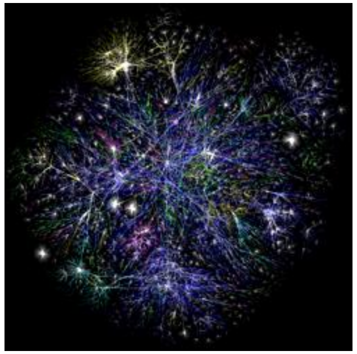
Information Networks

Generative processes for Power Laws and Scale-Free networks

Lecture 4



Announcement

- The first assignment will be handed out after the spring break
- The second assignment around middle of April

- You should send me an e-mail about the type of project that you intend to do within the next week.

Power Laws - Recap

- A (continuous) random variable X follows a power-law distribution if it has density function

$$p(x) = Cx^{-\alpha}$$

- A (continuous) random variable X follows a Pareto distribution if it has cumulative function

$$P[X \geq x] = Cx^{-\beta}$$

power-law with $\alpha = 1+\beta$

- A (discrete) random variable X follows Zipf's law if the frequency of the r-th largest value satisfies

$$p_r = Cr^{-\gamma}$$

power-law with $\alpha = 1+1/\gamma$

Power Laws – Generative processes

- We have seen that power-laws appear in various natural, or man-made systems
- What are the processes that generate power-laws?
- Is there a “universal” mechanism?

Combination of exponentials

- If variable Y is exponentially distributed

$$p(y) \sim e^{ay}$$

- If variable X is exponentially related to Y

$$X \sim e^{bY}$$

- Then X follows a power law

$$p(x) \sim x^{-(1+a/b)}$$

- Model for population of organisms

Monkeys typing randomly

- Consider the following generative model for a language [Miller 57]
  - The space appears with probability q_s
  - The remaining m letters appear with equal probability $(1-q_s)/m$

- Frequency of words follows a power law!
- Real language is not random. Not all letter combinations are equally probable, and there are not many long words

Least effort principle

- Let C_j be the cost of transmitting the j-th most frequent word

$$C_j \sim \log_m j$$

- The average cost is

$$C = \sum_{j=1}^{n} p_j C_j$$

- The average information content is

$$H = -\sum_{j=1}^{n} p_j \log_2 p_j$$

- Minimizing cost per information unit C/H yields

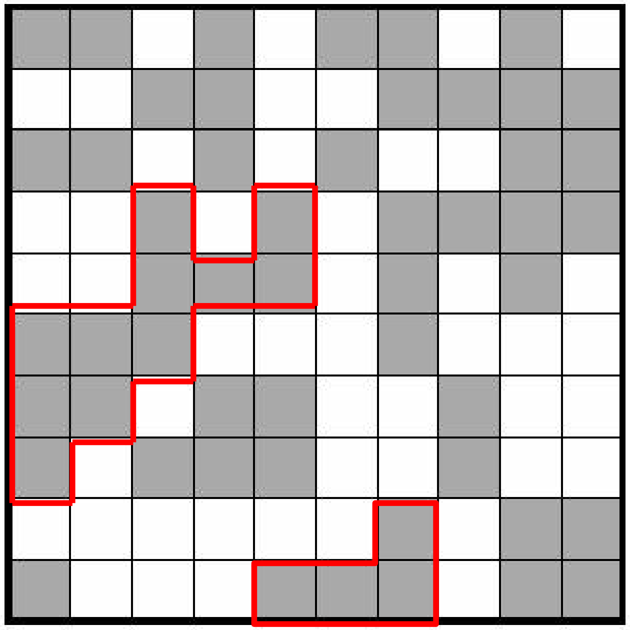
$$p_j \sim j^{-\alpha}$$

Critical phenomena

- When the characteristic scale of a system diverges, we have a phase transition.
- Critical phenomena happen at the vicinity of the phase transition. Power-law distributions appear

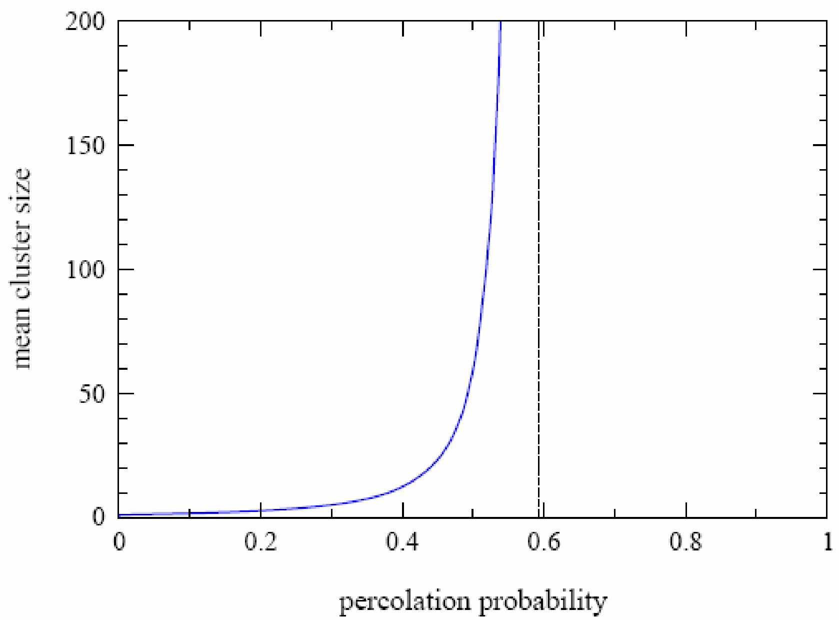
Percolation on a square lattice

- Each cell is occupied with probability p



- What is the mean cluster size?

Critical phenomena and power laws

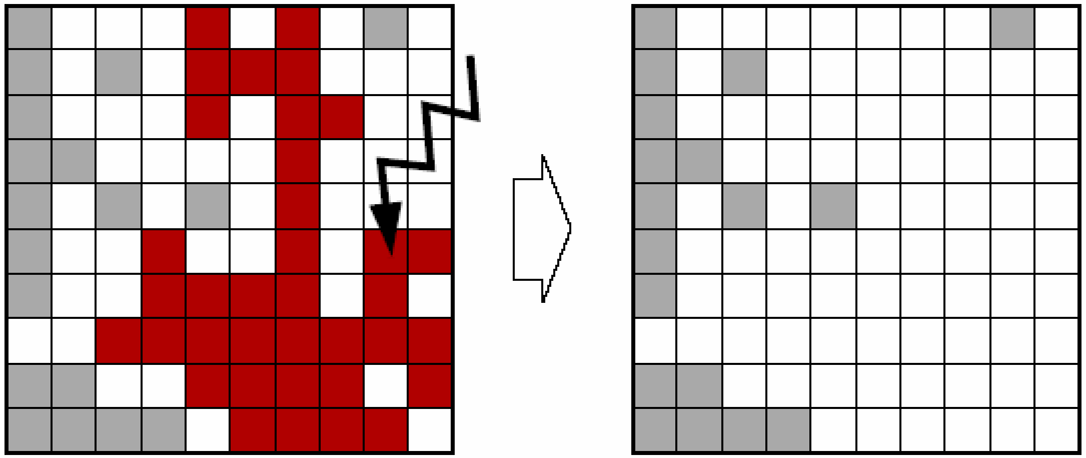


$p_c = 0.5927462\ldots$

- For $p < p_c$ mean size is independent of the lattice size
- For $p > p_c$ mean size diverges (proportional to the lattice size - percolation)
- For $p = p_c$ we obtain a power law distribution on the cluster sizes

Self Organized Criticality

- Consider a dynamical system where trees appear in each cell with probability p, and fires strike cells randomly



- The system eventually stabilizes at the critical point, resulting in power-law distribution of cluster (and fire) sizes

Preferential attachment

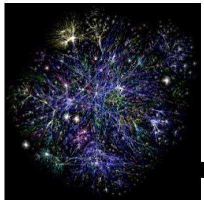
- The main idea is that “the rich get richer”
  - first studied by Yule for the size of biological genera
  - revisited by Simon
  - reinvented multiple times
- Also known as
  - Gibrat principle
  - cumulative advantage
  - Mathew effect

The Yule process

- A genus obtains species with probability proportional to its current size
- Every m new species, the m+1 species creates a new genus

- The sizes of genera follows a power law with

$$p_k \sim k^{-(2+1/m)}$$



Preferential Attachment in Networks

- First considered by [Price 65] as a model for citation networks
  - each new paper is generated with m citations (mean)
  - new papers cite previous papers with probability proportional to their indegree (citations)
  - what about papers without any citations?
    - each paper is considered to have a “default” citation
    - probability of citing a paper with degree k, proportional to k+1
- Power law with exponent $\alpha = 2+1/m$

Barabasi-Albert model

- Undirected(?) model: each node connects to other nodes with probability proportional to their degree
  - the process starts with some initial subgraph
  - each node comes with m edges
- Results in power-law with exponent $\alpha = 3$

The LCD model [Bollobas-Riordan]

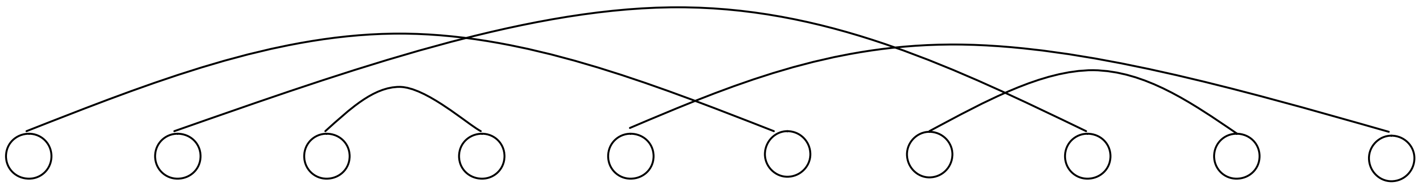
- Self loops and multiple edges are allowed
- A new vertex v, connects to a vertex u with probability proportional to the degree of u, counting the new edge.
- The m edges are inserted sequentially, thus the problem reduces to studying the single edge problem

Linearized Chord Diagram

- Consider $2n$ nodes labeled $\{1,2,\ldots,2n\}$ placed on a line in order.

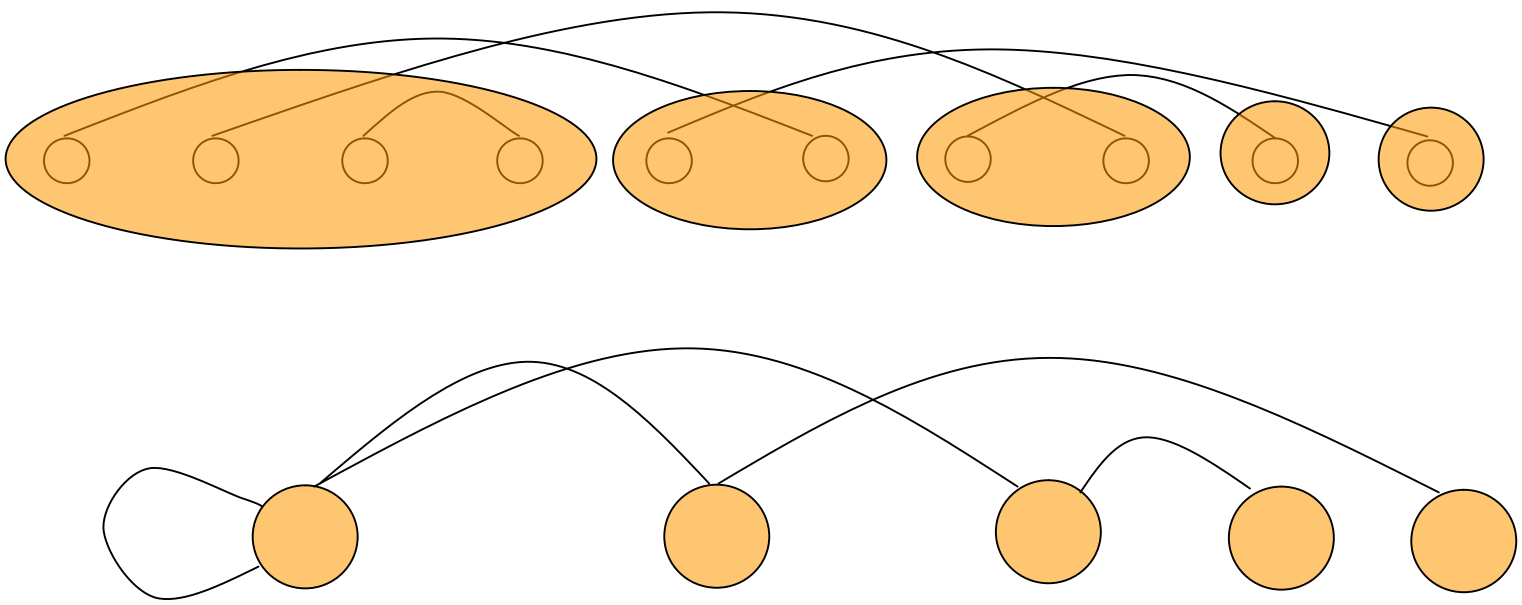
Linearized Chord Diagram

- Generate a random matching of the nodes.



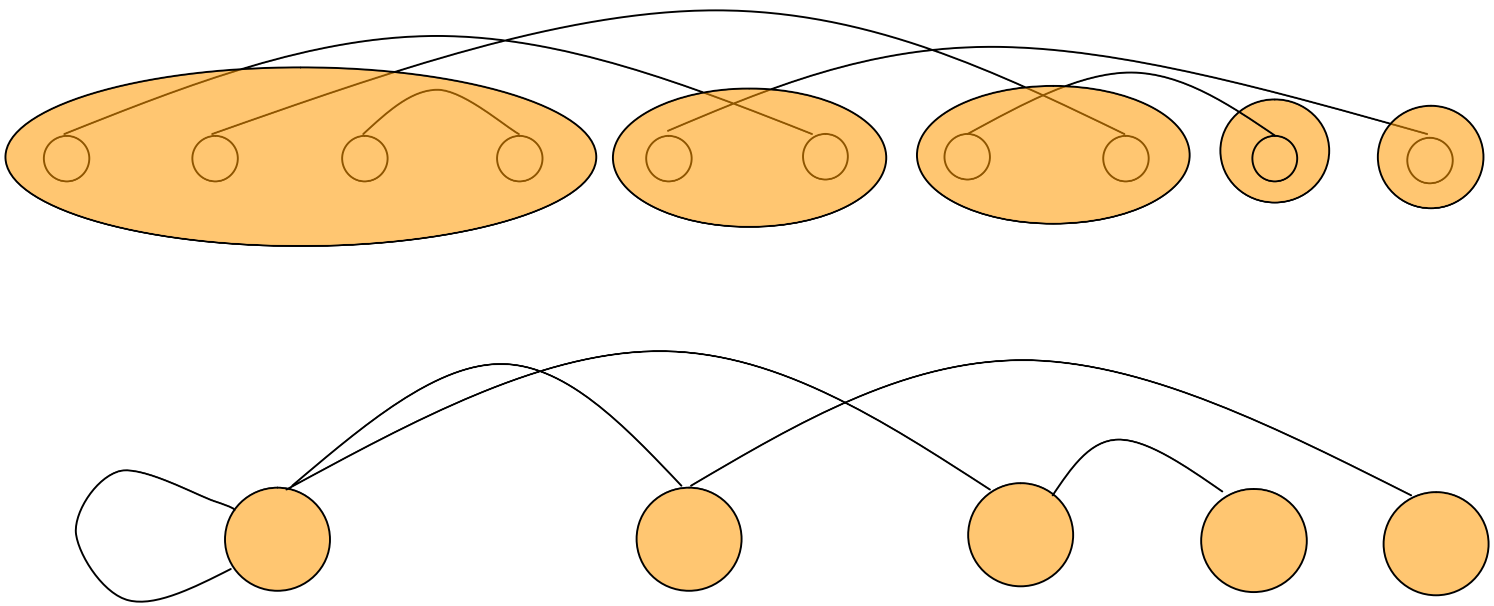
Linearized Chord Diagram

- Starting from left to right identify all endpoints until the first right endpoint. This is node 1. Then identify all endpoints until the second right endpoint to obtain node 2, and so on.



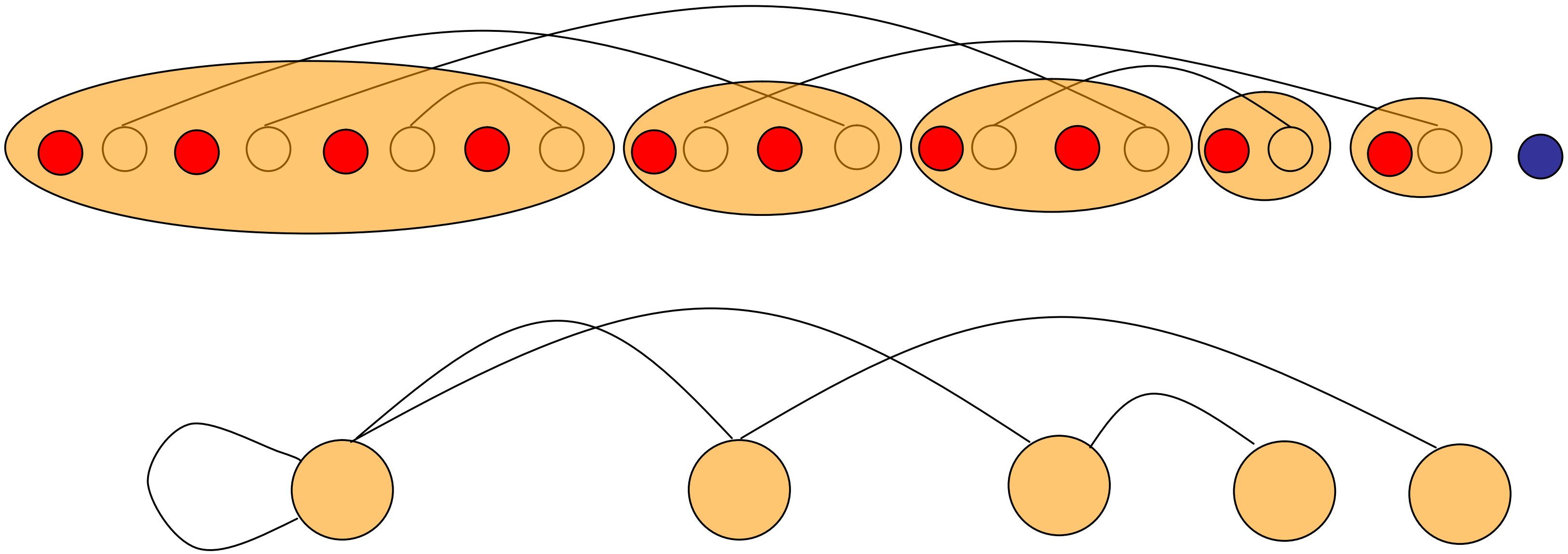
Linearized Chord Diagram

- Uniform distribution over matchings gives uniform distribution over all graphs in the preferential attachment model



Linearized Chord Diagram

- Uniform distribution over matchings gives uniform distribution over all graphs in the preferential attachment model
  - each supernode has degree proportional to the nodes it contains
  - consider adding a new chord, with the right endpoint being in the rightmost position and the left being placed uniformly

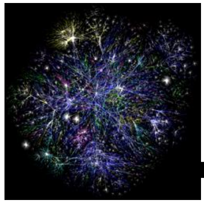


Preferential attachment graphs

- Expected diameter
  - if m = 1, the diameter is $\Theta(\log n)$
  - if m > 1, the diameter is $\Theta(\log n/\log\log n)$

- Expected clustering coefficient

$$E\left[C^{(2)}\right] = \frac{m-1}{8}\frac{\log^2 n}{n}$$



Weaknesses of the BA model

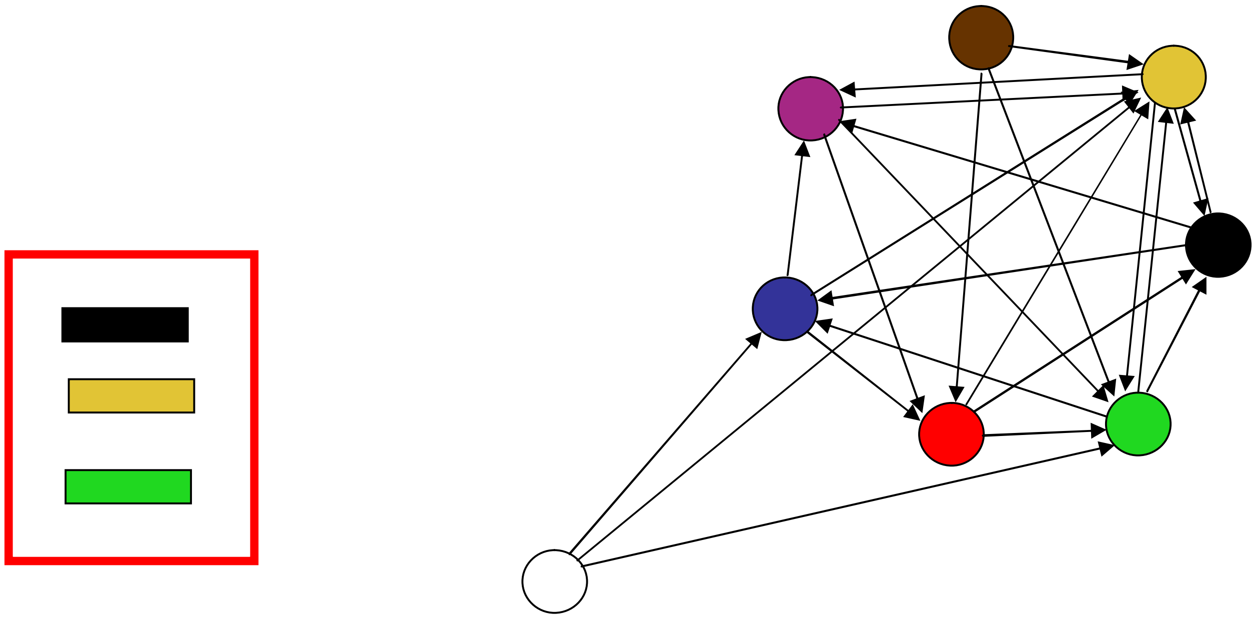
- It is not directed (not good as a model for the Web) and when directed it gives acyclic graphs
- It focuses mainly on the (in-) degree and does not take into account other parameters (out-degree distribution, components, clustering coefficient)
- It correlates age with degree which is not always the case
- Yet, it was a significant contribution in the network research

- Many variations have been considered some in order to address the above problems
  - edge rewiring, appearance and disappearance
  - fitness parameters
  - variable mean degree
  - non-linear preferential attachment

Copying model

- Each node has constant out-degree d
- A new node selects uniformly one of the existing nodes as a prototype
- For the i-th outgoing link
  - with probability α it copies the i-th link of the prototype node
  - with probability 1- α it selects the target of the link uniformly at random

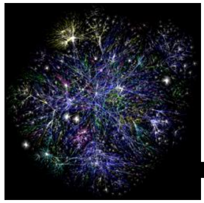
An example



Copying model properties

- Power law degree distribution with exponent $\beta = (2-\alpha)/(1-\alpha)$
- Number of bipartite cliques of size i x d is ne^{-i}

- The model was meant to capture the topical nature of the Web
- It has also found applications in biological networks



Other graph models

- Cooper Frieze model
  - multiple parameters that allow for adding vertices, edges, preferential attachment, uniform linking
- Directed graphs [Bollobas et al]
  - allow for preferential selection of both the source and the destination
  - allow for edges from both new and old vertices

Other interesting distributions

- The log-normal distribution
  - The variable Y = log X follows a normal distribution

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/2\sigma^2}$$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-(\ln x-\mu)^2/2\sigma^2}$$

$$\ln f(x) = -\frac{(\ln x)^2}{2\sigma^2} + \left(\frac{\mu}{\sigma^2} - 1\right)\ln x - \ln\sqrt{2\pi}\sigma - \frac{\mu^2}{2\sigma^2}$$

Lognormal distribution

- Generative model: Multiplicative process

$$X_j = F_j X_{j-1}$$

- Central Limit Theorem: If $X_1, X_2, \ldots, X_n$ are i.i.d. variables with mean m and finite variance s, then if $S_n = X_1 + X_2 + \ldots + X_n$

$$\frac{S_n - nm}{\sqrt{ns^2}} \sim N(0,1)$$

- When the multiplicative process has a reflective boundary, it gives a power law distribution

Double Pareto distribution

- Run the multiplicative process for T steps, where T is an exponentially distributed random variable

- Double Pareto: Combination of two Pareto distributions

References

- M. E. J. Newman, The structure and function of complex networks, SIAM Reviews, 45(2): 167-256, 2003
- M. E. J. Newman, Power laws, Pareto distributions and Zipf's law, *Contemporary Physics*
- M. Mitzenmacher, A Brief History of Generative Models for Power Law and Lognormal Distributions**,** Internet Mathematics
- R. Albert and L.A. Barabasi, Statistical Mechanics of Complex Networks, Rev. Mod. Phys. 74, 47-97 (2002).
- B. Bollobas, Mathematical Results in Scale-Free random Graphs