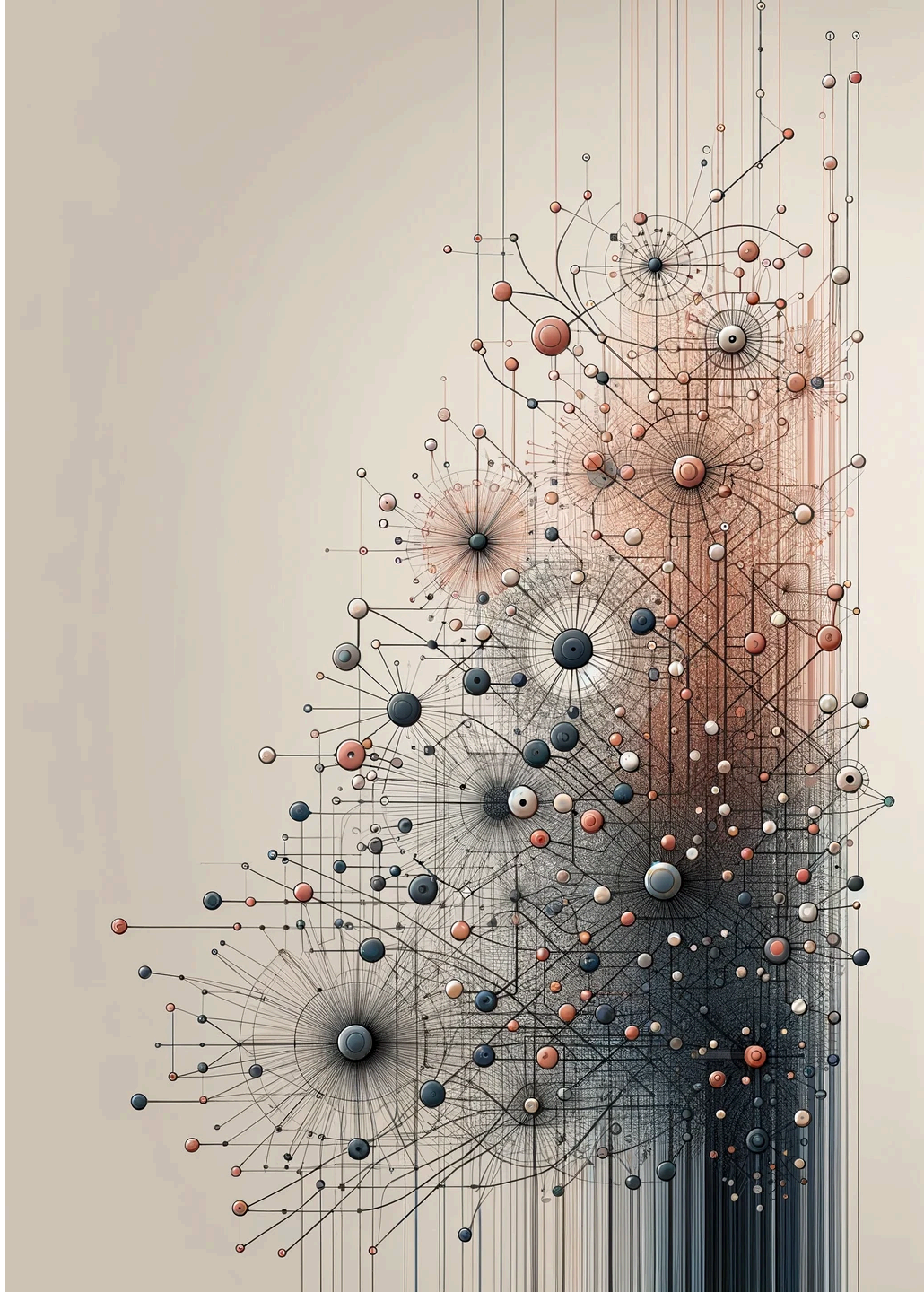
UNIVERSITÀ DI TORINO

Analisi e Visualizzazione delle Reti Complesse

NS10 - Power Laws and Rich-Get-Richer Phenomena

Prof. Rossano Schifanella



UNIVERSITÀ DI TORINO

Agenda

- Popularity as a Network Phenomenon
- Power Laws
- Rich-Get-Richer Models
- The Unpredictability of Rich-Get-Richer Effects
- The Long Tail
- The Effect of Search Tools and Recommendation Systems

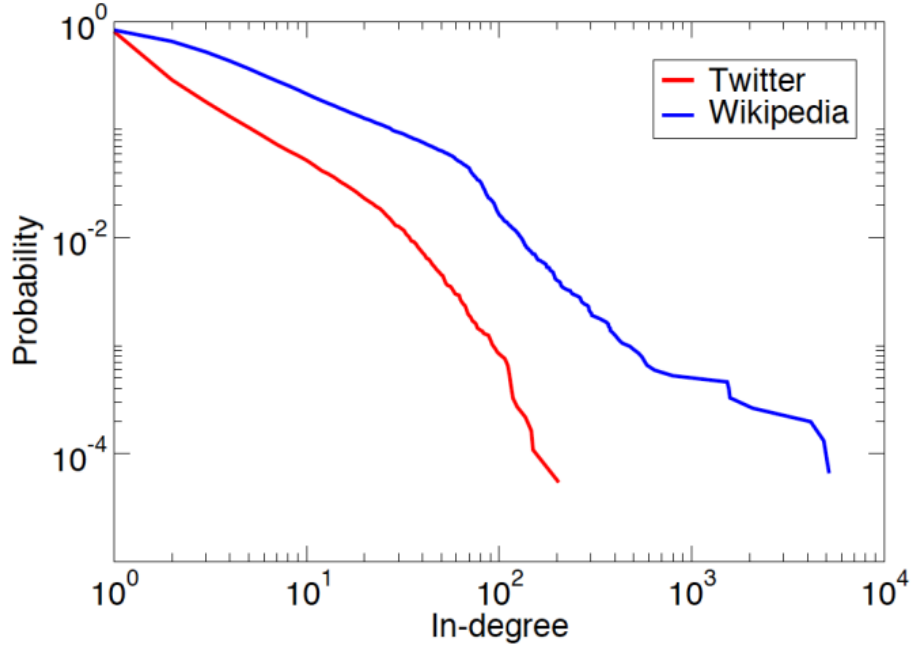
Popularity as a Network Phenomenon

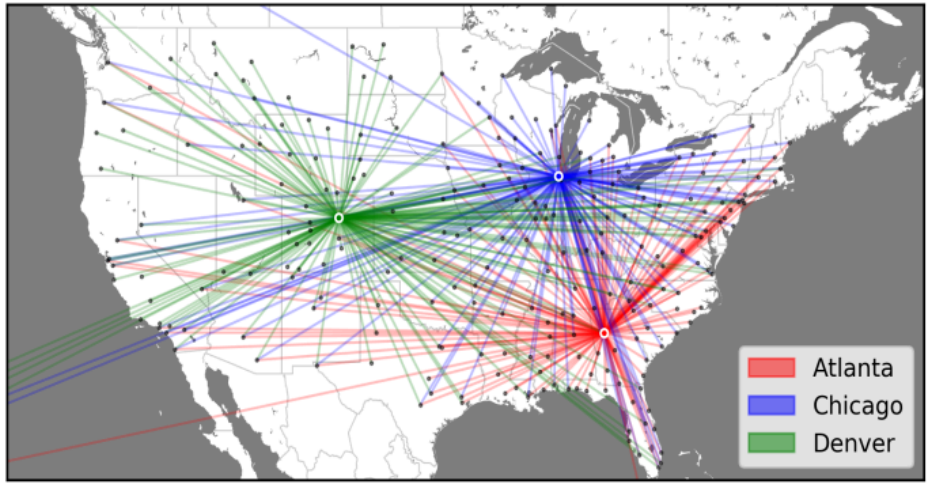
Popularity, heterogeneity and networks

Recap:

- Real networks are heterogeneous
- In-links as a measure of popularity
- The **heterogeneity** parameter as a **measure of distribution's broadness**

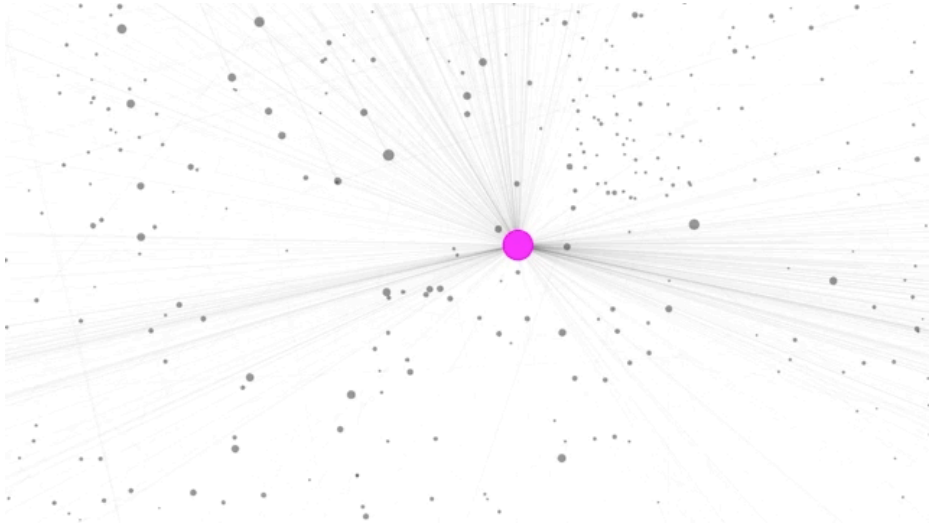
$$\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle^2}$$





Case study: the Web

- Characterizing popularity reveals imbalances (inequalities)
  - almost everyone is popular for very few people
  - very few people achieve high popularity
  - very, very few people achieve global popularity
- Why? Is this phenomenon intrinsic to the whole idea of popularity itself?
  - Have a look at the video for a dynamic view

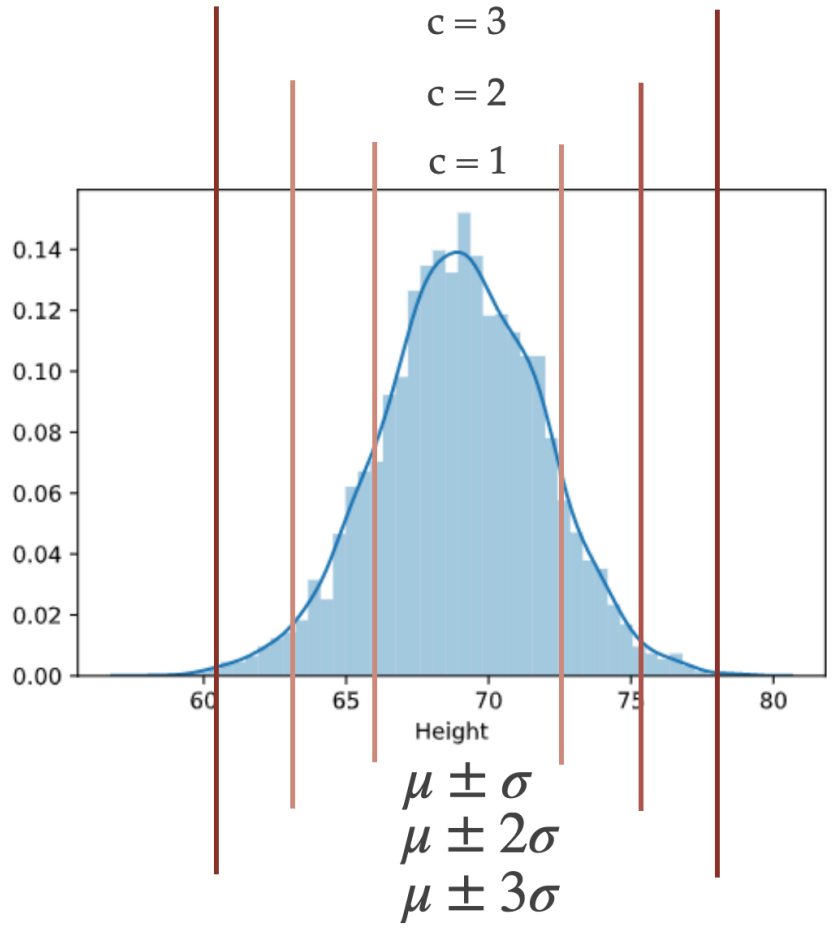


Looking for a popularity scale

- As a function of k, what fraction of (web) pages have k links?
- larger k corresponds to greater popularity
- First (and simple) hypothesis: **normal distribution**
- the mean defines a scale of the population: good for estimation/prediction

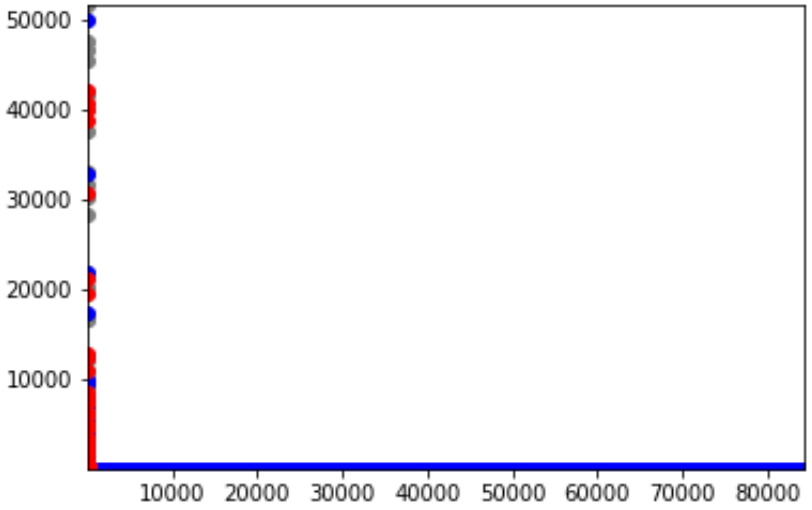
Example: Heights

- If we look at people's height distributions
- mean: = 69.03 (feet)
- std: = 2.86
- The probability of observing a value that exceeds the mean by more than c times the standard deviation decreases exponentially with c
- Amazingly high persons are very unlikely
- **Central Limit Theorem:** if we take any sequence of small **independent random quantities**, then in the limit their sum (or average) will be distributed according to the normal distribution.



Example: the Web (a sample)

- **Source dataset:**
  - Berkeley-Stanford web graph
  - Nodes: 68,5230
  - Edges: 7,600,595
- If we plot degree, in-degree, out-degree distribution, we find a different picture

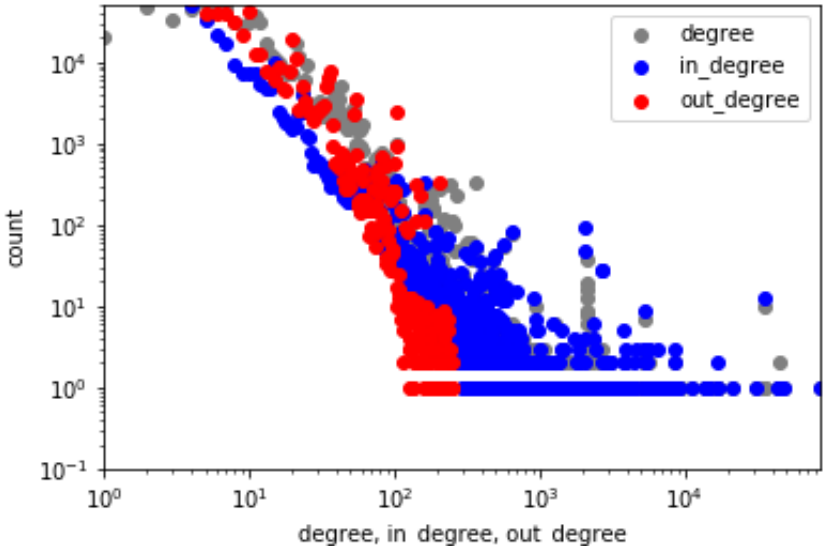


Example: the Web (a sample) in a log-log scale

- It turns out that the best way to plot heavy-tailed distributions is to use log-log scales

$$\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle^2} = \frac{81782.51}{22.18^2} = 166.18$$

- standard deviation $\sigma = \sqrt{\langle k^2 \rangle} = 285.98$
- the degree of a randomly chosen node is $22.18 - 285.98 \leq k_{in} \leq 22.18 + 285.98$
- not very informative



Power Laws

$$f(k) \propto \frac{1}{k^c} = k^{-c}$$

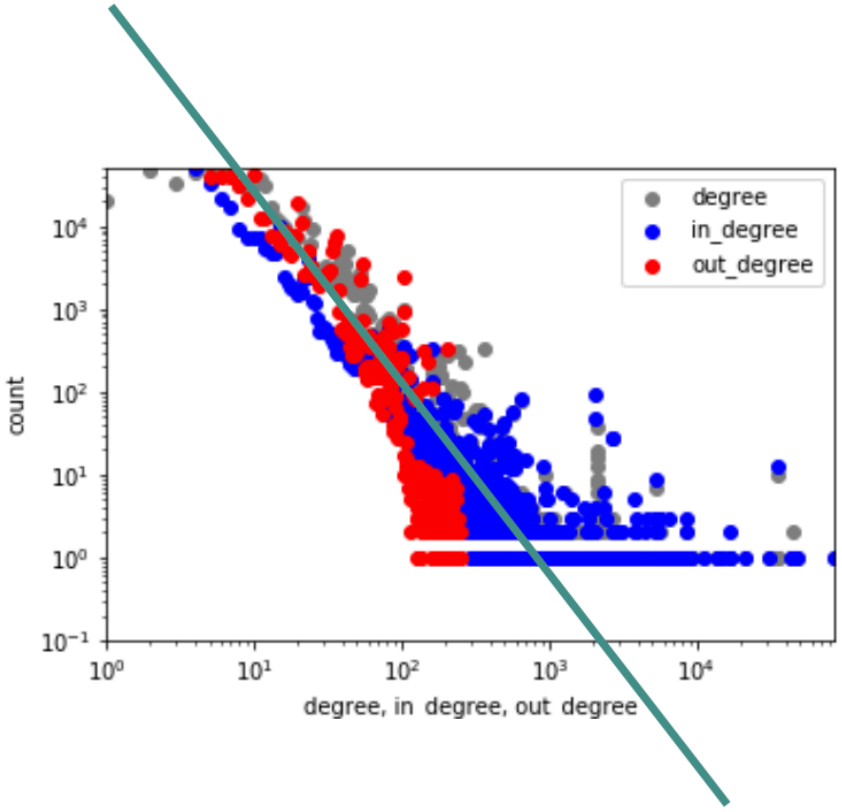
- It decreases much more slowly as k increases
- Emergence of hubs is likely
- It can be observed empirically in many domains
- E.g., in the Web the fraction of web pages that have k in-links follows a power law with $c = 2.1$
  - Pages with very large k are much more common than expected with the normal distribution
- In other networks $2 < c < 3$ very often
- In the $2 < c < 3$ regime we have that when $N \to \infty$ then $\langle k^2 \rangle \to \infty$

Fitting with a straight line

- Approximations of power laws are very common
- $f(k) = ak^{-c}$ for some constants a and c

$$log(f(k)) = log(ak^{-c})$$
$$log(f(k)) = log(a) - clog(k)$$
$$y = log(a) - cx$$

- In a log-log plot:
  - $log(a)$ is the **intercept**
  - $-c$ is the **slope**
- Refer to **this notebook in the course github** for a famous Python package to fit powerlaws



Why do hubs emerge?

- Let's accept that power laws represent many phenomena
- Why?
  - We are observing a kind of "order" emerging from chaos
  - Is there an underlying process that keeps the line so straight?
  - Like normal distributions arise from many independent random decisions averaging out, can we find something similar in this context?
    - we will find that power laws arise from the feedback introduced by correlated decisions across a population

UNIVERSITÀ DI TORINO

Rich-Get-Richer models

- **We assume that people tend to copy the decision of people who acted before them**
- Continuing with the Web scenario, we describe the network growth process as follows:

1. Pages are created in a sequence: $1, 2, \ldots, N$
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).
    i. 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.
    ii. With probability $1 - p$, page j chooses a page l uniformly at random from among all earlier pages, and **creates a link to the page that** i **points to**.
    iii. 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.

- To keep things simple, we will suppose that each page creates just one outbound link

UNIVERSITÀ DI TORINO

Relation to Preferential attachment

- The rich-get-richer model is a generalization of the preferential attachment model by Barabasi and Albert
- with probability $p \Rightarrow$ the selection of the end-point is random
- with probability $1-p \Rightarrow$ we can prove that rule 2.(ii) is equivalent to the following rule: "page j chooses a page with probability proportional to l's current number of in-links" $\Rightarrow$ that is preferential attachment
- This is a simple model: it does not explain everything, but it provides a natural explanation for the emergence of hubs
- Do not be surprised to observe skewed distributions in real data that resemble power law (usually from $k > k_{min}$ - other times exhibiting exponential cut-offs)

The Unpredictability of Rich-Get-Richer Effects

The fragility of popularity

- Power laws are produced by **feedback effects**
- The initial stages of the process that gives rise to popularity is a **relatively fragile state**
- Focusing on **cultural market**:
  - **Can we predict the popularity of a song, a movie, a book, etc.?**
- We can expect initial fluctuations
  - this brings **unpredictability**

Predicting hubs emergence

- We can predict that a **power law** and **hubs will emerge**
- **But which hubs?**
  - **Predicting the success of an individual item is not like predicting that some individual will have global success!**

The MusicLab experiment

[The MusicLab experiment paper] [pdf]

The MusicLab experiment

- **MusicLab**: a music download site, populated with 48 obscure songs of varying quality written by actual performing groups
- Visitors were presented with a list of original songs and given the opportunity to **listen** to them and to **download** them at the end of the session
- A table with the **download counts** was provided as **measure of popularity**
- Visitors randomly assigned to different 8 different sessions
- For every session's category a songs ranking was produced

UNIVERSITÀ DI TORINO

The MusicLab experiment

- **Rankings in all 8 sessions were considerably different!**
- **Very good songs did not end up at the bottom**
- **Very bad songs did not end up at the top**
- What about all the other songs?
  - They resulted in **very mixed positions**
- Observe that in some sessions, the order was established by means of popularity
- Social influence was found relevant at the end of the process
- But **initial fluctuations are unpredictable**

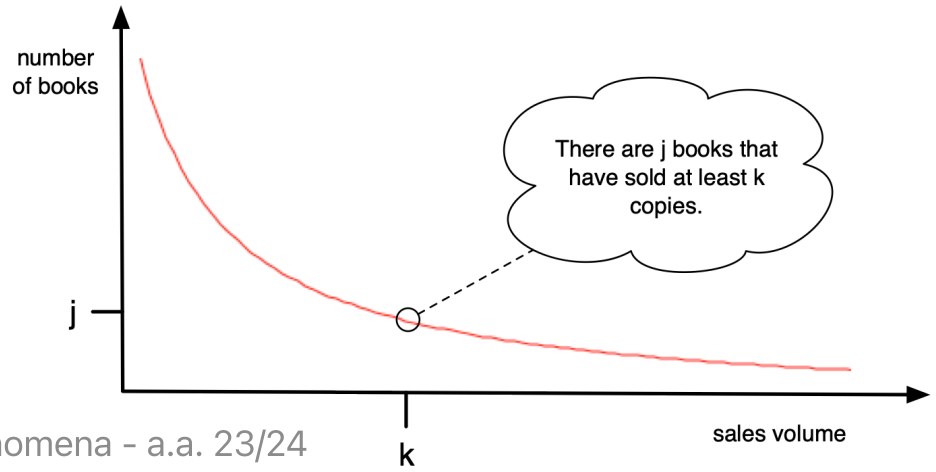
- Visitors in a nineth session of the experiment, When presented with the songs **without the download counts**, the final raking showed **significanlty less variability** in the distributon of popuylarity.

The Long Tail

- **Popularity can be characterized by power laws**
- That means that a small set of items is enormously popular
- In a media company for example, 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?
- If you could bet your money on "niches" or "hits", what would you do?
- Chris Anderson's idea:
  - **do not focus on hits, but try to estimate the market sales of all the niches**

Focus on hits

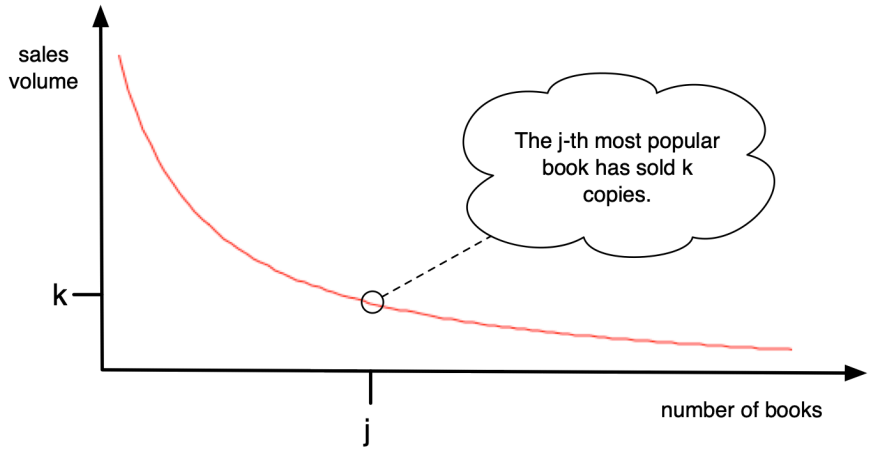
- Stereotype of media business is to focus only on **hits**
- Which area is bigger?
  - **unpopular vs popular items**
- As a function of k, how many items have sold at least k copies?
  - In the example, there are j books that have sold at least k copies.



Focus on niches

Switch the axes:

- Focus on the **long tail**: when we move to volumes of sales of many niche products. We need to compare if there is significantly more area under the left part of this curve (hits) or the right (niche products).
- As you look at less and less popular items, what sales volumes do you see?
  - The j^{th} most popular book has sold k copies.



- 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
- 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).

Zipf's law

- For the record: the previous plot is known as a **Zipf's plot**
- Introduced by **George Kinsley Zipf**, a Harvard linguistic professor
- Zipf's law usually refers to the size k of an occurrence of an event relative to it's rank j
  - it states hat the size of the j^{th} largest occurrence of the event is inversely proportional to its rank
  - $k \approx j^{-b}$, with b usually close to 1

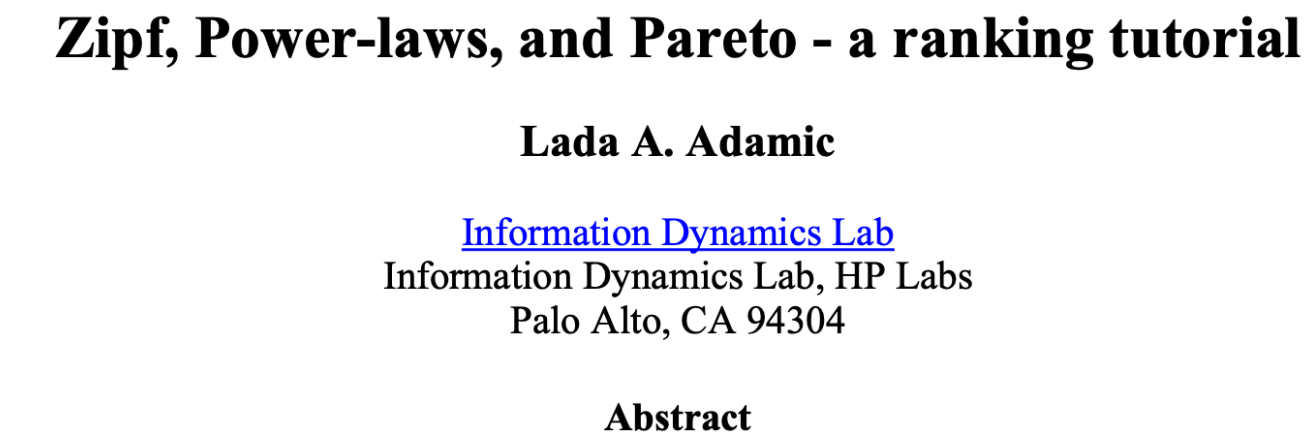
Pareto's law

- Many of you have probably found similarities with another famous law due to Pareto
- Wilfred Fritz Pareto (a former student at UniTo!) was interested in the distribution of income
- Instead of asking which was the j^{th} largest income, he asked how many people have an income greater than j
- Pareto's law is a cumulative probability distribution (cdf):
- $P(K > k) \approx k^{-\gamma}$

Three "similar" laws

- Zipf's law: $k \approx j^{-b}$
- Pareto's law: $P(K > k) \approx k^{-\gamma}$
- Power law: $f(k) \approx k^{-c}$
- They are all connected!
- It is possible to prove that $c = 1 + \gamma$ and that $\gamma = \frac{1}{b}$
- They are just three sides of the same coin!

For more information



Zipf, Power-laws, and Pareto - a ranking tutorial

Lada A. Adamic

Information Dynamics Lab
Information Dynamics Lab, HP Labs
Palo Alto, CA 94304

Abstract

- Reading material:
  - Zipf, Power-laws, and Pareto - a ranking tutorial (Adamic)
  - Power laws, Pareto distributions and Zipf's law (Newman)

UNIVERSITÀ DI TORINO

The Effect of Search Tools and Recommendation Systems

- Are Internet search tools making the richget-richer dynamics of popularity more extreme or less extreme?
  - **Two compelling but juxtaposed sides to this question**
- Search engines make the rich-get-richer dynamics more evident
  - e.g., highly-ranked pages are in turn the main ones that users see in order to formulate their own decisions about linking
- Other aspects that make the effect less extreme:
  - different queries, e.g., obscure queries, brings to different Search Engine results
  - targeted and personalized search can make unpopular items to surface
  - recommendation system's serendipity exploits the long tail argument

UNIVERSITÀ DI TORINO

Analysis of Rich-get-Richer processes

Objectives

- $f(k)$ fraction of nodes with degree k
- Goal: $f(k) \propto k^{-c}$
- Why does this happen with the Rich-Get-Richer model?
- What is the role of c?

Basic formalization

- $X_j(t)$ random variable that represents the number of links to j at a time step t
- $X_j(j) = 0$
  - since node j starts with no in-links when it is first created at time j
- We have that:

$$X_j(t+1) = X_j(t) + \frac{p}{t} + \frac{(1-p)X_j(t)}{t}$$

- where the term

$$\frac{p}{t} + \frac{(1-p)X_j(t)}{t}$$

- is **the expected change in** $X_j(t)$

UNIVERSITÀ DI TORINO

The deterministic argument

Let's suppose that:

- time runs continuously from 0 to N
- We approximate $X_j(t)$ with a continuous function $x_j(t)$
- It is like we are ignoring probabilities, and our idealized physical system just starts from a set of initial conditions and evolves in a fixed way over time
- Same initial condition $x_j(j) = 0$
- We model the rate of growth by the differential equation:

$$\frac{dx_j}{dt} = \frac{p}{t} + \frac{(1-p)x_j}{t}$$

- Rather than dealing with random variables $X_j(t)$ that move in small probabilistic **jumps** at discrete points in time, **we get to work with a quantity x_j that grows completely smoothly over time, at a rate tuned to match the expected changes in the corresponding random variables**.
- Let's set $q = 1 - p$

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

- Dividing both sides by $p + qx_j$, we get

$$\frac{1}{p + qx_j} \cdot \frac{dx_j}{dt} = \frac{1}{t}$$

- Integrating both sides:

$$\int \frac{1}{p+qx_j} \cdot \frac{dx_j}{dt} dt = \int \frac{1}{t} dt$$

$$\int \frac{1}{p+qx_j} \cdot dx_j = \int \frac{1}{t} dt$$

since $\int \frac{1}{x} \cdot dx = ln|x| + c$ and $\int \frac{1}{p+qx} \cdot dx = \frac{1}{q} ln|p+qx| + c$ (substitute $u = p + qx \Rightarrow du = q \cdot dx \Rightarrow dx = \frac{1}{q} \cdot du$)

$$q\left(\frac{\ln(p+qx_j)}{q} + c_1\right) = q(\ln t + c_2)$$

$$\ln(p+qx_j) = q\ln t + c$$

- Let us set $A = e^c$
- We can exponentiate both sides:

$$p + qx_j = At^q$$

$$x_j(t) = \frac{1}{q}(At^q - p)$$

- Recall initial condition: $x_j(j) = 0$

$$0 = x_j(j) = \frac{1}{q}(Aj^q - p)$$
$$Aj^q - p = 0$$
$$A = \frac{p}{j^q}$$

- We can substitute $A = \frac{p}{j^q}$ in $x_j(t) = \frac{1}{q}(At^q - p)$ obtaining

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

- So we solved the deterministic approximation:
- $x_j(t) = \frac{p}{q}\left[\left(\frac{t}{j}\right)^q - 1\right]$ is a closed form expression for how each x_j grows over time
- For a given value of k, and a time t, what fraction of all nodes have at least k in-links at time t?

Identifying a power law in the deterministic approximation

- For a given value of k and a time t, what fraction of all functions x_j satisfies $x_j \geq k$?

$$
\begin{aligned}
x_j(t) = \frac{p}{q}\left[\left(\frac{t}{j}\right)^q - 1\right] &\geq k \\
\left[\left(\frac{t}{j}\right)^q - 1\right] &\geq k\frac{q}{p} \\
\frac{t^q}{j^q} &\geq k\frac{q}{p} + 1 \\
t^q &\geq j^q \cdot \left(k\frac{q}{p} + 1\right) \\
j^q &\leq t^q \left(\frac{q}{p}k + 1\right) \\
j &\leq t\left(\frac{q}{p}k + 1\right)^{-\frac{1}{q}}
\end{aligned}
$$

- Out of all the functions $x_1, x_2, \ldots, x_t$ at time t, the **fraction** of values j that satisfies the above inequality is:

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

- We have the shape of a power law $F(k) \propto k^{-c}$:
- $\left(\frac{q}{p} k + 1 \right)$ is proportional to k
- $-\frac{1}{q}$ is a negative exponent

$F(x)$: fraction of nodes with **at least** in-degree k

but we aim at finding an approximation for

$f(k)$: fraction of nodes with **exactly** in-degree k

that means we can approximate $f(k)$ taking the derivative:

$$
\begin{aligned}
-\frac{dF}{dk} &= -\frac{d\left(\frac{q}{p}k+1\right)^{-\frac{1}{q}}}{dk} \\
&= \frac{1}{q}\cdot\frac{q}{p}\cdot\left(\frac{q}{p}k+1\right)^{-1-\frac{1}{q}} \\
&= \frac{1}{p}\cdot\left(\frac{q}{p}k+1\right)^{-1-\frac{1}{q}} \propto k^{-(1+\frac{1}{q})}
\end{aligned}
$$

Final step

The deterministic approximation of the model predicts that:

$$f(k) \propto k^{-\left(1+\frac{1}{q}\right)}$$

that is a power law with exponent:

$$1 + \frac{1}{q} = 1 + \frac{1}{1-p}$$

- When p is close to 1, link formation is mainly based on uniform random choices, and so the role of richget-richer dynamics is muted.

Meaning of the exponent

Let's study the behavior of the exponent:

$$\lim_{p \to 1} \left(1 + \frac{1}{1-p}\right) = \infty$$

- the exponent is infinity when link formation is mainly governed by uniform random choice $(p \to 1)$: very large numbers of in-degree are extremely rare

$$\lim_{p \to 0} \left(1 + \frac{1}{1-p}\right) = 2$$

- the growth is mainly governed by the preferential attachment process. The power law's exponent decreases toward 2, allowing for nodes with very large in-degree

UNIVERSITÀ DI TORINO

Conclusion

- Rich-Get-Richer processes explain the emergence of power laws and also exponents that in real scenarios are often slightly larger than 2
- Case Study: empirical findings in the Web showed that in-degree distributions can be fitted by a power law with exponent ≈ 2.1

UNIVERSITÀ DI TORINO

Some practical notes

Plotting empirical distributions

- When we download some data (or a sample), we have a collection of observations
- We should count the observations as a function of a given variable, then we can plot the empirical (probability) distribution
  - For example, we can count how many individuals in our sample have a given height
- If the variable has continuous values, we need to discretize these values into intervals (binning)

UNIVERSITÀ DI TORINO

Example: humans heights

- In this example, we read a dataset stored in a CSV file to create a pandas dataframe
- We can count every occurrence of heights values in different intervals, then divide every sum by the size of the sample (e.g., 10000)
- Python has a lot of pre-boiled methods and function to do that

```
hw_df = pd.read_csv("datasets/weight-height.csv")
hw_df
```

	Gender	Height	Weight
0	Male	73.847017	241.893563
1	Male	68.781904	162.310473
2	Male	74.110105	212.740856
3	Male	71.730978	220.042470
4	Male	69.881796	206.349801
...	...	...	...
9995	Female	66.172652	136.777454
9996	Female	67.067155	170.867906
9997	Female	63.867992	128.475319
9998	Female	69.034243	163.852461
9999	Female	61.944246	113.649103

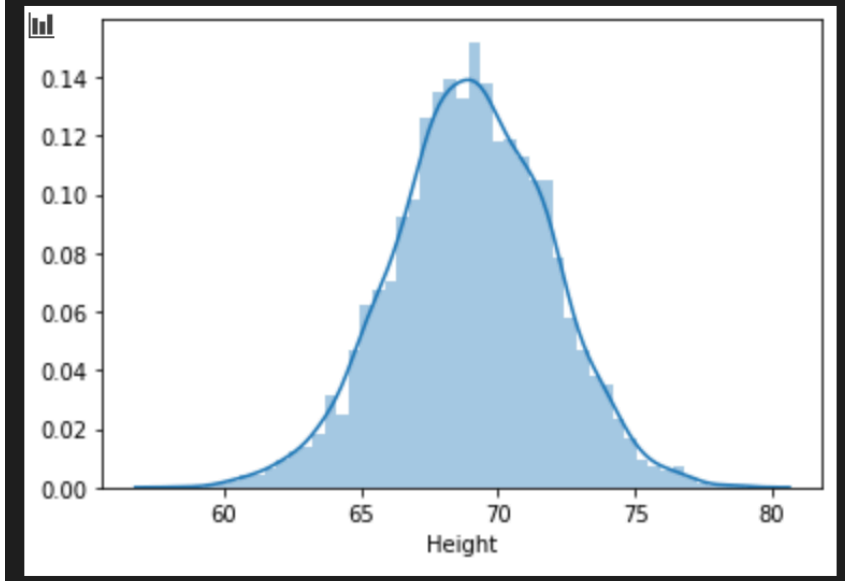
10000 rows × 3 columns

Histograms

- The histogram is a natural choice
- you can also try scatterplots
- Library seaborn has functions that make everything: counting, normalizing, binning, fitting

```
hw_df_male=hw_df[hw_df.Gender=='Male']
sns.distplot(hw_df_male.Height)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fcd711ef990>
```



The 3 sigma rule of thumb

- Given an empirical distribution, we can check where the observed data falls
- The three-sigma rule or 68-95-99.7 rule, is a statistical rule which states that for a normal distribution, almost all observed data will fall within three standard deviations (denoted by δ) of the mean or average (denoted by μ)

```
mu = hw_df_male.mean().Height
sigma = hw_df_male.std().Height
sample100 = hw_df_male.sample(100).Height
np.sum((sample100.values >= mu-sigma) & (sample100.values <= mu+sigma))/100
0.62

np.sum((sample100.values >= mu-2*sigma) & (sample100.values <= mu+2*sigma))/100
0.97

np.sum((sample100.values >= mu-3*sigma) & (sample100.values <= mu+3*sigma))/100
0.99
```

Exercises

- Download a sample of the Web graph, for example, from here
- Create a directed graph from the Web sample
- Generate a random graph with an equal number of nodes and edges for comparison
- Calculate degree distributions of both graphs, and plot them
- Estimate heterogeneity of both graphs
- Is the 3 sigma rule useful here?
- Can you say if some degree distribution would be fitted with a power law?

UNIVERSITÀ DI TORINO

Reading material

[ns2] **Chapter 18 (18.1 - 18.7) Power Laws and Rich-Get-Richer Phenomena**

Please check your general understanding of the topic completing the exercises at the end of the chapter

UNIVERSITÀ DI TORINO

Q&A