Power Laws and Rich-Get-Richer

CSC200 Lecture 31

February 10, 2016

Allan Borodin and Amirali Salehi-Abari

CSC200: Lecture 31

- Today:
  - Popularity, Rich-get-richer, Power Laws Ch.18.1-18.4
- Announcements
  - Term test 2: This Friday, Feb.12 (in this room) ; usual rules about one sheet of notes plus bring one empty sheet for possible calculations.
  - Over the reading week, I am planning to post the first set of questions for our Assignment 4, the last assignment
- Acknowledgement:
  - Some slides' materials are borrowed form the slides of the last offering of this course. Thanks to Professor Boutilier!

How Do Products Become Popular?

- A high-level look back: two reasons that products, services, information, etc. can become "popular" or widely used.
- **Information cascades (Ch.16):**
  - Choices made by others are informative for an agent's decision making process.
  - When a person *X* observes a person *Y* using the product, it *conveys information* about the quality of the product.
  - *X* combines this information with her own private information to make a rational decision (e.g. using Bayes Rule to determine the most likely probability).
- **Positive Externalities, Direct-Benefit or "Network Effects" (Ch.17):**
  - A person's utility depend on what other people do.
  - The more people use a product, the more benefit *X* derives from the product.
  - Each user *X* makes a rational decision whether to consume the product (e.g., adopt an OS, join clubs or a social network, read certain newspapers, etc.).

Popularity? (1)

- Popularity:
  - "A social phenomenon that dictates who or what is best liked"(Wikipedia)
  - "State of being liked, enjoyed, accepted, or done by a large number of people" (Merriam Webster Dictionary)
  - Most bought/read books: Harry Potter Series
  - Most watched Movie: Titanic
  - Most cited paper:
    - Protein measurement with the folin phenol reagent (Lowry et. al, 1951)
    - Number of Citation: **305,148** (data extracted on Oct. 2014)
    - Lowry's comments: "Although I really know it is not a great paper … I secretly get a kick out of the response"
    - Source: http://www.nature.com/news/the-top-100-papers-1.16224

Popularity? (2)

The questions that we try to address in next two lectures:

- How can we quantify popularity and imbalances that it cause?
- How is popularity distributed? And why does such a distribution arise?
- Do items, people, etc. become popular because of some intrinsic value or network process or to what extent is it based on chance?

Case Study: Web Graph

- We study popularity of the web pages but the idea here are applicable to any other contexts (e.g., social networks, movies, etc.)
- The popularity of a web page: the number of in-links.
  - Higher number of in-links, greater popularity
- Our popularity question:
  - How popularity is distributed over the set of webpages?
  - Or what fraction of web pages have k in-links?
- What is your guess?
  - A natural guess is Normal Distribution.

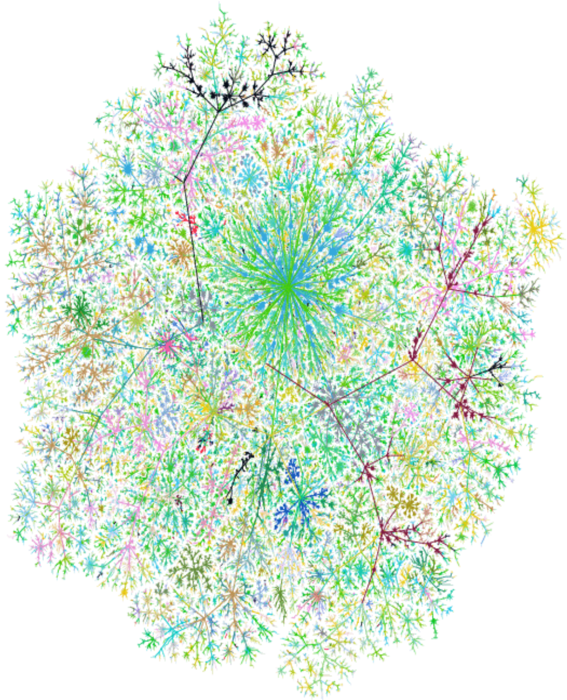


Image is taken from http://cheswick.com/ches/map/gallery/index.html

Normal Distribution

- Normal (or Gaussian) distribution (bell curve).
    - Ubiquitous in Nature.
- Characterized by mean μ and standard deviation σ

▪ Probability of seeing a specific sample average decreases *exponentially* with distance from mean μ.

▪ very large, or very small numbers are extremely unlikely.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty.$$

STANDARD DEVIATION OF THE MEAN

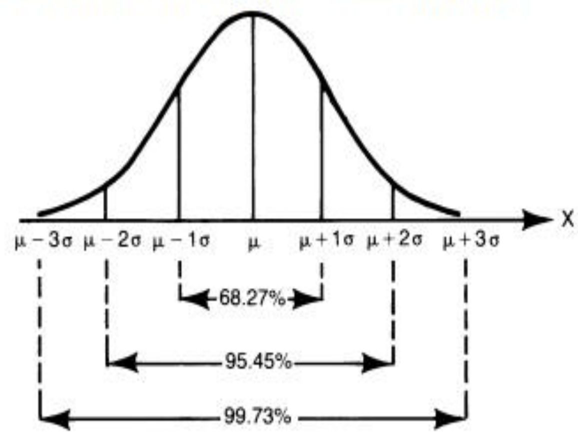


Figure 2

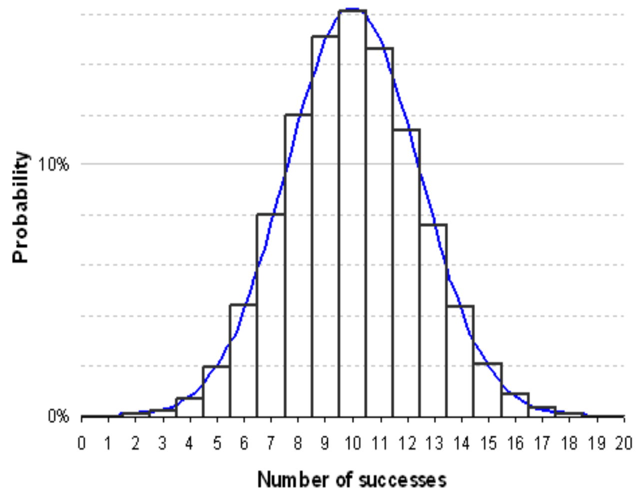
Percent	99.73%	99%	95.45%	95%	90%	80%	68.27%
No. of ± σ's	3.00	2.58	2.00	1.96	1.645	1.28	1.00

From: http://www.answers.com/topic/normal-distribution

Central Limit Theorem

- Central Limit Theorem explain how normal distribution arises:
  - Given a set of independent random variables. Their mean (or sum) will be approximately normally distributed.

- For instance, suppose each user from some population of N =25 users buys a certain book with probability $p=0.4$ *independent of what others do.*
- Then the *expected* number of books sold will be pN = 10
- The *distribution* of books sold will be approximately normal with mean pN =10 and variance p(1-p)N=6.
- What are the independent random variables in this example?



from http://www.quantdec.com/envstats/notes/class_06/properties.htm

Normal approximation to binomial distribution B(N=25,p=0.4); see here if not familiar with binomial distribution.

Back to our Web Graph Case Study

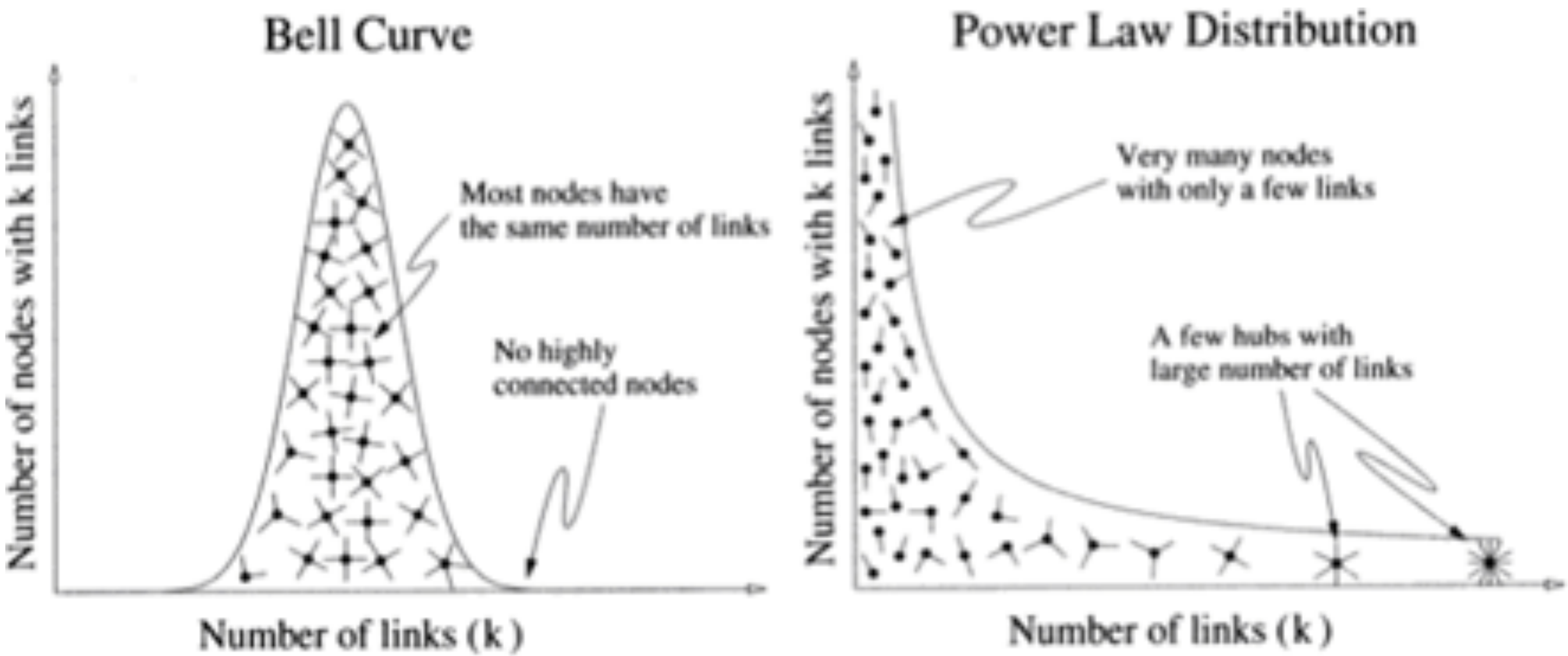
- What is the implication of Central Limit Theorem for webpages (or social networks)?
- If web pages (or people) decide independently at random on whether or not to link to any other web page, then what can you say about degree distribution?
- Based on Central Limit Theorem, as the number of in-links is the sum of many independent random quantities, the in-degree should be approximately normally distributed.
  - So, the number of pages with k in-links should decrease exponentially as k grows large.
  - Also, very large or small numbers of links are extremely unlikely.
- Does this happen in real-world?

  No.

Power Law Distribution.

- By crawling large sets of web pages and measuring the in-degree distributions, the recurring finding is that the fraction of web pages with *k* in-links is approximately proportional to k^{-2} (not to normal distribution).
  - k^{-2} decreases (with k) much slower than normal dist. does.
  - Small or large in-links values are more likely to occur compared to normal distribution. (We mentioned "heavy-tailed distributions" in Lecture 23.)
- A function that decreases proportionately with *k* to some fixed power is called a *power law*
  - e.g., number of web pages with *k* in-links: $f(k) = \alpha k^{-2} = \alpha\ 1/k^2$
  - α is a normalizing constant (varies with total number of pages, links)
- Power laws occur in:
  - Distribution of wealth follows a power law (Pareto distribution)
  - The fraction of books bought by *k* people $\propto k^{-3}$
  - The fraction of phones receiving *k* calls per day $\propto k^{-2}$
  - Citations to scientific articles, roughly $f(k) = \alpha/k^3$
  - Zipf observed that English word usage (in say a novel) follows a power law.
  - City sizes follow a power law.

Power Law vs. Normal Distribution

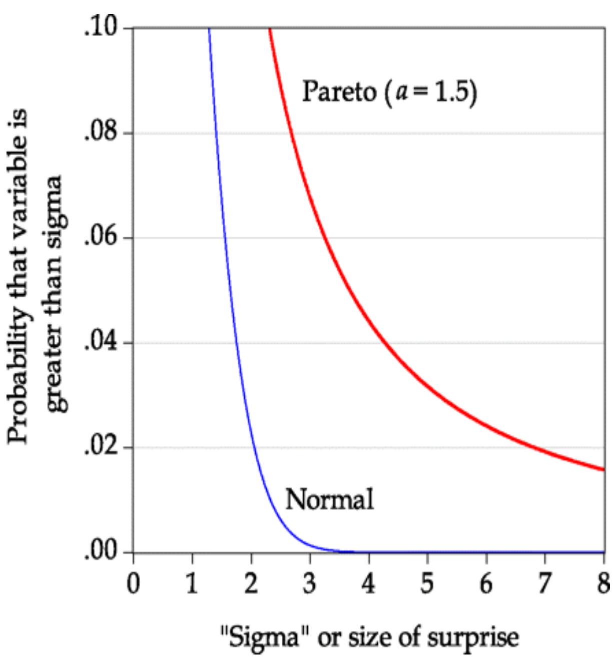


From "Linked: The New Science of Networks"

The “hub “terminology” here is inconsistent with E&K definition.

Power Laws are Scale-free

- The ratio of *f(k)* to *f(k')* depends only on the ratio *k/k'* not on their magnitude or "scale"
  - $f(k)/f(k') = \alpha k^{-c}/\alpha k'^{-c} = (k/k')^{-c}$
  - $f(2k)/f(2k') = \alpha 2k^{-c}/\alpha 2k'^{-c} = (k/k')^{-c}$
  - If you "slide along" on distribution, "relative picture" stays the same
  - Scale-free: the unit of measurement does not matter.
- They also have *long (or fat) tails*
  - significant numbers of events occur with large values of *k*
  - due to scale-free property: the relative reduction from *f(5)* to *f(10)* is same as from *f(50)* to *f(100)*, *f(1000)* to *f(2000)*, etc.: very stretched out!
  - compare Pareto distribution to normal distribution as *k* grows



Example 1: Links to Web Pages

- f(k): fraction of web pages with *k* in-links.
- f(k) = $\alpha\, k^{-2}$
  - So $\frac{f(1)}{f(2)} = 2^2 = 4$ times as many pages with 1 in-link as 2.
  - So $\frac{f(2)}{f(3)} = \frac{2^{-2}}{3^{-2}} = \frac{9}{4}$ times as many pages with 2 links as 3.
  - So $\frac{f(3)}{f(4)} = \frac{3^{-2}}{4^{-2}} = \frac{16}{9}$ times as many pages with 3 links as 4.
  - … 1.21 times as many pages with 10 links as 11 (ratio is 121/100)
  - … 1.02 times as many pages with 100 links as 101 (ratio is $101^2/100^2$)
- Notice that the *relative decrease* in number of pages with 1 additional links slows down very quickly, leaving a reasonable proportion of pages with large numbers of links
- BTW, can you quickly determine $\frac{f(6666666666)}{f(8888888888)}$ = ?

Example 2: What does $1/k^2$ Look Like?

- Suppose 1000 pages link to each other.
- Limit ourselves to maximum of *10 in-links* per page (for simplicity).
  - In-link distribution: $f(k) = \alpha\, k^{-2}$.
  - Table shows approx. number of pages with 1, 2, … 10 in-links
- Math is simple:
  - We know $\sum_{k=1}^{10} f(k) = 1$, so $\alpha \approx 0.645$
  - So, number of pages with
    - 1 in-link = $f(1) * 1000 \approx 645$
    - 2 in-link = $f(2) * 1000 = 0.645 * 2^{-2} * 1000 \approx 161$
    - …
- What is the number of edges?
  - 645*1+161*2+…+10* 6 = 1884
  - Edge density p = 1884/(1000*999) $\approx$ 0.0019

k (# of in links)	$1/k^2$	how many pages
1	1	645
2	1/4	161
3	1/9	72
4	1/16	40
5	1/25	26
6	1/36	18
7	1/49	13
8	1/64	10
9	1/81	8
10	1/100	6

Example 3: uniformly at random

- Table shows approx. number of pages with 1, 2, … 10 in-links for In-link distribution f(k) = $\alpha\, k^{-2}$
  - About 6 of 1000 pages have 10 in-links
- **Contrast:** suppose each page selects its out-link uniformly at random with p = 0.0019 (the same edge density in Example 2).
  - Each target page has p=19/10000 chance of being selected by a specific source page.
  - Chance of a specific page having 10 in-links is $= \binom{999}{10} p^{10}\ (1-p)^{989} \approx 2.44 * 10^{-5}$
  - Expected number of webpages with 10 in-links ≈ *0.0244.*
  - *Comparing this with 6 in power laws model, we see the power laws model have ≈ 6/.0244 ≈ 245 more times as many webpages with 10 in-links!!*

k (# of in links)	$1/k^2$	how many pages
1	1	645
2	1/4	161
3	1/9	72
4	1/16	40
5	1/25	26
6	1/36	18
7	1/49	13
8	1/64	10
9	1/81	8
10	1/100	6