hits) or the right (niche products).

- The debate over trends toward niche products becomes a question of whether this curve is changing shape over time, adding more area under the right at the expense of the left.

The Long Tail

- The curves of this type, where the axes are ordered so that the variable on the x-axis is rank rather than popularity, are called Zipf plots.

- The linguist George Kingsley Zipf, who produced such curves for a number of human activities [10].

- He identified the empirical principle known as Zipf's Law, that the frequency of the j^{th} most common word in English (or most other widespread human languages) is proportional to 1/j.

Effect of Search Tools & Recommendation Systems

- A further question that has been growing in importance as people consider popularity and its distribution:
- *Are Internet search tools making the rich-get-richer dynamics of popularity more extreme or less extreme?*
- On one side of this question, we've seen that a model in which people copy links from uniformly random Web pages already gives an advantage to popular pages.
- Search engines such as Google are using popularity measures to rank Web pages, and the highly-ranked pages are in turn the main ones that users see in order to formulate their own decisions about linking.

Effect of Search Tools & Recommendation Systems

- A further question that has been growing in importance as people consider popularity and its distribution:
- *Are Internet search tools making the rich-get-richer dynamics of popularity more extreme or less extreme?*

- On one side of this question, we've seen that a model in which people copy links from uniformly random Web pages already gives an advantage to popular pages.

- Search engines such as Google are using popularity measures to rank Web pages, and the highly-ranked pages are in turn the main ones that users see in order to formulate their own decisions about linking.

- A similar argument can be made for other media in which a handful of the most popular items have the potential to crowd out all others.

Effect of Search Tools & Recommendation Systems

- In simple models, this kind of feedback can accentuate rich-get-richer dynamics, producing even more inequality in popularity [11].

- There are other forces at work.

- Users type a very wide range of queries into Google,
  - by getting results on relatively obscure queries,
  - users are being led to pages that they are likely never to have discovered through browsing alone.
  - enabling people to find unpopular items more easily
  - and potentially counteracting the rich-get-richer dynamics.

- In order to make money from a giant inventory of niche products, a company crucially needs for its customers to be aware of these products, and to have some reasonable way to explore them [8].

Effect of Search Tools & Recommendation Systems

- Viewed in this light, the types of recommendation systems that companies like Amazon and Netflix have popularised can be seen as integral to their business strategies:

- They are essentially search tools designed to expose people to items that may not be generally popular, but which match user interests as inferred from their history of past purchases.

- Ultimately, the design of search tools is an example of a kind of higher-order feedback effect:
  - by causing people to process their available options in one way or another, we can
  - reduce rich-get-richer effects,
  - or amplify them,
  - or potentially steer them in different directions altogether.

References

1. https://www.cs.cornell.edu/home/kleinber/networks-book/
2. 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.
3. Bela Bollobas and Oliver Riordan. Mathematical results on scale-free random graphs. In Stefan Bornholdt and Hans Georg Schuster, editors, Handbook of Graphs and Networks, pages 1–34. John Wiley & Sons, 2005.
4. Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. Science, 286:509–512, 1999.
5. Herbert Simon. On a class of skew distribution functions. Biometrika, 42:425–440, 1955.
6. Stanislaw Cebrat, Jan P. Radomski, and Dietrich Stauffer. Genetic paralog analysis and simulations. In International Conference on Computational Science, pages 709– 717, 2004.
7. Matthew Salganik, Peter Dodds, and Duncan Watts. Experimental study of inequality and unpredictability in an artificial cultural market. Science, 311:854–856, 2006.
8. Chris Anderson. The long tail. Wired, October 2004.

References

9. Lada Adamic. Zipf, power-laws, and Pareto: A ranking tutorial, 2000. On-line at http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html.
10. George Kingsley Zipf. Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology. Addison Wesley, 1949.
11. Soumen Chakrabarti, Alan M. Frieze, and Juan Vera. The influence of search engines on preferential attachment. In Proc. 16th ACM-SIAM Symposium on Discrete Algorithms, pages 293–300, 2005.