PoCS What's The Story?

Principles of Complex Systems, CSYS/MATH 300
University of Vermont, Fall 2019
Assignment 4 • code name: A Fistful of Paintballs

Dispersed: Thursday, September 19, 2019.
Due: Friday, September 27, by 11:59 pm, 2019.
Some useful reminders:
Deliverator: Dr. Nick Allgaier (nicholas.allgaier@uvm.edu)
Assistant Deliverator: David Dewhurst (david.dewhurst@uvm.edu)
Office: To be disclosed
Office hours: To be emphatically determined through a democratic process
Course website: http://www.uvm.edu/pdodds/teaching/courses/2019-08UVM-300
Bonus course notes: http://www.uvm.edu/pdodds/teaching/courses/2019-08UVM-300/docs/dewhurst-pocs-notes.pdf

---

All parts are worth 3 points unless marked otherwise. Please show all your workingses clearly and list the names of others with whom you collaborated.

Please obey the basic life rule: Never use Excel. Or any Microsoft product except maybe Xbox (which sadly will likely not help you here.)

Graduate students are requested to use LaTeX (or related TeX variant).

email submission: 1. Please send to david.dewhurst@uvm.edu.
2. PDF only! Please name your file as follows (where the number is to be padded by a 0 if less than 10 and names are all lowercase): `CSYS300assignment%02d$firstname-$lastname.pdf` as in `CSYS300assignment06michael-palin.pdf`

Please submit your project's current draft in pdf format via email. Please use this file name format (all lowercase after CSYS):
`CSYS300project-$firstname-$lastname-YYYY-MM-DD.pdf` as in
`CSYS300project-lisa-simpson-1989-12-17.pdf` where the date is the date of submission (and not, say, your birthdate).

---

1. 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.

2. (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:

Direct link: `http://www.youtube.com/watch?v=OTzI5J5W1K0`

The hint's output including the bits not in the video:

PoCS 2013-09-23

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

$$n_k = \left[\frac{(k-1)(1-\rho)}{1+(1-\rho)k}\right] \overbrace{\left[\frac{(k-2)(1-\rho)}{1+(1-\rho)(k-1)}\right] n_{k-2}}^{n_{k-1}} \cdots$$

$$\left[\frac{(k-3)(1-\rho)}{1+(1-\rho)(k-2)}\right] n_{k-3}$$

$$\cdots\cdots \left[\frac{\star\times(1-\rho)}{1+(1-\rho)\star}\right] n_1$$

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

$$x = n+1 \qquad \Gamma(n+1) = n\,\Gamma(n) = \cdots = n!$$

example $0 < z < 1$

$$(1+zk)(1+z(k-1))\cdots\cdots(1+z1)$$

$$= z^k\left(\frac{1}{z}+k\right)\left(\frac{1}{z}+k-1\right)\cdots\left(\frac{1}{z}+1\right) = z^k\,\frac{\left(\frac{1}{z}+k\right)\left(\frac{1}{z}+k-1\right)\cdots}{\frac{1}{z}\cdot\left(\frac{1}{z}-1\right)\left(\frac{1}{z}-2\right)\cdots}$$

differ by 1

$$= z^k\,\frac{\Gamma\left(\frac{1}{z}+k+1\right)}{\Gamma\left(\frac{1}{z}+1\right)}.$$

3. 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.

4. (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.

   (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 $\rho_{\rm est}$ from an estimate of γ.

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

The data (links are clickable):

- Matlab file (`sortedcounts` = word frequency f in descending order, `sortedwords` = ranked words):
  http://www.uvm.edu/pdodds/teaching/courses/2019-08UVM-300/docs/ulysses.mat
- Colon-separated text file (first column = word, second column = word frequency f):
  http://www.uvm.edu/pdodds/teaching/courses/2019-08UVM-300/docs/ulysses.txt

Data taken from `http://www.doc.ic.ac.uk/~rac101/concord/texts/ulysses/`http://www.doc.ic.ac.uk/∼rac101/concord/texts/ulysses/. Note that some matching words with differing capitalization are recorded as separate words.

5. (3 + 3)

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 $k \geq 1$ and $2 < \gamma < 3$. (Note that k is discrete rather than continuous.)

(a) Estimate $\min k_{\rm max}$, the approximate minimum of the largest sample in the network, finding how it depends on N.

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

Hint—Some visual help on setting this problem up:

Direct link: `http://www.youtube.com/watch?v=4tqlEuXA7QQ`

(b) Determine the average value of samples with value $k \geq \min k_{\max}$ to find how the expected value of $k_{\max}$ (i.e., $\langle k_{\max} \rangle$) scales with N.

For language, this scaling is known as Heap's law.

6. (3 + 3)

Let's see how well your answer for the previous question works.

For $\gamma = 5/2$, generate $n = 1000$ sets each of $N = 10$, 10^2, 10^3, 10^4, 10^5, and 10^6 samples, using $P_k = ck^{-5/2}$ with $k = 1, 2, 3, \ldots$

How do we computationally sample from a discrete probability distribution?

Hint: You can use a continuum approximation to speed things up. In fact, taking the exact continuum version from the first two assignments will work.

(a) For each value of sample size N, plot the maximum value of the $n = 1000$ samples as a function of sample number (which is not the sample size N). So you should have $k_{\max}$ for $i = 1, 2, \ldots, n$ where i is sample number. These plots should give a sense of the unevenness of the maximum value of k, a feature of power-law size distributions.

(b) For each set, find the maximum value. Then find the average maximum value for each N. Plot $\langle k_{\max} \rangle$ as a function of N and calculate the scaling using least squares.

Does your scaling match up with your theoretical estimate?

How to sample from your power law distribution (and kinds of beasts):

We now turn our problem of randomly selecting from this distribution into randomly selecting from the uniform distribution. After playing around a little, $k = 10^6$ seems like a good upper limit for the number of samples we're talking about.

Using Matlab (or some ghastly alternative), we create a cdf for P_k for $k = 1, 2, \ldots, 10^6$ and one final entry $k > 10^6$ (for which the cdf will be 1).

We generate a random number x and find the value of k for which the cdf is the first to meet or exceed x. This gives us our sample k according to P_k and we repeat as needed. We would use the exactly normalized $P_k = \frac{1}{\zeta(5/2)} k^{-5/2}$ where ζ is the Riemann zeta function.

Now, we can use a quick and dirty method by approximating P_k with a continuous function $P(z) = (\gamma - 1)z^{-\gamma}$ for $z \geq 1$ (we have used the normalization coefficient found in assignment 1 for $a = 1$ and $b = \infty$). Writing $F(z)$ as the cdf for $P(z)$, we have $F(z) = 1 - z^{-(\gamma-1)} = 1 - z^{-3/2}$. Inverting, we obtain $z = [1 - F(z)]^{-2/3}$. We replace $F(z)$ with our random number x and round the value of z to finally get an estimate of k.