PoCS What's The Story?

Principles of Complex Systems, CSYS/MATH 300
University of Vermont, Fall 2017
Assignment 4 • code name: Only two hours from the beach

Dispersed: Friday, September 22, 2017.
Due: By 11:59 pm, Friday, September 29, 2017.
Some useful reminders:
Deliverator: Peter Dodds
Office: Farrell Hall, second floor, Trinity Campus
E-mail: peter.dodds+pocs@uvm.edu
Office hours: 1:15 pm to 2:30 pm on Tuesday, 1:15 pm to 4:45 pm Thursday
Course website: http://www.uvm.edu/pdodds/teaching/courses/2017-08UVM-300
Bonus course notes: http://www.uvm.edu/pdodds/teaching/courses/2017-08UVM-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.

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

Email submission: 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.

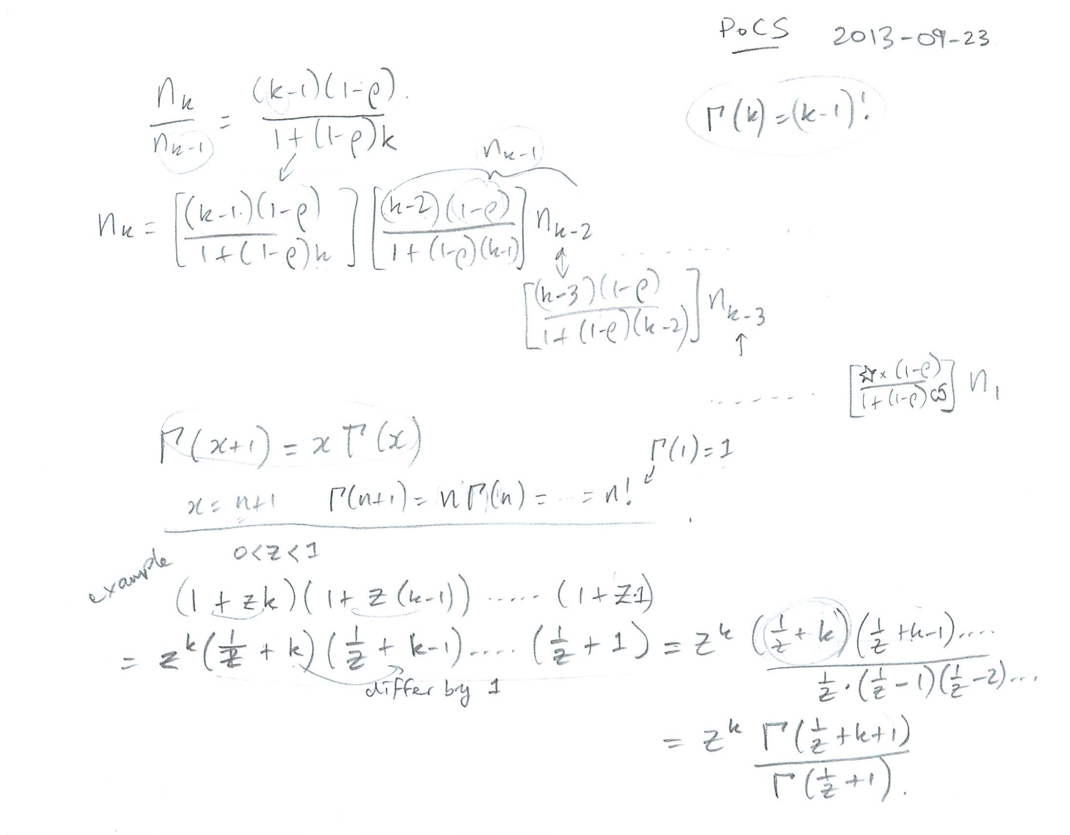
(b) From this exact form, determine the large k behavior for n_k ($\sim k^{-\gamma}$) and identify the exponent γ in terms of ρ. You are welcome to use the fact that $B(x,y) \sim x^{-y}$ for large x and fixed y (use Stirling's approximation or possibly Wikipedia).

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

Hint—Setting up Simon's model:

`http://www.youtube.com/watch?v=OTzI5J5W1KO`

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



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 ρ_{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/2017-08UVM-300/docs/ulysses.mat
- Colon-separated text file (first column = word, second column = word frequency f): http://www.uvm.edu/pdodds/teaching/courses/2017-08UVM-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.