P o C S | What's The Story? | **Principles of Complex Systems, Vols. 1 & 2, CSYS/MATH 300 and 303**
University of Vermont, Fall 2021
Assignment 06 • code name: Guest Starring John Noble ↗

Due: Friday, October 8, by 11:59 pm, 2021.

Relevant clips, episodes, and slides are listed on the assignment's page:

https://pdodds.w3.uvm.edu//teaching/courses/2021-2022principles-of-complex-systems//assignments/06/

Some useful reminders:

Deliverator: Prof. Peter Sheridan Dodds (contact through Teams)

Assistant Deliverator: Michael Arnold (contact through Teams)

Office: The Ether

Office hours: TBD

Course website:

https://pdodds.w3.uvm.edu//teaching/courses/2021-2022principles-of-complex-systems

---

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.

For coding, we recommend you improve your skills with Python, R, and/or Julia. The Deliverator uses Matlab.

Graduate students are requested to use LaTeX (or related TeX variant). If you are new to LaTeX, please endeavor to submit at least n questions per assignment in LaTeX, where n is the assignment number.

Assignment submission: Via Blackboard.

---

1. $(3 + 3)$

   Allotaxonometry.

   Rank-turbulence divergence (RTD) is defined as:

   $$D_\alpha^{\mathrm{R}}(R_1 \parallel R_2) = \sum_{\tau \in R_{1,2;\alpha}} \delta D_{\alpha,\tau}^{\mathrm{R}}(R_1 \parallel R_2)$$

   $$= \frac{1}{\mathcal{N}_{1,2;\alpha}} \frac{\alpha+1}{\alpha} \sum_{\tau \in R_{1,2;\alpha}} \left| \frac{1}{[r_{\tau,1}]^\alpha} - \frac{1}{[r_{\tau,2}]^\alpha} \right|^{1/(\alpha+1)}. \qquad (1)$$

   Find the limits of RTD for:

   (a) $\alpha \to 0$.

(b) $\alpha \to \infty$.

Leave $\frac{1}{\mathcal{N}_{1,2;\alpha}}$ as a constant.

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.

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:

$$\frac{n_k}{n_{k-1}} = \frac{(k-1)(1-\rho)}{1 + (1-\rho)k} \tag{2}$$

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

$$n_1 = \rho - (1-\rho)n_1 \tag{3}$$

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

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

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

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

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

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

$$\cdots\cdots \left[\frac{\star \times (1-\rho)}{1+(1-\rho)\cdot 2}\right]n_1$$

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

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

example $0 < z < 1$

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

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.

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.

   (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.

3

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 ($\texttt{sortedcounts}$ = word frequency f in descending order, $\texttt{sortedwords}$ = ranked words):
  https://pdodds.w3.uvm.edu//teaching/courses/2021-2022principles-of-complex-systems/docs/ulysses.mat

- Colon-separated text file (first column = word, second column = word frequency f):
  https://pdodds.w3.uvm.edu//teaching/courses/2021-2022principles-of-complex-systems/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 analyses 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.

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 ⬀.