CS 144 – Homework 3

Marco Yang

Using 2 late tokens for this.

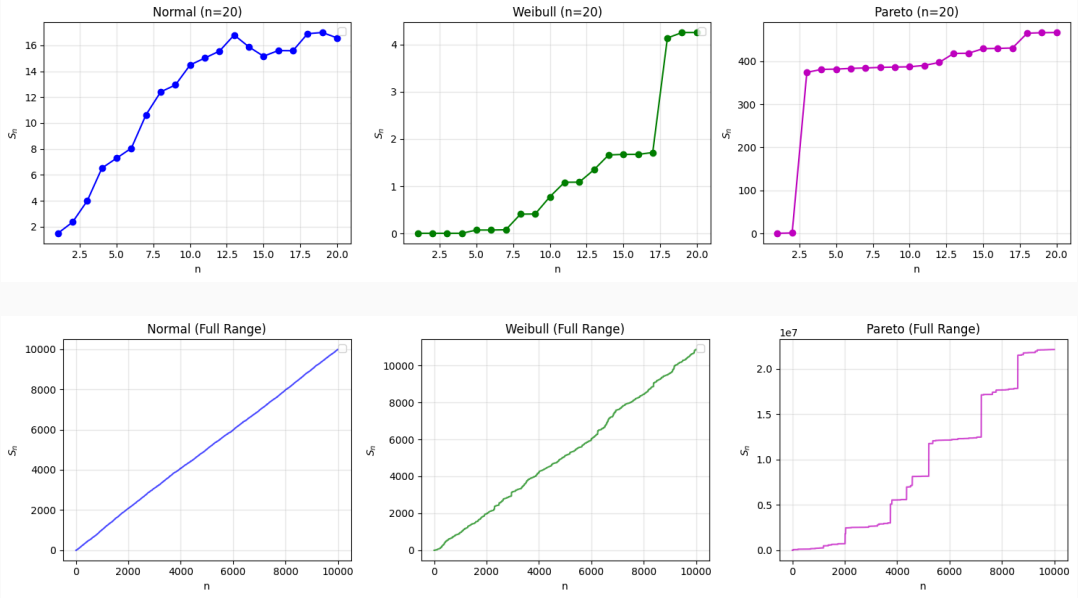
Coding + Data Analysis

1. **Heavy vs. Light**

   (a) **Law of Large Numbers**

   Plot S_n vs n for the Normal, Weibull, and Pareto distributions.

   Solution:

   For $n = 20$, there is a linear trend for the first two plots, but it's fairly noisy. When plotting all points, it's clear that the Normal distribution follows a line almost perfectly, and the Weibull also follows a line pretty well. This makes sense since for large n, the mean of the sample should converge to the expected value, which means $S_n = n\mathbb{E}[X]$ and the slope should be $\mathbb{E}[X]$. The Pareto distribution doesn't follow a line since it's heavy tailed and its expected value is infinite.

   (b) **CLT**

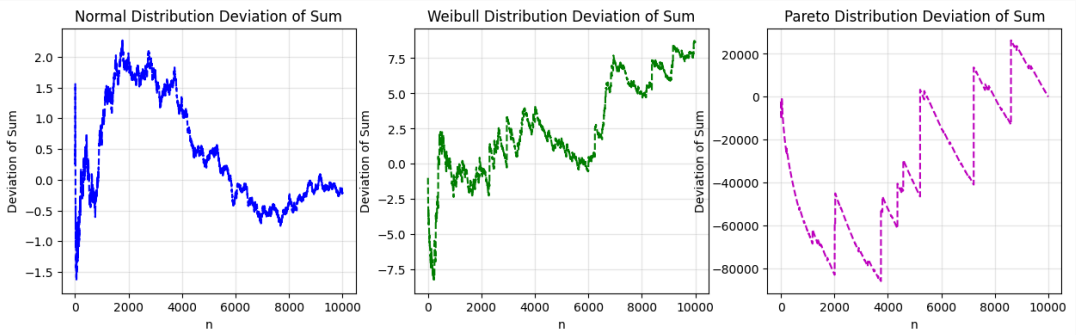
   Plot the number of standard deviations the empirical sum of n samples is from the expected sum of n samples.

   Solution: When plotting the deviation of the sum from the expected value, we get that for the normal distribution, it seems random at first but eventually evens converges around 0 over time. This makes sense because the CLT states that the distribution of sample means should approach a normal distribution centered at 0, so over time the mean should converge to 0.

   For the Weibull distribution, it crawls up over time, and ends at 7.7. In theory, it's still a light-tailed distribution so it should converge, but it doesn't seem to be for some reason.

Perhaps the variance of the Weibull distribution is much higher, so it might converge if I have more samples, but I tried 100,000 samples and it still doesn't work so I give up.

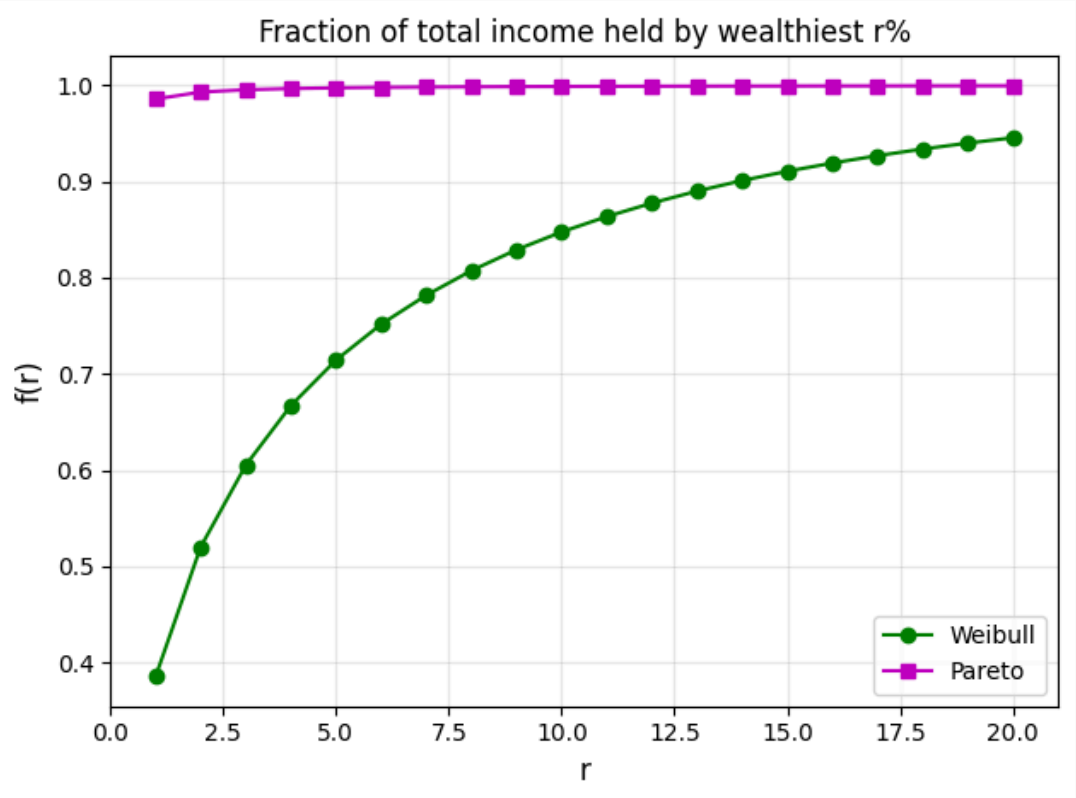
For the Pareto distribution, it increases and decreases in spikes, with the overall trend being positive and ending at 0. This makes sense since I used the sample mean (so at the end it should revert to 0), and the spiky trend is due to the fact that the distribution is heavy tailed, so every once in a while, we will get a wildly different value that completely messes up the sample mean. It also shouldn't converge to any specific value since the expected mean for a heavy-tailed distribution is infinity.



(c) **80-20 Rules**

Plot the fraction $f(r)$ of the total income of the city held by the wealthiest $r\%$ of the population for $r \in [1, 20]$.

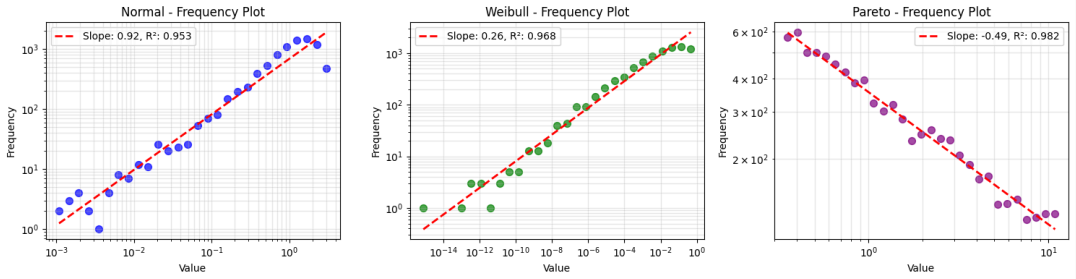
Solution:



The Weibull distribution's proportion of wealth held by the top $r\%$ increases with diminishing returns, which makes sense since it's light tailed. The probability of getting a singular person or a few data points with a value so large that it dominates to entire distribution is close to 0. However, the Pareto distribution is heavy-tailed, so one data point completely dominating the rest is somewhat probable. This is shown in the plot since the top 1% holds almost all of the wealth.

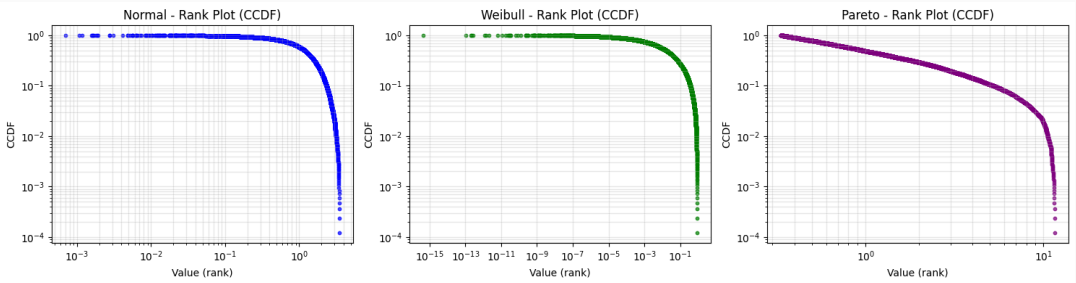
(d) **Identifying heavy tails**

Solution: We filter out the outliers before making the log-log plot because an outlier might be able to make a light-tailed distribution seem heavy tailed.



Based on my plots of the normal and weibull distributions above, to identify a light-tailed distribution, we should look for a positive linear trend in the log-log frequency plot with the domain ending around the mean. This makes sense since the x-axis grows on a log scale, so for a light-tailed distribution, there should be few values past the mean on the log scale that have a significant log frequency. It is safe to assume that the mean value of a sample would be expected given this plot since the "mode" of the data is near the mean, which is also the largest value on the graph.

In the Pareto frequency plot, we have a negative linear trend. This is typical of heavy-tailed distributions like we learned in lecture, and the fact that the log frequency is only linearly with the log value indicates that very large values (compared to the "mode" of the data, which is one of the much smaller values) will play a significant role in the mean.



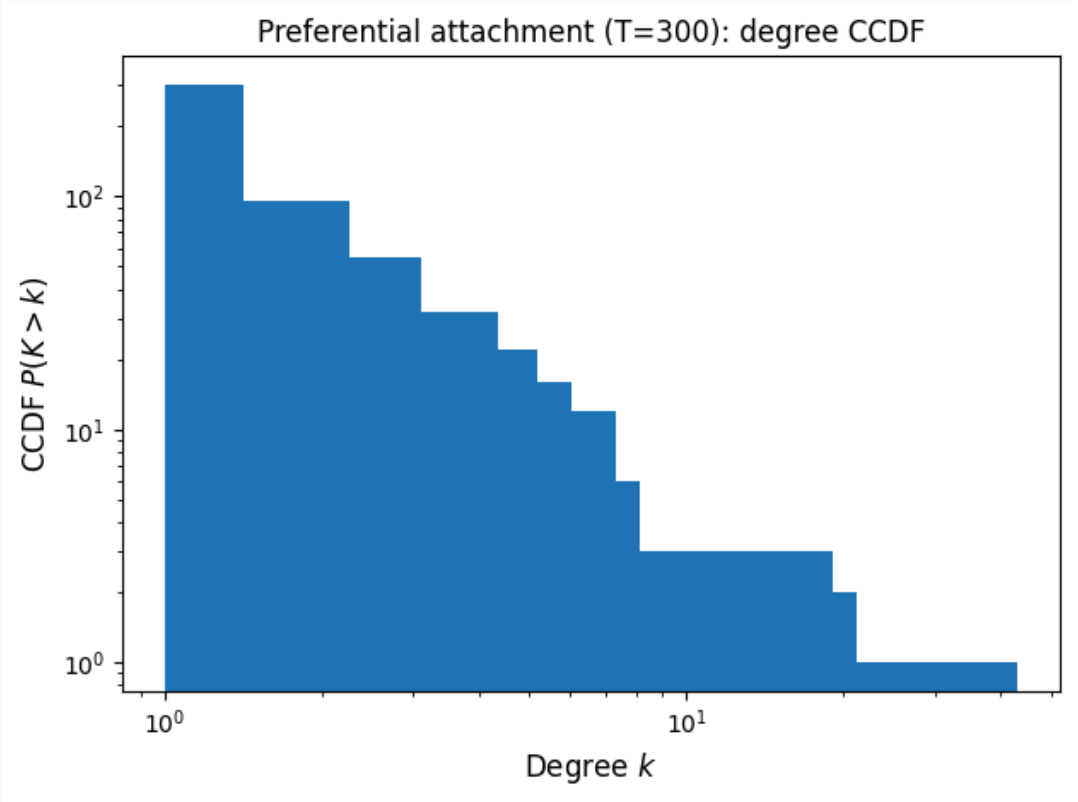
The log-log rank plots for the light-tailed distributions (normal and Weibull) show negative exponential trend, but the log-log rank plot of the heavy-tailed distribution (Pareto) is linear at the start before also becoming more like a negative exponential.

2. **The Devil is in the Details**

(a) **Preferential attachment model**

Code a preferential model generator and plot the empirical ccdf on a log-log scale.

Solution:



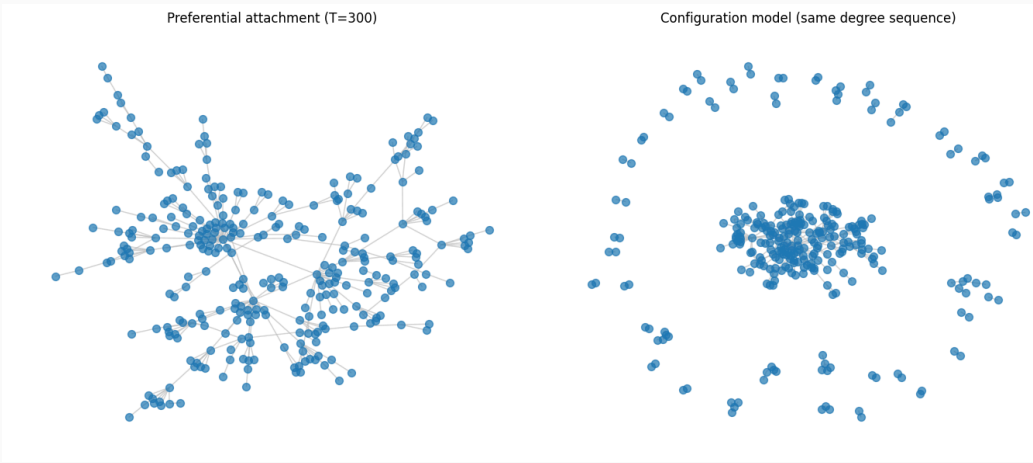
The CCDF on the log log scale has a negative linear trend, which makes sense since we also observed the same trend in the CCDF for the Pareto distribution, and both distributions are heavy-tailed.

(b) **Preferential attachment model**

Solution: Code is in notebook.

(c) Visualize and compare the above 2 models.

Solution: The preferential attachment model is one SCC, with many small individual cliques/clusters, while the configuration model has one massive cluster and a bunch of lonely nodes floating around.


Preferential attachment (T=300)
Configuration model (same degree sequence)


Preferential attachment — 4 instances
PA (seed=0)
PA (seed=1)
PA (seed=2)
PA (seed=3)

Theory

1. **When monkeys type**

   Imagine a monkey randomly typing away on a keyboard. The keyboard has n characters ($n > 1$ is a constant) and a space bar. The monkey hits the space bar with probability q and each of the n characters with equal probability $\frac{1-q}{n}$, with $\frac{1-q}{n} < q$. Each keystroke is independent.

   (a) What is the probability that the monkey types a particular c-letter word (specific c letters followed by a space).

   Solution:

   $$p(\text{word with } c \text{ letters}) = \left(\frac{1-q}{n}\right)^c q.$$

   (b) Rank all the words the monkey can type in decreasing order of probability. Assume for simplicity that rank 1 is assigned to the empty word of length 0.

   Let's take $n = 10$ as an example. The empty word has rank 1. There are 10 possible 1-leter words, which we assign ranks 2-11 in an arbitrary order. Similarly, there are 10^2 possible 2-

letter words, which we assign ranks 12-111, and so on. So any n-letter word will have some minimum and maximum rank.

Let P_r be the probability of occurence of a particular word with rank r. Show that

$$\lim_{r\to\infty} \frac{\log(P_r)}{\log(r)} = \log_n\left(\frac{1-q}{n}\right)$$

where $\log_n(\cdot)$ denotes the logarithm to the base n.

Hint: What is the relationship between the rank r and the length of the word? Think about the above example.

Solution: We know that a word w with c letters will have rank

$$\sum_{i=0}^{c-1} n^i < r \leq \sum_{i=0}^{c} n^i.$$

Since we know that $n^c > \sum_{i=1}^{c-1} n^i$ for any i, n,

$$n^{c-1} < r < n^{c+1}.$$

Then, the log of the rank r is in

$$\log n^{c-1} < \log r < \log n^{c+1}$$
$$(c-1)\log n < \log r < (c+1)\log n.$$

The log probability of w is

$$\log P_r = \log p(w) = \log\left(\frac{1-q}{n}\right)^c = c\left(\log\frac{1-q}{n}\right).$$

From our range for the log rank,

$$c > \frac{\log r}{\log n} - 1.$$

Thus, as $r \to \infty$, $c \to \infty$ as well. For large c, $c-1 \approx c \approx c+1$, so we can approximate the log rank as $\log r \approx c \log n$. Plugging this into the limit,

$$\lim_{r\to\infty} \frac{\log(P_r)}{\log(r)} = \lim_{r\to\infty} \frac{c\log\frac{1-q}{n}}{c\log n}$$
$$= \log_n \frac{1-q}{n}.$$

(c) Interpret the result in part (b). Specifically, what can you say about the distribution of r (i.e. the distribution P_r)? Attempt to explain the difference between this distribution and the result of part (a).

Solution: In our answer for part b, we showed that the P_r is linear with r on a log log scale as $r \to \infty$, which means that P_r is a heavy-tailed distribution. However, the distribution in part a is exponential, which we know is not heavy-tailed.

2. **Friendship Paradox, Part 2**

We started our discussion of the "Friendship Paradox" in HW 1, where you showed that, in a graph with average degree μ and variance σ^2, the average degree of a neighbor is $\mu + \frac{\sigma^2}{\mu}$. That is, your average friend has more friends than you do. This is surprising! How dare they? But, so far, you've only shown a "weak" version of the paradox – one that holds in expectation. The goal in this problem is to show that the paradox can be strengthened. In particular, one can prove that the ratio of the degree of a random neighbor of a node to the degree of the node is *very likely* to be large.

Strengthening the paradox will however depend on the network model we consider. For this problem, we will use the **configuration model** you explored in Problem 1.2.

The configuration model is defined as follows. The model generates a random graph on n nodes with a desired, fixed degree sequence, i.e., with the number of nodes with degree d matching a specified f_d. To do this, nodes are created with "stub" edges according to the desired degree. Then, the two stubs are chosen uniformly at random and connected to form an edge. (Note that this can lead to self loops and multiple edges for any pair of nodes.) This is repeated until no stubs remain. See Figure 2 for an illustration.

We will focus on configuration graphs with Pareto degree distributions, i.e., where $f_d = c \cdot d^{-\alpha}$ for some normalizing constant c and $\alpha \in (1, 2)$. You worked with the Pareto distribution in Problem 1.1, and here are some simple (and useful) facts to remember:

- Let δ be the highest degree in the graph. Since we consider Pareto degree distributions, we have that the highest degree satisfies $c\delta^{-\alpha} = 1$, and so $\delta = c^{\frac{1}{\alpha}}$.
- The number of nodes in the graph, n, is on the order of $\Theta(c)$.
- The number of edges in the graph, $m = |E|$, is on the order of $\Theta\left(c^{\frac{2}{\alpha}}\right)$.

The goal of this problem is to prove the following, stronger version of the friendship paradox: For a configuration graph with a Pareto degree distribution having $\alpha \in (1, 2)$, consider a random node, v_1, and a random neighbor of it, v_2. The ratio between the degree of v_2 and the degree of v_1 is $\Omega\left(n^{1-(\frac{1}{\alpha})-\varepsilon}\right)$ with constant probability.

Your task: To prove this result, we will take the following approach.

(i) Show that, the expected degree of a node chosen uniformly at random is $\Theta\left(c^{\frac{2}{\alpha}-1}\right)$.
(ii) Use (i) to show that, with a constant probability, a node selected uniformly at random will have a "small" degree, specifically, a degree that is $O\left(c^{\frac{2}{\alpha}-1}\right)$. *Hint: The Markov's inequality might be useful here.*
(iii) Show that the total number of stubs adjacent to nodes of "large" degree, i.e., degree $\geq c^{\frac{1}{\alpha}-\varepsilon}$, is $\Theta(m)$. *Hint: Consider summing up the number of stubs for high-degree nodes.*
(iv) Given (iii), what is the probability that a uniformly chosen node has a neighbor with degree greater than $c^{\frac{1}{\alpha}-\varepsilon}$?
(v) Combine (ii) and (iv) together to prove the result.

Solution: From the linearity of expectation, we know that the expected number of degrees is the same as the expected total degrees across all nodes divided by n, which is also equal to the sum of the expected nodes with each degree times each possible degree divided by n, which is

$$\begin{aligned}\mathbb{E}[d] &= \frac{1}{n}\sum_{d} f_d \cdot d \\ &= \frac{1}{n}\sum_{d=1}^{\delta} c \cdot d^{-\alpha} \cdot d \\ &= \frac{c}{n}\sum_{d=1}^{\delta} d^{1-\alpha}.\end{aligned}$$

Approximating this integral as a Riemann sum for large n,

$$\begin{aligned}\mathbb{E}[d] &\approx \frac{c}{n}\int_{1}^{\delta} x^{1-\alpha}\,\mathrm{d}x \\ &= \frac{c}{n}\left[\frac{x^{2-\alpha}}{2-\alpha}\right]_{1}^{\delta} \\ &= \frac{c}{n}\cdot\frac{\delta^{2-\alpha}-1}{2-\alpha}\end{aligned}$$

Since $\delta = c^{\frac{1}{\alpha}}, n = \Theta(c), \delta \gg 1$,

$$\begin{aligned}\mathbb{E}[d] &= \frac{c}{c}\cdot\frac{c^{\frac{2}{\alpha}-1}-1}{2-\alpha} \\ &= \Theta\left(c^{\frac{2}{\alpha}-1}\right).\end{aligned}$$

Now, applying Markov's inequality to the probability of selecting a node with a degree more than $O\left(c^{\frac{2}{\alpha}-1}\right)$,

$$P\left(d \geq O\left(c^{\frac{2}{\alpha}-1}\right)\right) \leq \frac{E[d]}{O\left(c^{\frac{2}{\alpha}-1}\right)}.$$

Since $E[d] = \Theta\left(c^{\frac{2}{\alpha}-1}\right) = kc^{\frac{2}{\alpha}-1}$ for some constant k, we know that the set of "probable" degrees is within $O\left(c^{\frac{2}{\alpha}}-1\right)$; otherwise, for $d' > O\left(c^{\frac{2}{\alpha}-1}\right) \Rightarrow d' \gg kc^{\frac{2}{\alpha}-1}$ for any constant k,

$$P(d \geq d') \leq \frac{kc^{\frac{2}{\alpha}-1}}{d'} \approx 0.$$

Summing up the total number of stubs adjacent to the high degree nodes,

$$
\begin{aligned}
S_{\text{high degree}} &= \sum_{d=c^{\frac{1}{\alpha}-\varepsilon}}^{\delta} f_d \cdot d \\
&\approx \int_{c^{\frac{1}{\alpha}-\varepsilon}}^{\delta} c \cdot x^{1-\alpha}\,\mathrm{d}x \\
&= c\left[\frac{x^{2-\alpha}}{2-\alpha}\right]_{c^{\frac{1}{\alpha}-\varepsilon}}^{\delta} \\
&= \frac{c\left(\left(c^{\frac{2}{\alpha}-1} - c^{\frac{2}{\alpha}-1-2\varepsilon+\varepsilon\alpha}\right)\right)}{2-\alpha} \\
&= \frac{c^{\frac{2}{\alpha}}\left(1 - c^{\varepsilon(\alpha-2)}\right)}{2-\alpha}.
\end{aligned}
$$

Since $n = \Theta(c)$, as $n \to \infty$, $c \to \infty$. With $\alpha - 2 < 0$, $c^{\varepsilon(\alpha-2)} \to 0$, so

$$
S_{\text{high degree}} \approx \frac{c^{\frac{2}{\alpha}}}{2-\alpha} = \Theta\left(c^{\frac{2}{\alpha}}\right).
$$

Since each stub corresponds to a neighbor, the probability that a randomly chosen node is a neighbor of a high degree node is the sum of all the stubs (edges leading to a high degree neighbor) divided by the total number of stubs (choosing a node uniformly at random and then picking an edge). Since the total number of edges is $m = \Theta\left(c^{\frac{2}{\alpha}}\right)$, the total number of stubs is $2m = 2\Theta\left(c^{\frac{2}{\alpha}}\right) = \Theta\left(c^{\frac{2}{\alpha}}\right)$. Thus, the probability of picking a high degree neighbor after choosing a node uniformly at random is

$$
p_{\text{high degree}} = \frac{S_{\text{high degree}}}{2m} = \frac{\Theta\left(c^{\frac{2}{\alpha}}\right)}{\Theta\left(c^{\frac{2}{\alpha}}\right)},
$$

which is constant.

Since both the probability of picking a “low degree node” and a “high degree node” are both constant, the ratio of the probability of getting a high degree node vs getting a low degree node is

$$
\frac{\Theta\left(c^{\frac{1}{\alpha}-\varepsilon}\right)}{O\left(c^{\frac{2}{\alpha}-1}\right)} = \Omega\left(c^{1-\frac{1}{\alpha}-\varepsilon}\right)
$$

with constant probability.

3. **Exploiting the Long Tail**

   Many companies have business models designed specifically to exploit the existence of heavy-tailed distributions. A nice pop-press book on the topic is “The Long Tail” by Chris Anderson.

   Probably the most cited example of this is Amazon. Compared to a brick-and-mortar store, which can sell only a small number of popular items; Amazon can sell a huge variety of items, including ones that are not particularly popular. If the world was light-tailed, this advantage would be small, but since the world is heavy-tailed, these unpopular items still make up a large revenue source in the aggregate.

In this problem, we'll study a simple characterization of consumer behavior to show why Amazon's business model is so profitable. This model should be very reminiscent of the preferential attachment model we studied in class.

The model: For simplicity, consider a market where at each time, t, a consumer arrives and makes two purchases: one based on individualistic behavior and one based on social behavior. These purchases are made simultaneously, so the product just chosen as an individual purchase should not be considered as a social purchase at the same time.

We will define $m_i(t)$ to be the sales volume of product i at time t. The individual purchase is a product that no one has ever purchased before. For convenience, we assume that the product i is chosen at time $t = i$ for individual purchase, which can be seen as just a relabelling of the items. In other words, $m_t(t) = 1$.

The social purchase is a product that many consumers have purchased. The likelihood that the consumer chooses a given product at time t for his social purchase is proportional to the volume at the previous time step, $m_i(t-1)$, of the product, i.e., this purchase is a consequence of peer pressure. Assume that the first customer (at $t = 1$) does not make a social purchase.

Your task: You will analyze the model above to determine the distribution of $m_i(t)$. The following steps should guide you.

(a) What is the expected increase in market size ($m_i(t)$) for product i between time t and $t+1$? It helps to split this into two possible cases: whether $i = t$ or $i < t$, which relates them to whether they are eligible for purchases by individualistic or social behavior.

> **Solution**: For $i = t$, we know that the market size increased from 0 to 1 since the product was never purchased before. For $i < t$, we have that
>
> $$\begin{aligned}\mathbb{E}[m_i(t+1) - m_i(t)] &= \frac{m_i(t)}{\sum_{j=1}^{i-1} m_j(t)} \\ &= \frac{m_i(t)}{2t-1}\end{aligned}$$
>
> since the probability of product i having been bought is proportional to the market share it had and the total number of items bought before the individual purchase at time t is $2t-1$ (2 items bought per unit of time up until now, minus one since we haven't made the individual purchase for time t yet).

(b) Using mean value and continuous time approximation, as we did for the preferential attachment model in class, argue that the following differential equation describes the rate of change of volume of product i at time t:

$$\frac{dm_i(t)}{dt} = \frac{m_i(t)}{2t-1} \text{ for } t > i$$

> **Solution**: If we approximate the function $m_i(t)$ as a continuous one and let $\delta t = 1 = \mathrm{d}t$, we have that the change between $m_i(t+1)$ and $m_i(t)$ is the derivative of m_i at time t. Since the change between timestamps is is random in the original process, we replace it with the mean value (expected value of the increase) to get the equation

$$\frac{dm_i(t)}{dt} = \frac{m_i(t)}{2t-1} \text{ for } t > i.$$

(c) Solve the above differential equation for $m_i(t)$ using the boundary conditions to determine the constant(s). In this step, make the approximation $\log(2t-1) \approx \log(2t)$ for convenience of computation.

Solution:

$$\frac{dm_i(t)}{dt} = \frac{m_i(t)}{2t-1}$$
$$\frac{1}{m_i(t)} dm_i(t) = \frac{1}{2t-1} dt$$
$$\int \frac{1}{m_i(t)} dm_i(t) = \int \frac{1}{2t-1} dt$$
$$\log m_i(t) = \frac{1}{2}\log(2t-1) + C_1 = \frac{1}{2}\log(2t) + C_1$$
$$m_i(t) = C_2(2t)^{\frac{1}{2}}$$
$$m_i(t) = C\sqrt{2t}.$$

The boundary condition is $m_i(i) = 1$. Then,

$$m_i(i) = C\sqrt{2i} = 1 \Rightarrow C = \frac{1}{\sqrt{2i}}.$$

Our final approximation is

$$m_i(t) = \frac{1}{\sqrt{2i}}\sqrt{2t} = \sqrt{\frac{t}{i}}.$$

(d) What is the market share, $\frac{m_p(t)}{\sum_{j=1}^{t} m_j(t)}$, of the most popular product p? How does this relate to the success of Amazon?

Solution: It's pretty obvious that at a given time t, the expected most popular product is just the first one that was ordered since $m_i(t)$ is inversely proportional to the square root its index i. Plugging in $p = 1$ and total market size being $2t$,

$$\frac{m_1(t)}{\sum_{j=1}^{t} m_j(t)} = \frac{\sqrt{t}}{2t}$$
$$= \frac{1}{2\sqrt{t}}.$$

As time progresses, the most popular product's market share becomes 0. This shows that no matter how dominant one product is, the new items that come later on continue to contribute significant earnings? TBH I'm not entirely sure how this relates to the success of Amazon.