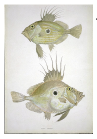
P o C S  What's The Story?

Principles of Complex Systems, Vols. 1, 2, & 3D
CSYS/MATH 300, 303, & 394
University of Vermont, Fall 2022
Solutions to Assignment 06
The Truth Shall Make Ye Fret

Name: Krishna Kannan Srinivasan

Conspirators: Cameron L, Adam O, Cailin G

---

1. More on the peculiar nature of distributions of power law tails:

   Consider a set of N samples, randomly chosen according to the probability distribution $P_k = ck^{-\gamma}$ where $k = 1, 2, 3, \ldots$.

   Estimate $\min k_{\max}$, the approximate minimum of the largest sample in the system, finding how it depends on N.

   (Hint: we expect on the order of 1 of the N samples to have a value of $\min k_{\max}$ or greater.)

   Hint—Some visual help on setting this problem up:
   http://www.youtube.com/watch?v=4tqlEuXA7QQ

   Solution:

   We have the probability distribution,

   $$P_k = ck^{-\gamma}$$

   If $\min k_{\max}$ is the minimum of largest sample from many experiments,

   $$\sum_{k=mink_{max}}^{\infty} P_k = \frac{1}{N}$$

   and

   $$\sum_{k=-\infty}^{mink_{max}} P_k = \frac{N-1}{N}$$

   Approximating sum as an integral,

   $$\int_{x=mink_{max}}^{\infty} P_k \, dx = \frac{1}{N}$$

   $$\int_{x=mink_{max}}^{\infty} cx^{-\gamma} \, dx = \frac{1}{N}$$

$$cx^{-\gamma}\Big|_{mink_{max}}^{\infty} = \frac{1}{N}$$

$$\frac{c}{\gamma - 1}\left[mink_{max}^{-(\gamma-1)}\right] = \frac{1}{N}$$

$$mink_{max} = \left[\frac{c}{\gamma - 1}\right]^{\frac{1}{\gamma-1}} N^{\frac{1}{\gamma-1}}$$

$$mink_{max} \propto N^{\frac{1}{\gamma-1}}$$

□

Notes:

- For language, this scaling is known as Heaps' law (Stigler's Law applies again).
- In a later assignment, we will test this scaling by (thoughtfully) sampling from power-law size distributions.

2. Code up Simon's rich-gets-richer model.

   Show Zipf distributions for $\rho = 0.10$, 0.01, and 0.001. and perform regressions to test $\alpha = 1 - \rho$.

   Run the simulation for long enough to produce decent scaling laws (recall: three orders of magnitude is good).

   Averaging over simulations will produce cleaner results so try 10 and then, if possible, 100.

   Note the first mover advantage.

   Solution:

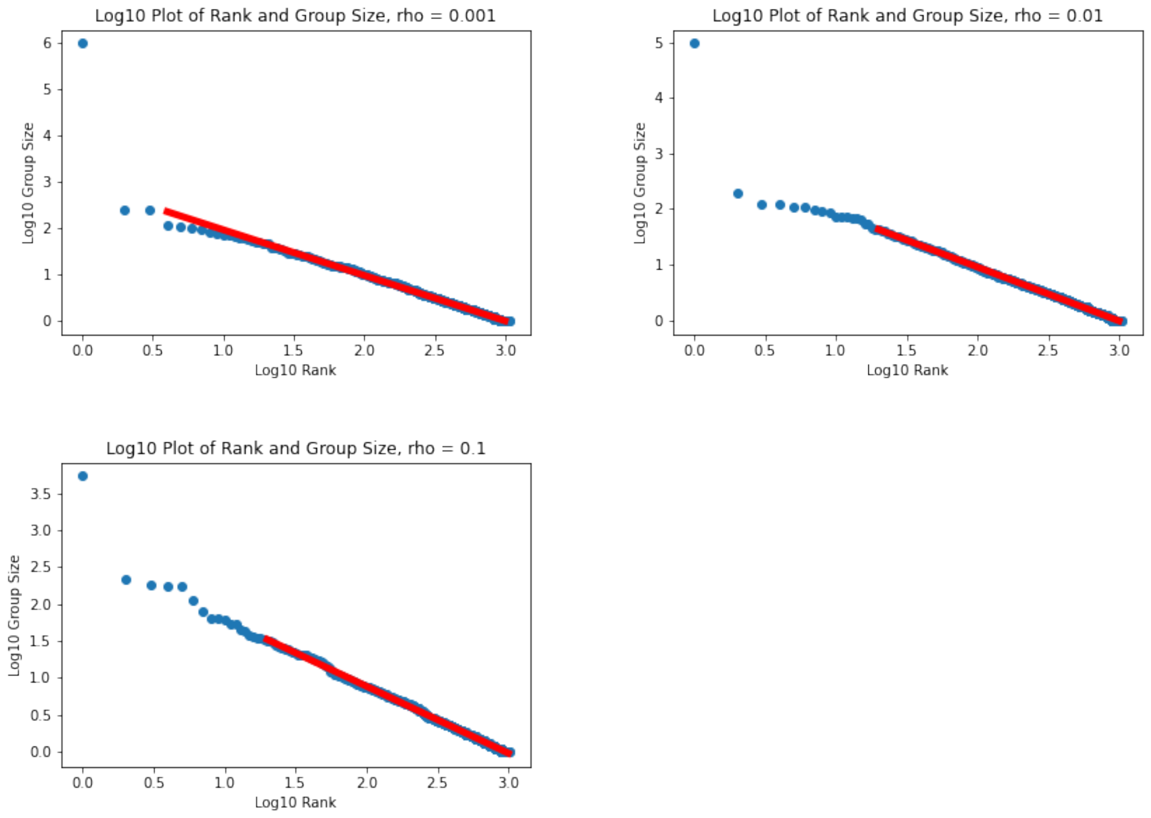


Figure 1: Rank vs Group Size for different innovation rates

(a) $\rho = 0.001$, from regression, $\alpha_{reg} = 0.99$, $\alpha_{est} = 1 - \rho = 0.999$ which agrees

(b) $\rho = 0.01$, from regression, $\alpha_{reg} = 0.97$, $\alpha_{est} = 1 - \rho = 0.99$ which agrees

(c) $\rho = 0.1$,from regression, $\alpha_{reg} = 0.91$, $\alpha_{est} = 1 - \rho = 0.9$ which agrees

□

3. (3 + 3 + 3 points) For Herbert Simon's model of what we've called Random Competitive Replication, we found in class that the normalized number of groups in the long time limit, n_k, satisfies the following difference equation:

$$\frac{n_k}{n_{k-1}} = \frac{(k-1)(1-\rho)}{1+(1-\rho)k} \tag{1}$$

where $k \geq 2$. The model parameter ρ is the probability that a newly arriving node forms a group of its own (or is a novel word, starts a new city, has a unique flavor, etc.). For $k = 1$, we have instead

$$n_1 = \rho - (1-\rho)n_1 \tag{2}$$

which directly gives us n_1 in terms of ρ.

(a) Derive the exact solution for n_k in terms of gamma functions and ultimately the beta function.

(b) From this exact form, determine the large k behavior for n_k ($\sim k^{-\gamma}$) and identify the exponent γ in terms of ρ. You are welcome to use the fact that $B(x,y) \sim x^{-y}$ for large x and fixed y (use Stirling's approximation or possibly Wikipedia).

Note: Simon's own calculation is slightly awry. The end result is good however.

Hint—Setting up Simon's model:
http://www.youtube.com/watch?v=OTzI5J5W1K0

Solution:

(a)
We have,

$$\frac{n_k}{n_{k-1}} = \frac{(k-1)(1-\rho)}{1+(1-\rho)k}$$

$$n_k = \frac{(k-1)(1-\rho)}{1+(1-\rho)k}\frac{(k-2)(1-\rho)}{1+(1-\rho)(k-1)}\frac{(k-3)(1-\rho)}{1+(1-\rho)(k-2)}\cdots\frac{(1)(1-\rho)}{1+(1-\rho)2}n_1$$

$$n_k = \frac{(k-1)!(1-\rho)^{k-1}}{(1-\rho)^{k-1}\left[\frac{1}{1-\rho}+k\right]\left[\frac{1}{1-\rho}+k-1\right]\ldots\left[\frac{1}{1-\rho}+2\right]}n_1$$
$$n_k = \frac{(k-1)!}{\left[\frac{1}{1-\rho}+k\right]\left[\frac{1}{1-\rho}+k-1\right]\ldots\left[\frac{1}{1-\rho}+2\right]}n_1$$

The Gamma function for integers is defined by,

$$\Gamma(n+1) = n!$$

Then,

$$\Gamma(k) = (k-1)!$$

We also know that,

$$n(n-1)\ldots 2 = \frac{n!}{2!}$$
$$n(n-1)\ldots 2 = \frac{\Gamma(n+1)}{\Gamma(2+1)}$$

Then,

$$\left[\frac{1}{1-\rho}+k\right]\left[\frac{1}{1-\rho}+k-1\right]\ldots\left[\frac{1}{1-\rho}+2\right] = \frac{\Gamma(1+\frac{1}{1-\rho}+k)}{\Gamma(1+\frac{1}{1-\rho}+2})$$

with $\Gamma(1) = 1$. Writing in terms of Gamma functions,

$$n_k = \frac{\Gamma(k)\Gamma(1 + \frac{1}{1-\rho} + 2)}{\Gamma(1 + \frac{1}{1-\rho} + k)} n_1$$

Also for rational numbers,

$$\Gamma(x + 1) = x\Gamma(x)$$

Then,

$$\Gamma(1 + \frac{1}{1-\rho} + k) = \left[\frac{1}{1-\rho} + k\right] \Gamma(\frac{1}{1-\rho} + k)$$

$$\Gamma(1 + \frac{1}{1-\rho} + 2) = \left[\frac{1}{1-\rho} + 1\right] \Gamma(\frac{1}{1-\rho} + 1)$$

Substituting we have,

$$n_k = \frac{\Gamma(k)\left[\frac{1}{1-\rho} + 1\right] \Gamma(\frac{1}{1-\rho} + 1)}{\Gamma(1 + \frac{1}{1-\rho} + k)} n_1$$

Using the relationship between Beta Functions and Gamma Functions,

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x + y)}$$

,

$$B(k, 1 + \frac{1}{1-\rho}) = \frac{\Gamma(k)\Gamma(1 + \frac{1}{1-\rho})}{\Gamma(1 + \frac{1}{1-\rho} + k)}$$

$$n_k = \left[\frac{1}{1-\rho} + 1\right] B(k, 1 + \frac{1}{1-\rho}) n_1$$

Simplifying,

$$n_k = \left[\frac{2-\rho}{1-\rho}\right] B(k, 1 + \frac{1}{1-\rho}) n_1$$

And since,

$$n_1 = \rho - (1 - \rho)n_1$$

$$n_1 = \frac{\rho}{1-\rho}$$

Substituting we have the final expression for n_k in terms of Beta functions and innovation rate ρ.

$$n_k = \left[\frac{\rho}{1-\rho}\right] B(k, 1 + \frac{1}{1-\rho})$$

(b) Large k behavior and γ in terms of ρ.
We have,

$$n_k = \left[\frac{\rho}{1-\rho}\right] B(k, 1 + \frac{1}{1-\rho})$$

Using approximation for Beta function from wikipedia,

$$B(x, y) = \Gamma(y) x^{-y}$$

For Large x and fixed y,

$$n_k = \left[\frac{\rho}{1-\rho}\right] \Gamma(1 + \frac{1}{1-\rho}) k^{-(1+\frac{1}{1-\rho})}$$

Comparing with

$$n_k = ck$$

,

$$c = \frac{\rho}{1-\rho}\Gamma(1 + \frac{1}{1-\rho})$$

and,

$$\gamma = 1 + \frac{1}{1-\rho}$$

with $\rho \neq 1$ since $c \to \infty$

For large k, it looks like $n_k \to 0$.

□

4. What happens to γ in the limits $\rho \to 0$ and $\rho \to 1$? Explain in a sentence or two what's going on in these cases and how the specific limiting value of γ makes sense.

Solution: We have,

$$\gamma = 1 + \frac{1}{1-\rho}$$

(a) When $\rho \to 0$,$\gamma \to 2$ This means the power law distribution will have infinite mean and variance. This is analogous to having a single large category.

(b) When $\rho \to 1$, $\gamma \to \infty$, This is not a power law distribution, Infinite categories of size 1. Hence the mean and variance are finite ($\gamma > 3$)

□

5. (6 + 3 + 3 points)

In Simon's original model, the expected total number of distinct groups at time t is ρt. Recall that each group is made up of elements of a particular flavor.

In class, we derived the fraction of groups containing only 1 element, finding

$$n_1^{(g)} = \frac{N_1(t)}{\rho t} = \frac{1}{2-\rho}$$

(a) (3 + 3 points)

Find the form of $n_2^{(g)}$ and $n_3^{(g)}$, the fraction of groups that are of size 2 and size 3.

Solution: We have,

$$n_k = \frac{(k-1)(1-\rho)}{1+(1-\rho)k}\frac{(k-2)(1-\rho)}{1+(1-\rho)(k-1)}\frac{(k-3)(1-\rho)}{1+(1-\rho)(k-2)}\cdots\frac{(1)(1-\rho)}{1+(1-\rho)2}n_1$$

$$n_k = \frac{(k-1)}{\frac{1}{1+(1-\rho)}+k}\frac{(k-2)}{\frac{1}{1+(1-\rho)}+k-1}\cdots\frac{(2)}{\frac{1}{1+(1-\rho)}+3}\frac{(1)}{\frac{1}{1+(1-\rho)}+1}n_1$$

Also,

$$n_1^{(g)} = \frac{N_1(t)}{\rho t} = \frac{1}{2-\rho}$$

$$n_1 = \rho N_1(t) = \frac{\rho}{2-\rho}$$

Since $N_k(t) = n_k t$,

$$n_2^{(g)} = \frac{N_2(t)}{\rho t} = \frac{n_2}{\rho}$$
$$n_2^{(g)} = \frac{1}{\rho}\frac{1}{\frac{1}{1-\rho}+2}\frac{\rho}{2-\rho} \quad (3)$$
$$n_2^{(g)} = \frac{1-\rho}{(2-\rho)(3-2\rho)}$$

$$n_3^{(g)} = \frac{N_3(t)}{\rho t} = \frac{n_3}{\rho}$$
$$n_3^{(g)} = \frac{1}{\rho}\frac{2}{\frac{1}{1-\rho}+3}\frac{1}{\frac{1}{1-\rho}+2}\frac{\rho}{2-\rho} \tag{4}$$
$$n_3^{(g)} = \frac{2(1-\rho)^2}{(2-\rho)(3-2\rho)(4-3\rho)}$$

□

(b) Using data for James Joyce's Ulysses (see below), first show that Simon's estimate for the innovation rate $\rho_{\text{est}} \simeq 0.115$ is reasonably accurate for the version of the text's word counts given below.

Hint: You should find a slightly higher number than Simon did.

Hint: Do not compute ρ_{est} from an estimate of γ.

Solution:

```
with open("ulysses.txt", 'r') as f:
    strings = f.read().split()
strings = [s.replace(":", "") for s in strings]
ulysses = {}
for k in range(0, len(strings)-2,2):
    key = strings[k]
    value = int(strings[k+1])
    ulysses[key] = value
num_unique_words = len(list(ulysses.keys()))
num_all_words = sum(list(ulysses.values()))
rho_est = num_unique_words/num_all_words
print("rho_est = ", rho_est)
```

We get $\rho_{est} = 0.119$ which matches up with 0.115.

□

(c) Now compare the theoretical estimates for $n_1^{(g)}$, $n_2^{(g)}$, and $n_3^{(g)}$, with empirical values you obtain for Ulysses.

Solution:

i. From the estimated ρ_{est} for the dataset, awe have $n_1^g = 0.53$, $n_2^g = 0.17$, $n_3^g = 0.082$

ii. From the data, $n_{1-data}^g = 0.564$ $n_{2-data}^g = 0.15$ $n_{3-data}^g = 0.07$

□

The data (links are clickable):

- Matlab file (`sortedcounts` = word frequency f in descending order, `sortedwords` = ranked words):
  https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse/docs/ulysses.mat
- Colon-separated text file (first column = word, second column = word frequency f):
  https://pdodds.w3.uvm.edu/teaching/courses/2022-2023pocsverse/docs/ulysses.txt

Data taken from http://www.doc.ic.ac.uk/~rac101/concord/texts/ulysses/.

Note that some matching words with differing capitalization are recorded as separate words.

6. (3 + 3)

   Repeat the preceding data analysis for Ulysses for Jane Austen's "Pride and Prejudice" and Alexandre Dumas' "Le comte de Monte-Cristo" (in the original French), working this time from the original texts.

   For each text, measure the fraction of words that appear only once, twice, and three times, and compare them with the theoretical values offered by Simon's model.

   Download text (UTF-8) versions from https://www.gutenberg.org:

   - Pride and Prejudice: https://www.gutenberg.org/ebooks/42671.
   - Le comte de Monte-Cristo: https://www.gutenberg.org/ebooks/17989.

   You will need to parse and count words using your favorite/most-hated language (Python, R, Perl-ha-ha, etc.).

   Gutenberg adds some (non-uniform) boilerplate to the beginning and ends of texts, and you should remove that first. Easiest to do so by inspection for just two texts.

   For a curated version of Gutenberg, see this paper by Gerlach and Font-Clos: https://arxiv.org/abs/1812.08092.

   Solution:

```
Pride and Prejudice

num all words =  111753
num_unique_words =  6843
rho_est =  0.06123325548307428
n1est =  0.5157918057074812
n2est =  0.16827195795906952
n3est =  0.08278600136223944
num_1_words =  3013
num_2_words =  947
num_3_words =  592
n1_g_data =  0.44030396025135177
n2_g_data =  0.13838959520678065
n3_g_data =  0.08651176384626626


Monte Cristo

num all words =  124471
num_unique_words =  13501
rho_est =  0.10846703248146154
n1est =  0.5286717266746234
n2est =  0.16935576961783247
n3est =  0.0821783578875855
num_1_words =  7166
num_2_words =  2015
num_3_words =  1046
n1_g_data =  0.530775498111251
n2_g_data =  0.14924820383675283
n3_g_data =  0.0774757425375898
```

$\square$