Rich get Richer

Power Laws, Long Tails and Preferential Attachment Models in World Wide Web and Social Networks

Arindam Pal
arindam.pal1@tcs.com

TCS Innovation Labs Kolkata

June 21, 2013

Agenda

- Popularity and the Rich get Richer phenomena
- Power laws in social networks
- Preferential attachment model
- Emergence of long tails
- Effect of search engines and recommendation systems
- Analysis of the preferential attachment model
- Conclusion

Popularity

Here are some questions about popularity.

- Why do some people or things become more popular than others?
- Why do popular objects get even more popular?
- How can we quantify these imbalances?
- Why do they arise?
- Are they intrinsic to the notion of popularity?

We will try to answer some of these questions.

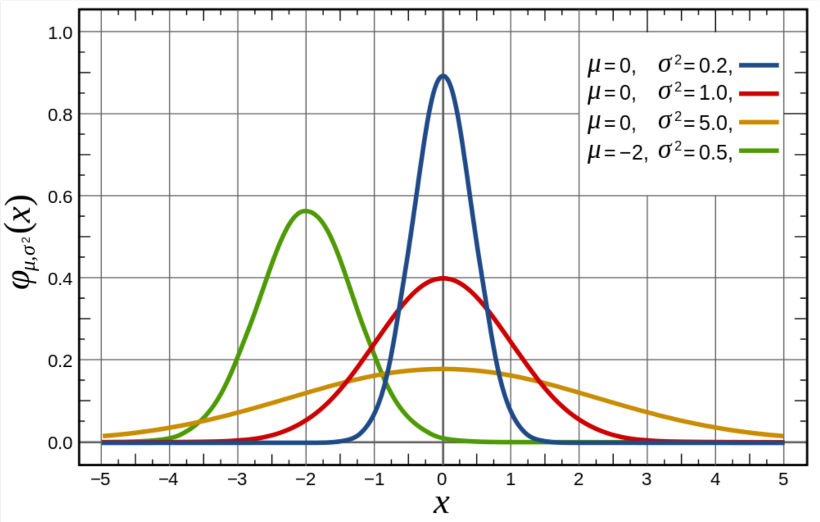
Popularity on the Web and social networks

- We can consider these networks as graphs, where there is a directed edge between two nodes whenever a page links to another page or an undirected edge when two users are friends.
- Counting the number of incoming edges is a measure of popularity.
- This is known as the *in-degree* of a node.
- As a function of k, what fraction of pages on the web has in-degree k?
- This is a measure of how popularity is distributed among web pages.
- This is called the *in-degree distribution* of a graph.
- What kind of probability distribution is this?

The Normal distribution

- The Normal (Gaussian) distribution is specified by two parameters – the mean (μ) and the standard deviation (σ) from the mean.
- The *probability density function* is given by $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.
- We write $X \sim \mathcal{N}(\mu, \sigma^2)$.
- Typically it is scaled (normalized) so that $\mu = 0$ and $\sigma = 1$.
- $\mathbf{Pr}[|X - \mu| \geq c\sigma] \leq e^{-\alpha c}$, for some $\alpha > 0$.
- The probability of observing a value that exceeds the mean by more than c times the standard deviation decreases exponentially with c.

The Normal curve



The Central Limit Theorem

- Let $X_1, \ldots, X_n$ be a sequence of independent and identically distributed random variables with $\mathbf{E}[X_i] = \mu$ and $\mathbf{Var}[X_i] = \sigma^2$.
- If
$$S_n = \frac{1}{n} \sum_{i=1}^{n} X_i,$$
- Then
$$\lim_{n \to \infty} S_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$
- In other words, in the limit the sum (or average) of any sequence of independent and identically distributed random variables is distributed according to the normal distribution.

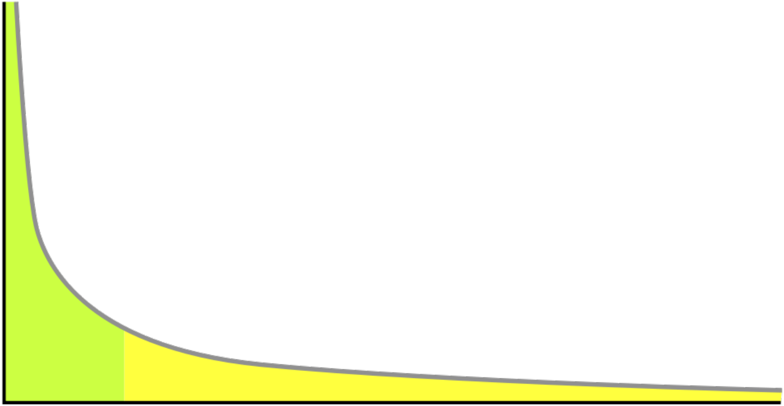
Predicted vertex degree distribution

- If we assume 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 and identically distributed random quantities.
- Hence, the number of in-links should be normally distributed.
- So, the number of pages with k in-links should decrease exponentially in k, as k grows large.
- Let X be the random variable denoting the in-degree of a page.
- $\mathbf{Pr}[X = k] = A \cdot e^{-\alpha k}$ for some constants A and α.

Actual vertex degree distribution

- It has been observed that the fraction of web pages having in-degree k is approximately proportional to $\frac{1}{k^2}$.
- $\mathbf{Pr}[X = k] = A \cdot k^{-c}$, for some constants A and c.
- So it is more likely to have pages with large in-degree than what is predicted by the normal distribution.
- These are also called *scale-free networks*.
- This is not unique for web pages. This also happens for telephone networks, friendship networks, citation networks and many other networks.

Power laws and long tails



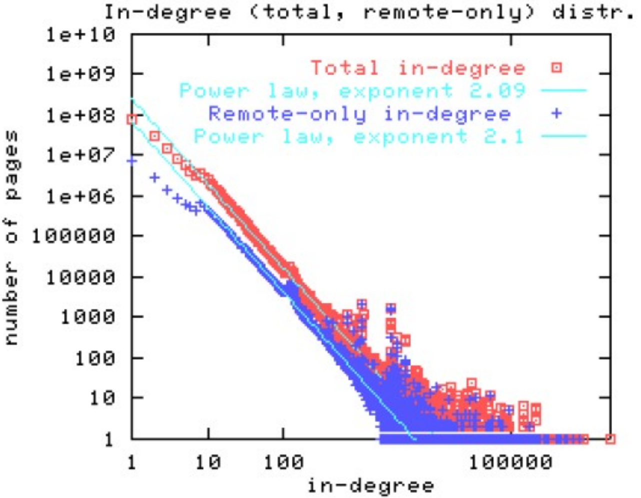
Some examples of power law distributions

- The fraction of web pages that are linked by k web pages is approximately proportional to $\frac{1}{k^2}$.
- The fraction of telephone numbers that receive k calls per day is approximately proportional to $\frac{1}{k^2}$.
- The fraction of books that are bought by k people is approximately proportional to $\frac{1}{k^3}$.
- The fraction of scientific papers that receive k citations is approximately proportional to $\frac{1}{k^3}$.

How to check if a distribution follows power law

- Let $P(k)$ be the fraction of items having value k.
- Suppose we want to test whether $P(k) = A \cdot k^{-c}$, for some constants A and c.
- Then, $\log P(k) = \log A - c \log k$.
- So, if we plot $\log P(k)$ as a function of $\log k$, we should get a straight line whose slope is $-c$ and whose intercept on the y-axis is $\log A$.
- A log-log plot provides a quick way to figure out if the data exhibits an approximate power law distribution.

Power law distribution plotted on a log-log scale



The Erdős-Rényi random graph model

- There are two ER models: the $\mathcal{G}(n, p)$ model and the $\mathcal{G}(n, m)$ model.
- In the $\mathcal{G}(n, p)$ model, there are n nodes.
- Each of the $\binom{n}{2}$ edges is included with probability p.
- The expected number of edges in a graph $G \in \mathcal{G}(n, p)$ is $\binom{n}{2}p$.
- Let $P(k)$ be the probability of a vertex having degree k.

$$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}.$$

$$\lim_{n \to \infty} P(k) = \frac{c^k e^{-c}}{k!}, \text{ if } np = c.$$

- Hence, the vertex degree distribution for an ER graph is binomial, which is Poisson for large n.

Problem with the Erdős-Rényi model

- It is a static model. There is no mechanism to allow vertex additions/deletions.
- The vertex degree distribution does not follow a power law distribution, even in the limit of large n.
- So where is the power law coming from?
- We need a new generative model to explain this behavior.

The preferential attachment model

- Here is a simple stochastic process for creation of links on web pages.
- Pages are created in the order $1, \ldots, N$.
- When page j is created, it links to an existing page using the following probabilistic rule:
  1. With probability p, page j chooses a page i *uniformly at random* from among all existing pages, and creates a link to this page i.
  2. With probability $1 - p$, page j chooses a page i *uniformly at random* from among all earlier pages, and creates a link to *the page that* i *points to*.
- This is known as the Barabási–Albert model.

An alternate formulation

- The probability of linking to some page ℓ is directly proportional to the total number of pages that currently link to ℓ.
- An alternate way to state rule (2) is:
  2a With probability $1 - p$, page j creates a link to a page ℓ with probability proportional to ℓ's current in-degree.
- Note that in rule (2), we are copying the decision made by another page, while in rule (2a), we are selecting a page based on its popularity, although the rules are equivalent.

A few comments

- This is called *rich get richer*, because the probability that the popularity of a page increases is directly proportional to it's current popularity.
- Links are formed *preferentially* to pages that already have high popularity.
- In this model, the probability of a page having in-degree k will be proportional to $\frac{1}{k^c}$, where the value of c depends on p.
- As p gets smaller, copying becomes more frequent. As a result c gets smaller, and we are more likely to see extremely popular pages.

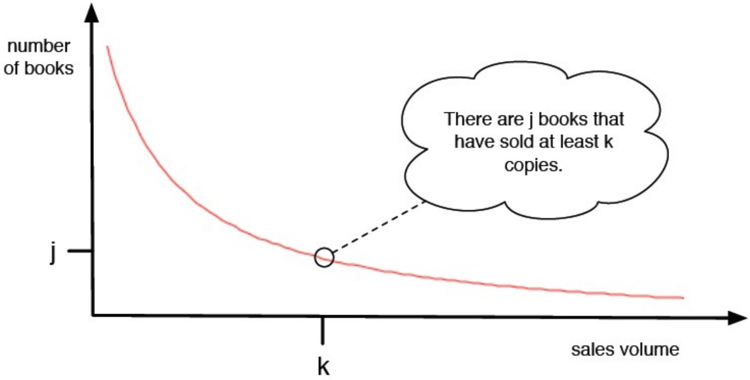
The Long Tail

- Consider a media company with a large inventory of books or music.
- The important question is: are most sales being generated by a small set of items that are very 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" – a small number of blockbusters that create huge revenues.
- In the latter case, the company is basing its success on a multitude of "niche products," each of which appeals to a small segment of the audience.

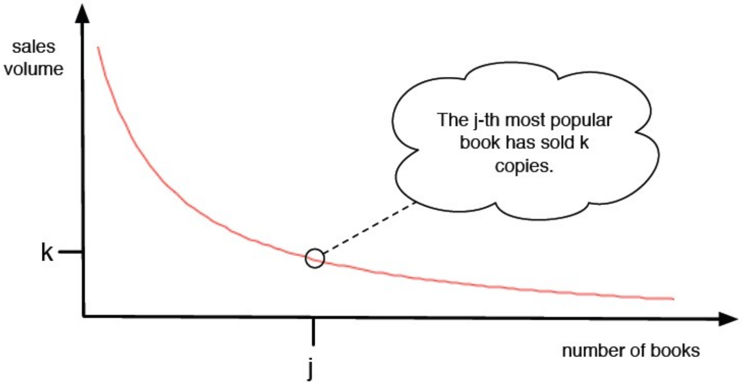
Properties of the Long Tail

- We are interested in the following question – As a function of k, how many items have popularity at least k?
- A point (k, j) on this curve means there are j books that have sold at least k copies.
- Now we want to ask the inverse question – As a function of j, how many copies of the j^{th} most popular item has been sold?
- A point (j, k) on this curve means k copies of the j^{th} most popular item has been sold.

Frequency distribution



Rank distribution



Long Tail and Zipf's law

- The area under the curve from some point j to the right is the total volume of sales generated by all items of sales rank j and higher.
- For a particular set of products, whether there is significantly more area under the left part of this curve (hits) or the right part (niche products)?
- It has been observed that there is significant probability mass under the right part, showing that items which are not so popular generate significant amount of sale.
- Curves of the type where the variable on the x-axis represents rank and y-axis represents frequency have a long history.
- Zipf's law says that the frequency of the j^{th} most common word in English is proportional to $\frac{1}{j}$, which is a power law.

Effect of search engines and recommendation systems

- Are search engines making the rich get richer dynamics of popularity more extreme or less extreme?
- On one hand, Google is using popularity measures to rank Web pages, and the highly-ranked pages are the ones that users see in order to formulate their own decisions about linking.
- On the other hand, by getting results on relatively obscure queries, users are finding pages that they are unlikely to have discovered through browsing alone.
- In order to make money from a giant inventory of niche products, customers should be able to find these products.
- Recommendation systems used by companies like Amazon and Netflix are search tools designed to expose people to items which match user interests as inferred from their history of past purchases.

- Pages are created in the order $1, \dots, N$.
- When page j is created, it links to an existing page using the following probabilistic rule:
  1. With probability p, page j chooses a page i *uniformly at random* from among all existing pages, and creates a link to this page i.
  2. With probability $1 - p$, page j creates a link to a page ℓ with probability proportional to ℓ's current in-degree.

The discrete process

- Let $X_j(t)$ be the in-degree of a node j at time $t \geq j$, for $1 \leq j \leq N$.
- **The initial condition**: Since node j starts with no in-links when it is first created at time j, we know that $X_j(j) = 0$.
- **The expected change to X_j at time $t+1$**: Node j gets an in-link at $t+1$ if the link from the newly created node $t+1$ points to it.
- With probability p, node $t+1$ creates a link to a node chosen uniformly at random among all existing nodes. The probability that j is this node is $\frac{1}{t}$.
- With probability $1-p$, node $t+1$ creates a link to node j with probability proportional to j's in-degree. Since the total number of nodes is t and in-degree of j is $X_j(t)$, this probability is $\frac{X_j(t)}{t}$.

The probabilistic recurrence relation for $X_j(t)$

- The recurrence relation for $X_j(t)$ is given by

$$\mathbf{E}[X_j(t+1) - X_j(t)] = \frac{p}{t} + \frac{(1-p)X_j(t)}{t},$$

$$\mathbf{E}[X_j(t+1)] = \mathbf{E}[X_j(t)] + \frac{p}{t} + \frac{(1-p)X_j(t)}{t}.$$

- Since it is complicated to solve this probabilistic recurrence, we will analyze a closely related but simpler process.
- The idea in formulating the simpler model is to make it deterministic.
- In this model there are no probabilities; instead, everything evolves in a fixed way over time.

The continuous process

- Time t runs continuously from 0 to N.
- We approximate $X_j(t)$ by a continuous function of time $x_j(t)$.
- **The initial condition**: Since $X_j(j) = 0$, we define $x_j(j) = 0$.
- **The rate of change of x_j at time t**:

$$\text{Since, } \mathbf{E}[X_j(t+1) - X_j(t)] = \frac{p}{t} + \frac{(1-p)X_j(t)}{t},$$

$$\text{We define, } \frac{dx_j}{dt} = \frac{p}{t} + \frac{(1-p)x_j}{t}.$$

- Rather than dealing with random variables $X_j(t)$ that move in small probabilistic jumps at discrete points in time, we work with a quantity $x_j(t)$ that changes smoothly over time, at a rate tuned to match the expected changes in the corresponding random variables.

Analyzing the continuous process

- Setting $q = 1 - p$ for conciseness we get,

$$\frac{dx_j}{dt} = \frac{p + qx_j}{t},$$
$$\int \frac{dx_j}{p + qx_j} = \int \frac{dt}{t}.$$

- Solving this differential equation along with the initial condition $x_j(j) = 0$, we get

$$x_j(t) = \frac{p}{q}\left[\left(\frac{t}{j}\right)^q - 1\right].$$

Power law from the deterministic approximation

- For a given value of k and a time t, what fraction of all nodes have at least k in-links at time t?
- Equivalently, for a given value of k and a time t, what fraction of all functions $x_j(t)$ satisfies $x_j(t) \geq k$?

$$\frac{p}{q}\left[\left(\frac{t}{j}\right)^{q} - 1\right] \geq k,$$

$$j \leq t\left(\frac{qk}{p} + 1\right)^{-\frac{1}{q}}.$$

- Out of all the functions $x_1, \ldots, x_t$ at time t, the fraction of values j that satisfy this is

$$\frac{1}{t} \cdot t\left(\frac{qk}{p} + 1\right)^{-\frac{1}{q}} = \left(\frac{qk}{p} + 1\right)^{-\frac{1}{q}}.$$

- Hence, the fraction of x_j that are at least k is proportional to $k^{-\frac{1}{q}}$.

From at least k to exactly k

- Suppose $f(x)$ is the *probability density function* of a continuous random variable X.
- Then, $\mathbf{Pr}[a \leq X \leq b] = \int_a^b f(x)dx$.
- Let $F(x)$ be the *cumulative distribution function* of X.
- We know that $F(x) = \mathbf{Pr}[X \leq x] = \int_{-\infty}^x f(t)dt$.
- Equivalently, $f(x) = F'(x) = \frac{dF}{dx}$.
- Since in our case we have, $G(k) = \mathbf{Pr}[X \geq k] = 1 - F(k)$, the required function is $f(k) = \frac{dF}{dk} = -\frac{dG}{dk}$.
- Note that since X is a continuous random variable, $f(k) = 0$. This is an approximation to the actual value of $\mathbf{Pr}[X = k]$.

Power law arising from the deterministic model

$$\text{Since, } G(k) = \left(\frac{qk}{p} + 1\right)^{-\frac{1}{q}},$$

$$\text{We have, } -\frac{dG}{dk} = \frac{1}{q} \cdot \frac{q}{p} \left(\frac{qk}{p} + 1\right)^{-\left(1+\frac{1}{q}\right)}$$

$$\text{Hence, } \mathbf{Pr}[X = k] = \frac{1}{p} \left(\frac{qk}{p} + 1\right)^{-\left(1+\frac{1}{q}\right)}.$$

- The deterministic model predicts that the fraction of nodes with k in-links is proportional to $k^{-\left(1+\frac{1}{q}\right)}$, which is a power law with exponent $c = 1 + \frac{1}{1-p}$.

Remarks

- Subsequent analysis of the original probabilistic model showed that, with high probability over the random formation of links, the fraction of nodes with k in-links is proportional to $k^{-\left(1+\frac{1}{1-p}\right)}$.
- The heuristic argument given by the deterministic approximation to the model provides a simple way to see where this power law exponent comes from.
- $\lim_{p\to 1} c = \infty$. Hence, link formation is mainly based on uniform random choices and the power law exponent tends to infinity.
- In this case, nodes with very large numbers of in-links become increasingly rare.
- $\lim_{p\to 0} c = 2$. Hence, the network is highly influenced by the copying behavior.
- The fact that 2 is a natural limit for the exponent also tallies with the fact that many power law exponents in real networks is close to 2.

Conclusion

- In this talk, we discussed about how popularity evolves in social networks.
- We talked about a common phenomenon called *rich get richer*.
- We saw how power law emerges and how the preferential attachment model can give a mathematical explanation of this.
- We also saw how long tails and search engines can affect the dynamics of sells for e-commmerce companies.
- New ideas and mathematical techniques are needed to analyze global effects observed in social networks.
- This includes results from random graphs, percolation theory, spectral graph theory and probabilistic methods.

Moral of the story

The rich get richer and the smart get smarter!

Questions?