PoCS What's The Story?

# Principles of Complex Systems, Vols. 1, 2, & 3D
# CSYS/MATH 6701, 6713, & a pretend number
# University of Vermont, Fall 2023
# Assignment 07
## "Cool, cool, cool"

**Due:** Wednesday, October 25, by 11:59 pm
https://pdodds.w3.uvm.edu/teaching/courses/2023-2024pocsverse/assignments/07/
*Some useful reminders:*
**Deliverator:** Prof. Peter Sheridan Dodds (contact through Teams)
**Assistant Deliverator:** Chris O'Neil (contact through Teams)
**Office:** The Ether
**Office hours:** See Teams calendar
**Course website:** https://pdodds.w3.uvm.edu/teaching/courses/2023-2024pocsverse
**Overleaf:** LaTeX templates and settings for all assignments are available at https://www.overleaf.com/read/tsxfwwmwdgxj.

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

For coding, we recommend you improve your skills with Python, R, and/or Julia. The (evil) Deliverator uses (evil) 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 Brightspace or other preferred death vortex.

**Please submit your project's current draft** in pdf format via Brightspace by the same time specified for this assignment. For teams, please list all team member names clearly at the start.

See assignment 9 for instructions including details for the first presentation.

1. (6 + 3 + 3 points)

   In Simon's original model, the expected total number of distinct groups at time $t$ is $\rho t$. 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 $\rho_{\rm est} \simeq 0.115$ 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 $\gamma$.

(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): https://pdodds.w3.uvm.edu/teaching/courses/2023-2024pocsverse/docs/ulysses.mat
- Colon-separated text file (first column = word, second column = word frequency $f$): https://pdodds.w3.uvm.edu/teaching/courses/2023-2024pocsverse/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.

2. (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 ↗.