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 30
"The Darkest Timeline"

Due: Never
https://pdodds.w3.uvm.edu/teaching/courses/2023-2024pocsverse/assignments/30/
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.

The questions you don't have to do!

Some are open ended madnesses.

1. (3 + 3 points) *Zipfarama via Optimization:*

   Complete the Mandelbrotian derivation of Zipf's law by minimizing the function

$$\Psi(p_1, p_2, \ldots, p_n) = F(p_1, p_2, \ldots, p_n) + \lambda G(p_1, p_2, \ldots, p_n)$$

   where the 'cost over information' function is

$$F(p_1, p_2, \ldots, p_n) = \frac{C}{H} = \frac{\sum_{i=1}^{n} p_i \ln(i + a)}{-g \sum_{i=1}^{n} p_i \ln p_i}$$

and the constraint function is

$$G(p_1, p_2, \ldots, p_n) = \sum_{i=1}^{n} p_i - 1 \quad (= 0)$$

to find

$$p_j = e^{-1-\lambda H^2/gC}(j+a)^{-H/gC}.$$

Then use the constraint equation, $\sum_{j=1}^{n} p_j = 1$ to show that

$$p_j = (j+a)^{-\alpha}.$$

where $\alpha = H/gC$.

3 points: When finding λ, find an expression connecting λ, g, C, and H.

The Perishing Monks who have returned say the way is sneaky. Before collapsing, one monk mumbled something about substituting the form you find for $\ln p_i$ into H's definition (but do not replace p_i).

Note: We have now allowed the cost factor to be $(j+a)$ rather than $(j+1)$.

2. Carrying on from the previous problem:

   For $n \to \infty$, use some computation tool (e.g., Matlab, an abacus, but not a clever friend who's really into computers) to determine that $\alpha \simeq 1.73$ for $a = 1$. (Recall: we expect $\alpha < 1$ for $\gamma > 2$)

3. Do not do! Solution provided. For finite n, find an approximate estimate of a in terms of n that yields $\alpha = 1$.

   (Hint: use an integral approximation for the relevant sum.)

   What happens to a as $n \to \infty$?

   Solution:

   For finite n and a burning desire that $\alpha = 1$, we can approximate the above sum with an integral

$$1 \simeq \int_{x=1}^{n} (x+a)^{-1}.$$

   Barging ahead:

$$\begin{aligned} 1 &\simeq \ln(x+a)\big|_1^n \\ &= [\ln(n+a) - \ln(1+a)] \\ &= \ln\frac{(n+a)}{(1+a)}. \end{aligned}$$

Next, we isolate a:

$$\begin{aligned} e &\simeq \frac{(n+a)}{(1+a)} \\ &= \frac{(1+a)}{(1+a)} + \frac{(n-1)}{(1+a)} \\ &= 1 + \frac{(n-1)}{(1+a)} \end{aligned}$$

Some twists and turns give:

$$a \simeq \frac{n-1}{e-1} - 1 = \frac{n-e}{e-1} \sim \frac{1}{e-1} n.$$

So we see that a grows linearly with n and that that $\alpha = 1$ is an impossibility in the $n = \infty$ limit.

Simon's model also has the $\alpha = 1$ case in a peculiar limit (no new arrivals). While Mandelbrot's model is perhaps the least realistic in the mechanism detail, the broader story of optimization remains plausible. There is much subsequent work over the next fifty years that attempts to improve upon both Mandelbrot and Simon.

4. The 1-d theoretical percolation problem:

   Consider an infinite 1-d lattice forest with a tree present at any site with probability p.

   (a) Find the distribution of forest sizes as a function of p. Do this by moving along the 1-d world and figuring out the probability that any forest you enter will extend for a total length ℓ.

   (b) Find p_c, the critical probability for which a giant component exists.

   Hint: One way to find critical points is to determine when certain average quantities explode. Compute $\langle l \rangle$ and find p such that this expression goes boom (if it does).

5. Show analytically that the critical probability for site percolation on a triangular lattice is $p_c = 1/2$.

   Hint—Real-space renormalization gets it done.:
   http://www.youtube.com/watch?v=JlkbU5U7QqU

6. (3 + 3)

   Coding, it's what's for breakfast:

(a) Percolation in two dimensions (2-d) on a simple square lattice provides a classic, nutritious example of a phase transition.

Your mission, whether or not you choose to accept it, is to code up and analyse the L by L square lattice percolation model for varying L.

Take $L =$ 20, 50, 100, 200, 500, and 1000.

(Go higher if you feel $L =$ 1000 is for mere mortals.)

(Go lower if your code explodes.)

Let's continue with the tree obsession. A site has a tree with probability p, and a sheep grazing on what's left of a tree with probability $1 - p$.

Forests are defined as any connected component of trees bordered by sheep, where connections are possible with a site's four nearest neighbors on a lattice.

Each square lattice is to be considered as a landscape on which forests and sheep co-exist.

Do not bagelize (or doughnutize) the landscape (no periodic boundary conditions—boundaries are boundaries).

(Note: this set up is called site percolation. Bond percolation is the alternate case when all links between neighboring sites exist with probability p.)

Steps:

i. For each L, run N_{tests}=100 tests for occupation probability p moving from 0 to 1 in increments of 10^{-2}. (As for L, you may use a smaller or larger increment depending on how things go.)

ii. Determine the fractional size of the largest connected forest for each of the N_{tests}, and find the average of these, S_{avg}.

iii. On a single figure, for each L, plot the average S_{avg} as a function of p.

(b) Comment on how $S_{\mathrm{avg}}(p; N)$ changes as a function of L and estimate the critical probability p_c (the percolation threshold).

For the few Matlabbers, a helpful reuse of code (intended for black and white image analysis): You can use Matlab's bwconncomp to find the sizes of components. Very nice.

7. (3 + 3)

(a) Using your model from the previous question and your estimate of p_c, plot the distribution of forest sizes (meaning cluster sizes) for $p \simeq p_c$ for the largest L your code and psychological makeup can withstand. (You can average the distribution over separate simulations.)

Comment on what kind of distribution you find.

(b) Repeat the above for $p = p_c/2$ and $p = p_c + (1 - p_c)/2$, i.e., well below and well above p_c.

Produce plots for both cases, and again, comment on what you find.

8. Show that the Gini coefficient G for our idealized power-law size distribution of wealth is:

$$G = \begin{cases} 1 & \text{if } 1 < \gamma \leq 2, \\ \frac{1}{1+2(\gamma-2)} & \text{if } \gamma > 2. \end{cases} \tag{1}$$

Having developed a sense of what values of γ mean, and because of the simplicity of the relationship between G and γ, we can convert a real-world wealth distribution's value of G to γ for the equivalent idealized power-law size distribution:

$$\begin{array}{rl} \gamma \leq 2 & \text{if } G = 1, \\ \gamma = \frac{1}{2}\left(\frac{1}{G} + 3\right) & \text{if } G < 1. \end{array} \tag{2}$$

For example, what does a Gini coefficient of 1/2 mean for an idealized power law?

Eq. 2 gives $\gamma = 5/2$, which we recognized as coming from the Bad Place of finite mean and infinite variance.

9. (3 + 3 + 3 + 3 + 3)

We take a look at the 80/20 rule, 1 per centers, and similar concepts.

Take x to be the wealth held by an individual in a population of n people, and the number of individuals with wealth between x and $x + \mathrm{d}x$ to be approximately $N(x)\mathrm{d}x$.

Given a power-law size frequency distribution $N(x) = cx^{-\gamma}$ where $x_{\min} \ll x \ll \infty$, determine the value of γ for which the so-called 80/20 rule holds.

In other words, find γ for which the bottom 4/5 of the population holds 1/5 of the overall wealth, and the top 1/5 holds the remaining 4/5.

Note that inherent in our construction of the wealth frequency distribution is that the population is ordered by increasing wealth.

Assume the mean is finite, i.e., $\gamma > 2$.

(a) Determine the total wealth W in the system given $\int_{x_{\min}}^{\infty} \mathrm{d}x N(x) = n$.

(b) Imagine that the bottom $100\,\theta_{\mathrm{pop}}$ percent of the population holds $100\,\theta_{\mathrm{wealth}}$ percent of the wealth.

Show γ depends on θ_{pop} and θ_{wealth} as

$$\gamma = 1 + \frac{\ln \frac{1}{(1-\theta_{\mathrm{pop}})}}{\ln \frac{1}{(1-\theta_{\mathrm{pop}})} - \ln \frac{1}{(1-\theta_{\mathrm{wealth}})}}. \tag{3}$$

(c) Given the above, is every pairing of θ_{pop} and θ_{wealth} possible?

(d) Find γ for the 80/20 requirement ($\theta_{\mathrm{pop}} = 4/5$ and $\theta_{\mathrm{wealth}} = 1/5$).

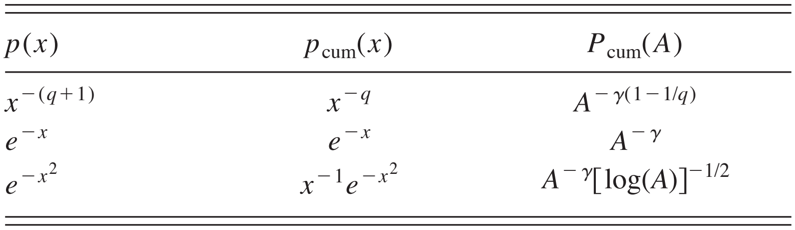
(e) For the "80/20" γ you find, determine the fraction of wealth θ_{wealth} that the bottom fraction θ_{pop} of the population possesses as a function of θ_{pop} and plot the result.

10. (3 + 3 + 3)

Highly Optimized Tolerance:

This question is based on Carlson and Doyle's 1999 paper "Highly optimized tolerance: A mechanism for power laws in design systems" [1]. In class, we made our way through a discrete version of a toy HOT model of forest fires. This paper revolves around the equivalent continuous model's derivation. You do not have to perform the derivation but rather carry out some manipulations of probability distributions using their main formula.

Our interest is in Table I on p. 1415:



$p(x)$	$p_{\mathrm{cum}}(x)$	$P_{\mathrm{cum}}(A)$
$x^{-(q+1)}$	x^{-q}	$A^{-\gamma(1-1/q)}$
e^{-x}	e^{-x}	$A^{-\gamma}$
e^{-x^2}	$x^{-1}e^{-x^2}$	$A^{-\gamma}[\log(A)]^{-1/2}$

and Equation 8 on the same page:

$$P_{\geq}(A) = \int_{p^{-1}(A^{-\gamma})}^{\infty} p(\mathbf{x})\mathrm{d}\mathbf{x} = p_{\geq}\left(p^{-1}\left(A^{-\gamma}\right)\right),$$

where $\gamma = \alpha + 1/\beta$ and we'll write $P_{\geq}$ for P_{cum}.

Please note that $P_{\geq}(A)$ for $x^{-(q+1)}$ is not correct. Find the right one!

Here, $A(\mathbf{x})$ is the area connected to the point $\mathbf{x}$ (think connected patch of trees for forest fires). The cost of a 'failure' (e.g., lightning) beginning at $\mathbf{x}$ scales as $A(\mathbf{x})^{\alpha}$ which in turn occurs with probability $p(\mathbf{x})$. The function p^{-1} is the inverse function of p.

Resources associated with point $\mathbf{x}$ are denoted as $R(\mathbf{x})$ and area is assumed to scale with resource as $A(\mathbf{x}) \sim R^{-\beta}(\mathbf{x})$.

Finally, $p_{\geq}$ is the complementary cumulative distribution function for p.

As per the table, determine $p_{\geq}(x)$ and $P_{\geq}(A)$ for the following (3 pts each):

(a) $p(x) = cx^{-(q+1)}$,

(b) $p(x) = ce^{-x}$, and

(c) $p(x) = ce^{-x^2}$.

Note that these forms are for the tails of p only, and you should incorporate a constant of proportionality c, which is not shown in the paper.

11. The discrete version of HOT theory:

From lectures, we had the following.

Cost: Expected size of 'fire' in a d-dimensional lattice:

$$C_{\text{fire}} \propto \sum_{i=1}^{N_{\text{sites}}} p_i a_i$$

where a_i = area of ith site's region, and p_i = avg. prob. of fire at site i over a given time period.

The constraint for building and maintaining $(d-1)$-dimensional firewalls in d-dimensions is

$$C_{\text{firewalls}} \propto \sum_{i=1}^{N_{\text{sites}}} a_i^{(d-1)/d} a_i^{-1},$$

where we are assuming isometry.

Using Lagrange Multipliers, and, optionally, safety goggles, rubber gloves, a pair of tongs, and a maniacal laugh, determine that:

$$p_i \propto a_i^{-\gamma} = a_i^{-(1+1/d)}.$$

12. (3 + 3 + 3 + 3)

A courageous coding festival:

Code up the discrete HOT model in 2-d. Let's see if we find any of these super-duper power laws everyone keeps talking about. We'll follow the same approach as the $N = L \times L$ 2-d forest discussed in lectures.

Main goal: extract yield curves as a function of the design D parameter as described below.

Suggested simulations elements:

- Take $L = 32$ as a start. Once your code is running, see if $L = 64$, 128, or more might be possible. (The original sets of papers used all three of these values.) Use a value of L that's sufficiently large to produced useful statistics but not prohibitively time consuming for simulations.
- Start with no trees.

- Probability of a spark at the (i, j)th site: $P(i, j) \propto e^{-i/\ell}e^{-j/\ell}$ where (i, j) is tree position with the indices starting in the top left corner ($i, j = 1$ to L). (You will need to normalize this properly.) The quantity ℓ is the characteristic scale for this distribution. Try out $\ell = L/10$.
- Consider a design problem of $D = 1$, 2, L, and L^2. (If L and L^2 are too much, you can drop them. Perhaps sneak out to $D = 3$.) Recall that the design problem is to test D randomly chosen placements of the next tree against the spark distribution.
- For each test tree, compute the average forest fire size over the full spark distribution:

  $$\sum_{i,j} P(i, j)S(i, j),$$

  where $S(i, j)$ is the size of the forest component at (i, j). Select the tree location with the highest average yield and plant a tree there.
- Add trees until the 2-d forest is full, measuring average yield as a function of trees added.
- Only trees within the cluster surrounding the ignited tree burn (trees are connected through four nearest neighbors).

(a) Plot the forest at (approximate) peak yield.

(b) Plot the yield curves for each value of D, and identify (approximately) the peak yield and the density for which peak yield occurs for each value of D.

(c) Plot Zipf (or size) distributions of tree component sizes S at peak yield. Note: You will have to rebuild forests and stop at the peak yield value of D to find these distributions. By recording the sequence of optimal tree planting, this can be done without running the simulation again.

(d) Extra level: Plot Zipf (or size) distributions for $D = L^2$ for varying tree densities $\rho = 0.10, 0.20, \ldots, 0.90$. This will be an effort to reproduce Fig. 3b in [2].

Hint: Working on un-treed locations will make choosing the next location easier.

13. Plot time series for the rank of the following baby names in the US over all years in the census data.

    Do so for raw ranks and $\log_{10}$ ranks.

    - Shirley.
    - Desmond.

- Madison.
- Aiden.
- A name of your choice.

Note that if you plotted relative frequency rather than rank, you would need to know (or estimate) the overall number of babies born. Ranks are both easy simple to work with and easy to understand.

14. **The complex geographies of fairness, greed, belief.**

    Let's start connecting people to places.

    Now: Source census population data as a function of location with corresponding map shape files.

    Goal: We will want to be able to connect density of people in regions with density of specific facilities.

    So the shape files should be as usefully fine in scale as possible. For the census, we have block, block groups, and tracts.

    Please do this collectively by discussing and sharing links/data in the assignments channel on Teams.

    Depending on the software you use, much of this data may be well curated.

15. From lectures on Supply Networks:

    Show that for large V and $0 < \epsilon < 1/2$

    $$\min V_{\text{net}} \propto \int_{\Omega_{d,D}(V)} \rho \, ||\vec{x}||^{1-2\epsilon} \, \mathrm{d}\vec{x} \sim \rho V^{1+\gamma_{\max}(1-2\epsilon)}$$

    Reminders: we defined $L_i = c_i^{-1} V^{\gamma_i}$ where $\gamma_1 + \gamma_2 + \ldots + \gamma_d = 1$, $\gamma_1 = \gamma_{\max} \geq \gamma_2 \geq \ldots \geq \gamma_d$., and $c = \prod_i c_i \leq 1$ is a shape factor.

    Assume the first k lengths scale in the same way with $\gamma_1 = \ldots = \gamma_k = \gamma_{\max}$, and write $||\vec{x}|| = (x_1^2 + x_2^2 + \ldots + x_d^2)^{1/2}$.

16. (3 + 3 points) **Supply networks and allometry:**

    This question's calculation is a specific, exactly-solvable case of the general result that you may attack (with optional relish and other condiments) in a nearby question.

    Consider a set of rectangular areas with side lengths L_1 and L_2 such that $L_1 \propto A^{\gamma_1}$ and $L_2 \propto A^{\gamma_2}$ where A is area and $\gamma_1 + \gamma_2 = 1$. Assume $\gamma_1 > \gamma_2$ and that $\epsilon = 0$.

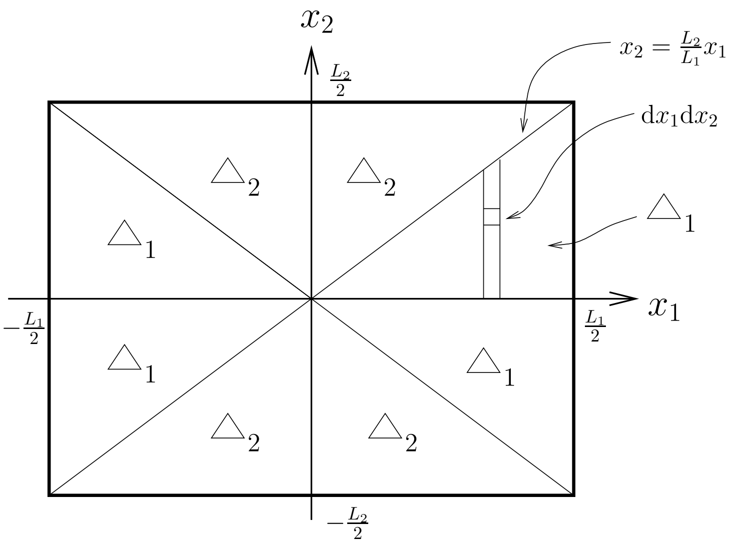
Now imagine that material has to be distributed from a central source in each of these areas to sinks distributed with density $\rho(A)$, and that these sinks draw the same amount of material per unit time independent of L_1 and L_2.

(a) Find an exact form for how the volume of the most efficient distribution network scales with overall area $A = L_1 L_2$. (Hint: you will have to set up a double integration over the rectangle.)

(b) If network volume must remain a constant fraction of overall area, determine the maximal scaling of sink density ρ with A.

Extra hints:

- Integrate over triangles as follows.
- You need to only perform calculations for one triangle.



17. Open:

    Derive a scaling law for the number of side branches that doesn't use stream ordering.

    How many parameters do we need? 3?

18. Come up with a microscopic description of branching river networks that builds from the outlet of the basin rather than the smallest streams.

    For bodies, move from aorta to capillaries.

19. (3 + 3 + 3)

    Estimating the rare:

Google's raw data is for word frequency $k \geq 200$ so let's deal with that issue now.

From Assignment 2, we had for word frequency in the range $200 \leq k \leq 10^7$, a fit for the CCDF of

$$N_{\geq k} \sim 3.46 \times 10^8 k^{-0.661},$$

ignoring errors.

(a) Using the above fit, create a complete hypothetical N_k by expanding N_k back for $k = 1$ to $k = 199$, and plot the result in double-log space (meaning log-log space).

(b) Compute the mean and variance of this reconstructed distribution.

(c) Estimate:

i. The hypothetical total number and fraction of unique words in Google's data set (think at the species or type level now),

ii. The hypothetical fraction of words that appear once out of all words (think of words as organisms or tokens here),

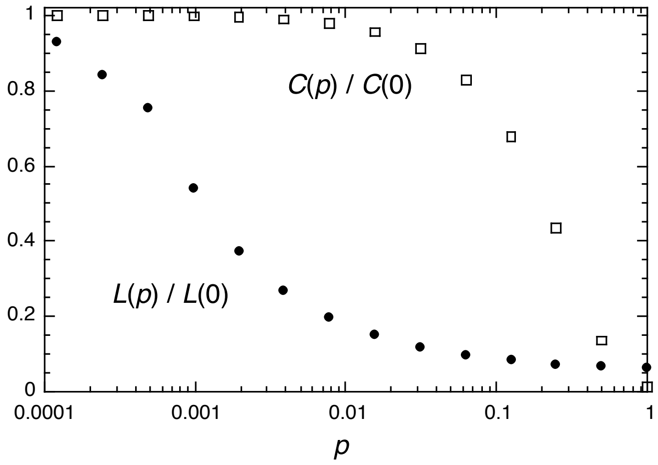
iii. And what fraction of total words are left out of the Google data set by providing only those with counts $k \geq 200$ (back to words as organisms or tokens).

20. Simulate the small-world model and reproduce Fig. 2 from the 1998 Watts-Strogatz paper showing how clustering and average shortest path behave with rewiring probability p [3].

Please find and use any suitable code online, and feel free to share with each other via Slack.

Use $N = 1000$ nodes and $k = 10$ for average degree, and vary p from 0.0001 to 1, evenly spaced on a logarithmic scale (there are only 14 values used in the paper).

Here's the figure you're aiming for:



21. (3 + 3 + 3 + 3 + 3 + 3 pts) **Generalized entropy and diversity:**

For a probability distribution of $i = 1, \ldots, n$ entities with the ith entity having probability of being observed p_i, Shannon's entropy is defined as [4]: $H = -\sum_{i=1}^{n} p_i \ln p_i$. There are other kinds of entropies and we'll explore some aspects of them here.

Let's use the setting of words in a text (another meaningful framing is abundance of species in an ecology). So we have word i appearing with probability p_i and there are n words.

Now, a useful quantity associated with any kind of entropy is diversity, D [5]. Given a text T with entropy H, we define D to be the number of words in another hypothetical text T' which (1) has the same entropy, and (2) where all words appear with equal frequency $1/D$. In text T', we have $p_i = 1/D$ for $i = 1, \ldots, D$.

Diversity is thus a number, and behaves in number-like ways that are more intuitive to grasp than entropy. (Entropy is still the primary thing here.)

Determine the diversity D in terms of the probabilities $\{p_i\}$ for the following:

(a) Simpson concentration:

$$S = \sum_{i=1}^{n} p_i^2.$$

(b) Gini index:

$$G \equiv 1 - S = 1 - \sum_{i=1}^{n} p_i^2.$$

Please note any connections between diversity for the Simpson and Gini indices.

(c) Shannon's entropy:

$$H = -\sum_{i=1}^{n} p_i \ln p_i.$$

(d) Renyi entropy:

$$H_q^{(\mathrm{R})} = \frac{1}{q-1}\left(-\ln \sum_{i=1}^{n} p_i^q\right),$$

where $q \neq 1$.

(e) The generalized Tsallis entropy:

$$H_q^{(\mathrm{T})} = \frac{1}{q-1}\left(1 - \sum_{i=1}^{n} p_i^q\right),$$

where $q \neq 1$.

Please note any connections between diversity for Renyi and Tsallis.

(f) Show that in the limit $q \to 1$, the diversity for the Tsallis entropy matches up with that of Shannon's entropy.

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

23. (3 + 3)

    Allotaxonometry.

    Rank-turbulence divergence (RTD) is defined as:

$$\begin{aligned} D^{\mathrm{R}}_{\alpha}(R_1 \,\|\, R_2) &= \sum_{\tau \in R_{1,2;\alpha}} \delta D^{\mathrm{R}}_{\alpha,\tau}(R_1 \,\|\, R_2) \\ &= \frac{1}{\mathcal{N}_{1,2;\alpha}} \frac{\alpha+1}{\alpha} \sum_{\tau \in R_{1,2;\alpha}} \left| \frac{1}{\left[r_{\tau,1}\right]^{\alpha}} - \frac{1}{\left[r_{\tau,2}\right]^{\alpha}} \right|^{1/(\alpha+1)}. \end{aligned} \tag{4}$$

    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.

24. For finite cutoffs a and b with $a \ll b$, which cutoff dominates the expression for the nth moment as a function of γ and n?

    Note: both cutoffs may be involved to some degree.

25. (a) A parent has two children, not twins, and one is a girl born on a Tuesday. What's the probability that both children are girls?

    See if you can produce both a calculation of probabilities and a visual explanation with shapes (e.g., discs and pie pieces).

    Once you have the answer, can you improve our intuition here? Why does adding the more detailed piece of information of the Tuesday birth change the probability from $1/3$?

    (Assume 50/50 birth probabilities.)

    (b) Same as the previous question but we now know that one is a girl born on December 31. Again, what's the probability that both are girls?

26. Computational Pareidolia.

    A peculiar class project.

    As a team, figure out how to gather, curate, and analyze pictures of the front of cars as they have evolved over time.

Upper limit of insanity: All cars ever sold in the US (types) combined with sales (tokens).

(a) Photos should be from the front.

(b) Ideally, we have photos

(c) Figure out how to assess the emotional content expressed by a car's 'face'.

(d) May be purely computational, may need to use people's assessments. We can use Mechanical Turk for example.

(e) Suggest setting up a single Github repository for the work.

Some articles:

- The faces thing: https://www.smithsonianmag.com/smart-news/for-experts-cars-really-do-have-faces-57005307/.
- Sinisterness: https://www.latimes.com/business/autos/la-hy-sinister-faces-pg-photogallery.html.
- Brain imaging: "High-resolution imaging of expertise reveals reliable object selectivity in the fusiform face area related to perceptual performance" https://www.pnas.org/content/early/2012/09/27/1116333109.abstract.

27. *"Any good idea can be stated in fifty words or less."—Stanisław Ulam.*[1]

    Things have sped up since Ulam made his claim.

    The top of the narrative hierarchy:

    Read through Anderson's seminal paper "More is different" [6] and generate three descriptions of complexification with exactly the following lengths:

    (a) 1–3 words,

    (b) 4–6 words,

    (c) and 7–12 words.

    The 1–3 words one: Try to improve on "More is different".

28. For class discussion, read "Will a large complex system be stable?" by Robert May [7].

    Put together three comments and/or questions.

[1] At the very least, Ulam's claim is self-consistent.

29. (3 + 3 + 3 + 3) This question is all about pure finite and infinite random networks

    We'll define a finite random network as follows. Take N labelled nodes and add links between each pair of nodes with probability p.

    (a) i. For a random node i, determine the probability distribution for its number of friends k, $P_k(p, N)$.

        ii. What kind of distribution is this?

        iii. What does this distribution tend toward in the limit of large N, if p is fixed?

        (No need to do calculations here; just invoke the right Rule of the Universe.)

    (b) Using $P_k(p, N)$, determine the average degree. Does your answer seem right intuitively?

    (c) Show that in the limit of $N \to \infty$ but with mean held constant, we obtain a Poisson degree distribution.

        Hint: to keep the mean constant, you will need to change p.

    (d) i. Compute the clustering coefficients C_1 and C_2 for standard finite random networks (N nodes).

        ii. Explain how your answers make sense.

        iii. What happens in the limit of an infinite random network with finite mean?

30. (3 + 3)

    Determine the clustering coefficient for toy model small-world networks [3] as a function of the rewiring probability p. Find C_1, the average local clustering coefficient:

$$C_1(p) = \left\langle \frac{\sum_{j_1 j_2 \in \mathcal{N}_i} a_{j_1 j_2}}{k_i(k_i - 1)/2} \right\rangle_i = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{j_1 j_2 \in \mathcal{N}_i} a_{j_1 j_2}}{k_i(k_i - 1)/2}$$

    where N is the number of nodes, $a_{ij} = 1$ if nodes i and j are connected, and $\mathcal{N}_i$ indicates the neighborhood of i.

    As per the original model, assume a ring network with each node connected to a fixed, even number m local neighbors ($m/2$ on each side). Take the number of nodes to be $N \gg m$.

    Start by finding $C_1(0)$ and argue for a $(1 - p)^3$ correction factor to find an approximation of $C_1(p)$.

    Hint 1: you can think of finding C_1 as averaging over the possibilities for a single node.

Hint 2: assume that the degree of individual nodes does not change with rewiring but rather stays fixed at m. In other words, take the average degree of individuals as the degree of a randomly selected individual.

For what value of p is $C_1(p)/C_1(0) \simeq 1/2$?

Does this seem reasonable given your simulation?

(3 points for set up, 3 for solving.)

31. (3 + 3):

Consider a modified version of the Barabàsi-Albert (BA) model [8] where two possible mechanisms are now in play. As in the original model, start with m_0 nodes at time $t = 0$. Let's make these initial guys connected such that each has degree 1. The two mechanisms are:

M1: With probability p, a new node of degree 1 is added to the network. At time $t + 1$, a node connects to an existing node j with probability

$$P(\text{connect to node } j) = \frac{k_j}{\sum_{i=1}^{N(t)} k_i} \tag{5}$$

where k_j is the degree of node j and $N(t)$ is the number of nodes in the system at time t.

M2: With probability $q = 1 - p$, a randomly chosen node adds a new edge, connecting to node j with the same preferential attachment probability as above.

Note that in the limit $q = 0$, we retrieve the original BA model (with the difference that we are adding one link at a time rather than m here).

In the long time limit $t \to \infty$, what is the expected form of the degree distribution P_k?

Do we move out of the original model's universality class?

Different analytic approaches are possible including a modification of the BA paper, or a Simon-like one (see also Krapivsky and Redner [9]).

Hint: You can attempt to solve the problem exactly and you'll find an integrating factor story.

Another hint, moment of mercy: Approximate the differential equation by considering large t (this will simplify the denominators).

(3 points for set up, 3 for solving.)

32. (3 + 3)

Using Gleeson and Calahane's iterative equations below, derive the contagion condition for a vanishing seed by taking the limit $\phi_0 \to 0$ and $t \to \infty$. In lectures, we derived the discrete evolution equations for the fraction of infected nodes ϕ_t and the fraction of infected edges θ_t as follows:

$$\phi_{t+1} = \phi_0 + (1 - \phi_0) \sum_{k=0}^{\infty} P_k \sum_{j=0}^{k} \binom{k}{j} \theta_t^j (1 - \theta_t)^{k-j} B_{kj},$$

$$\theta_{t+1} = G(\theta_t; \phi_0) = \phi_0 + (1 - \phi_0) \sum_{k=1}^{\infty} \frac{k P_k}{\langle k \rangle} \sum_{j=0}^{k-1} \binom{k-1}{j} {\theta_t}^j (1 - \theta_t)^{k-1-j} B_{kj},$$

where $\theta_0 = \phi_0$, and B_{kj} is the probability that a degree k node becomes active when j of its neighbors are active.

Recall that by contagion condition, we mean the requirements of a random network for macroscopic spreading to occur.

To connect the paper's model and notation to those of our lectures, given a specific response function F and a threshold model, the B_{kj} are given by $B_{kj} = F(j/k)$.

Allow B_{k0} to be arbitrary (i.e., not necessarily 0 as for simple threshold functions).

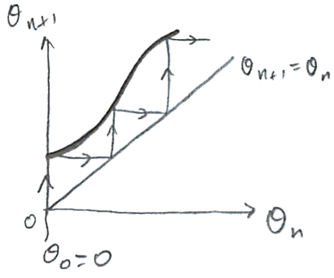
We really only need to understand how θ_t behaves. Write the corresponding equation as $\theta_{t+1} = G(\theta_t; \phi_0)$ and determine when

(a) $G(0; 0) > 0$ (spreading is for free).

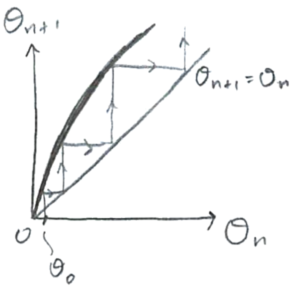
(b) $G(0; 0) = 0$ and $G'(0; \phi_0) > 1$ meaning $\phi = 0$ is a unstable fixed point.

Here's a graphical hint for the three cases you need to consider as $\theta_0 \to 0$:

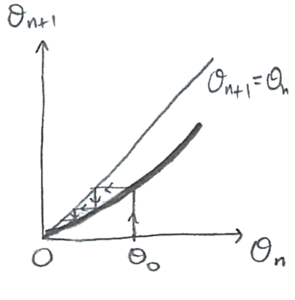
Success:



Sucesss:



Fail:



33. (3 + 3 + 3) Optional:

Solve Krapivsky-Redner's model for the pure linear attachment kernel $A_k = k$.

Starting point:

$$n_k = \frac{1}{2}(k-1)n_{k-1} - \frac{1}{2}kn_k + \delta_{k1}$$

with $n_0 = 0$.

(a) Determine n_1.

(b) Find a recursion relation for n_k in terms of n_{k-1}.

(c) Now find

$$n_k = \frac{4}{k(k+1)(k+2)}$$

for all k and hence determine γ.

34. (3 + 3) Optional:

From lectures:

(a) Starting from the recursion relation

$$n_k = \frac{A_{k-1}}{\mu + A_k} n_{k-1},$$

and $n_1 = \mu/(\mu + A_1)$, show that the expression for n_k for the Krapivsky-Redner model with an asymptotically linear attachment kernel A_k is:

$$\frac{\mu}{A_k}\prod_{j=1}^{k}\frac{1}{1+\frac{\mu}{A_j}}.$$

(b) Now show that if $A_k \to k$ for $k \to \infty$ (or for large k), we obtain $n_k \to k^{-\mu-1}$.

35. (3 + 3 + 3)

From lectures, complete the analysis for the Krapivsky-Redner model with attachment kernel:

$$A_1 = \alpha \text{ and } A_k = k \text{ for } k \geq 2.$$

Find the scaling exponent $\gamma = \mu + 1$ by finding μ. From lectures, we assumed a linear growth in the sum of the attachment kernel weights $\mu t = \sum_{k=1}^{\infty} N_k(t)A_k$, with $\mu = 2$ for the standard kernel $A_k = k$.

We arrived at this expression for μ which you can use as your starting point:

$$1 = \sum_{k=1}^{\infty}\prod_{j=1}^{k}\frac{1}{1+\frac{\mu}{A_j}}$$

(a) Show that the above expression leads to

$$\frac{\mu}{\alpha} = \sum_{k=2}^{\infty} \frac{\Gamma(k+1)\Gamma(2+\mu)}{\Gamma(k+\mu+1)}$$

Hint: you'll want to separate out the $j = 1$ case for which $A_j = \alpha$.

(b) Now use result that [9]

$$\sum_{k=2}^{\infty} \frac{\Gamma(a+k)}{\Gamma(b+k)} = \frac{\Gamma(a+2)}{(b-a-1)\Gamma(b+1)}$$

to find the connection

$$\mu(\mu-1) = 2\alpha,$$

and show this leads to

$$\mu = \frac{1+\sqrt{1+8\alpha}}{2}.$$

(c) Interpret how varying α affects the exponent γ, explaining why $\alpha < 1$ and $\alpha > 1$ lead to the particular values of γ that they do.

36. Yes, even more on power law size distributions. It's good for you.

For the probability distribution $P(x) = cx^{-\gamma}$, $0 < a \leq x \leq b$, compute the mean absolute displacement (MAD), which is given by $\langle |X - \langle X \rangle| \rangle$ where $\langle \cdot \rangle$ represents expected value. As always, simplify your expression as much as possible.

MAD is a more reasonable estimate for the width of a distribution, but we like variance σ^2 because the calculations are much prettier. Really.

37. In the limit of $b \to \infty$, show that MAD asymptotically behave as:

$$\langle |X - \langle X \rangle| \rangle = \frac{2(\gamma-2)^{(\gamma-3)}}{(\gamma-1)^{(\gamma-2)}} a.$$

How does this compare with the behavior of the variance? (See the last question of Assignment todo???.)

38. *Simon's model II:*

A missing piece from the lectures: Obtain γ in terms of ρ by expanding Eq. **??** in terms of $1/k$. In the end, you will need to express n_k/n_{k-1} as $(1 - 1/k)^{\theta}$; from here, you will be able to identify γ. Taylor expansions and Procrustean truncations will be in order.

This (dirty) method avoids finding the exact form for n_k.

39. A spectacularly optional extra.

Warning:

- Only attempt if using registered safety equipment including welding goggles and a lead apron.
- Make sure to back up your brain in at least two geographically distant places beforehand (e.g., on different planets).

Dangerous feature:

- If you make it out, you will be very happy.

In lectures on lognormals and other heavy-tailed distributions, we came across a super fun and interesting integral when considering organization size distributions arising from growth processes with variable lifespans.

Show that

$$P(x) = \int_{t=0}^{\infty} \lambda e^{-\lambda t} \frac{1}{x\sqrt{2\pi t}} \exp\left(-\frac{(\ln \frac{x}{m})^2}{2t}\right) \mathrm{d}t$$

leads to:

$$P(x) \propto x^{-1} e^{-\sqrt{2\lambda(\ln \frac{x}{m})^2}},$$

and therefore, surprisingly, two different scaling regimes. Enjoyable suffering may be involved. Really enjoyable suffering. But many monks have found a way so you should follow their path laid out below.

Hints and steps:

- Make the substitution $t = u^2$ to find an integral of the form (excluding a constant of proportionality)

$$I_1(a, b) = \int_0^{\infty} \exp\left(-au^2 - b/u^2\right) \mathrm{d}u$$

  where in our case $a = \lambda$ and $b = (\ln \frac{x}{m})^2/2$.

- Substitute $au^2 = t^2$ into the above to find

$$I_1(a, b) = \frac{1}{\sqrt{a}} \int_0^{\infty} \exp\left(-t^2 - ab/t^2\right) \mathrm{d}t$$

- Now work on this integral:

$$I_2(r) = \int_0^{\infty} \exp\left(-t^2 - r/t^2\right) \mathrm{d}t$$

  where $r = ab$.

- Differentiate I_2 with respect to r to create a simple differential equation for I_2. You will need to use the substitution $u = \sqrt{r}/t$ and your differential equation should be of the (very simple) form

  $$\frac{\mathrm{d}I_2(r)}{\mathrm{d}r} = -(\text{something})I_2(r).$$

- Solve the differential equation you find. To find the constant of integration, you can evaluate $I_2(0)$ separately:

  $$I_2(0) = \int_0^\infty \exp(-t^2)\mathrm{d}t,$$

  where our friend $\Gamma(frac12)$ comes into play.

A collection of questions from earlier seasons of PoCS, Vol 2 (also variously known as CoNKs, CocoNuTs, and Complex Networks).

This is all a big soup and some questions may be poorly constructed or repeated.

- The first series of questions will explore real networks by performing some key measurements introduced in Principles of Complex Systems, Vol. 1.
- For general coherence with other humans, you are encouraged to use Python. Also very good: Unix command line tools, R, Julia, Matlab. But you can of course use whatever system you like.
- Data is available in two compressed formats:
  - Matlab + text (tgz): https://pdodds.w3.uvm.edu/teaching/courses/2023-2024pocsverse/data/303complexnetworks-data-package.tgz
  - Matlab + text (zip): https://pdodds.w3.uvm.edu/teaching/courses/2023-2024pocsverse/data/303complexnetworks-data-package.zip

  and can also be found on the course website (helpfully) under data.
- The main Matlab file containing everything is networkdata_combined.mat.
- For directed networks, the ijth entry of the adjacency matrix represents the weight of the link from node i to node j. Adjacency matrices for undirected networks are symmetric.
- For all questions below, treat each network as undirected unless otherwise instructed.

- For this assignment, convert all weights on links to 1, if the network is weighted.
- You do not have to use Matlab for your basic analyses. Python would be a preferred route for many.
- The supplied text versions may be of use for visualization using gml.
- The Matlab command spy will give you a quick plot of a sparse adjacency matrix.
- Real data sets used here are taken from Mark Newman's compilation (and linked-to sites) at http://www-personal.umich.edu/~mejn/netdata/.

1. Record in a table the following basic characteristics:
   - N, the number of nodes;
   - m, the total number of links;
   - Whether the network is undirected or directed based on the symmetry of the adjacency matrix;
   - $\langle k \rangle$, the average degree ($\langle k_{\rm in} \rangle$ and $\langle k_{\rm out} \rangle$ if the network is directed);
   - The maximum degree $k^{\rm max}$ (for both out-degree and in-degree if the network is directed);
   - The minimum degree $k^{\rm min}$ (for both out-degree and in-degree if the network is directed).
2. (3+3)
   (a) Plot the degree distribution P_k as a function of k. In the case that P_k versus k is uninformative, also produce plots that are clarifying. For example, $\log_{10} P_k$ versus $\log_{10} k$.
   (Note: Always use base 10.)
   (b) See if you can characterize the distributions you find (e.g., exponential, power law, etc.).
3. Measure the clustering coefficient C_2 where

$$C_2 = \frac{3 \times \#\text{triangles}}{\#\text{triples}}.$$

For directed networks, transform them into undirected ones first.

One approach is to compute C_2 as

$$C_2 = \frac{3 \times \frac{1}{6}\mathrm{Tr}A^3}{\frac{1}{2}\left(\sum_{ij}[A^2]_{ij} - \mathrm{Tr}A^2\right)}.$$

Note: avoiding computing A^3 is important and can be done.

4. For each of our main six networks, compute and present distributions of the shortest path length between all pairs of nodes. Notation: $d_{i,j}$ is the shortest distance between i and j.

   Also compute the average shortest path length, $\langle d \rangle$.

5. Generate ensembles of random networks of the same 'size' as the six networks. Process 1 random network and then scale up as computing power/time/sanity permits. 1000 random networks would be good.

   Size here means having the same number of nodes and the same number of edges.

   As for the real networks, compute the shortest path lengths for these random networks and present frequency distributions.

6. Determine how well/poorly random networks produce the shortest path distributions of real world networks.

   Using whatever tests you like, show how well both the average shortest path length and the full distributions compare between the real network and their random counterparts.

7. Given N labelled nodes and allowing for all possible number of edges m, what's the total number of undirected, unweighted networks we can construct?

   How does this number scale with N?

8. Given N labelled nodes and a variable number of m edges, for what value of m do we obtain the largest diversity of networks? And for this m, how does the number of networks scale with N?

9. We've seen that large random networks have essentially no clustering, meaning that locally, random networks are pure branching networks. Nevertheless, a finite, non-zero number of triangles will be present.

   For pure random networks, with connection probability $p = \langle k \rangle/(N-1)$, what is the expected total number of triangles as $N \to \infty$?

10. Repeat the preceding calculation for cycles of length 4 and 5 (triangles are cycles of length 3).

11. Show that the second moment of the Poisson distribution is

$$\langle k^2 \rangle = \langle k \rangle^2 + \langle k \rangle.$$

   and hence that the variance is $\sigma^2 = \langle k \rangle$.

12. We've figured out in class that for large enough N (and $\langle k \rangle$ fixed), a random network always has a Poisson degree distribution:

$$P(k; \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$$

where $\lambda = \langle k \rangle$. And as we've discussed, we don't find these networks in the real world (they don't arise due to simple mechanisms). Let's investigate this oddness a little further.

Compute the expected size of the largest degree in an infinite random network given $\langle k \rangle$ and as a function of increasing sample size N. In other words, in selecting (with replacement) N degrees from a pure Poisson distribution with mean $\langle k \rangle$, what's the expected minimum value of the largest degree $\min k_{\max}$?

A good way to compute $k_{\max}$ is to equate it to the value for which we expect $1/N$ of our random selections to exceed. (We had a question in 300 along these lines for power-law size distributions.)

Hint—Of course we'll be using Stirling's Approximation.:
http://www.youtube.com/watch?v=uK5yakuX59M

13. Generating functions and giant components: In this question, you will use generating functions to obtain a number of results we found in class for standard random networks.

    (a) For an infinite standard random network (Erdös-Rényi/ER network) with average degree $\langle k \rangle$, compute the generating function F_P for the degree distribution P_k.
    (Recall the degree distribution is Poisson: $P_k = e^{-\langle k \rangle} \langle k \rangle^k / k!$, $k \geq 0$.)

    (b) Show that $F_P'(1) = \langle k \rangle$ (as it should).

    (c) Using the joyous properties of generating functions, show that the second moment of the degree distribution is $\langle k^2 \rangle = \langle k \rangle^2 + \langle k \rangle$.

14. (a) Continuing on from Q1 for infinite standard random networks, find the generating function $F_R(x)$ for the $\{R_k\}$, where R_k is the probability that a node arrived at by following a random direction on a randomly chosen edge has k outgoing edges.

    (b) Now, using $F_R(x)$ determine the average number of outgoing edges from a randomly-arrived-at-along-a-random-edge node.

    (c) Given your findings above and the condition for a giant component existing in terms of generating functions, what is the condition on $\langle k \rangle$ for a standard random network to have a giant component?

15. (a) Find the generating function for the degree distribution P_k of a finite random network with N nodes and an edge probability of p.

    (b) Show that the generating function for the finite ER network tends to the generating function for the infinite one. Do this by taking the limit $N \to \infty$ and $p \to 0$ such that $p(N-1) = \langle k \rangle$ remains constant.

16. (a) Prove that if random variables U and V are distributed over the non-negative integers then the generating function for the random variable $W = U + V$ is

$$F_W(x) = F_U(x) F_V(x).$$

    Denote the specific distributions by $\mathbf{Pr}(U = i) = U_i$, $\mathbf{Pr}(V = i) = V_i$, and $\mathbf{Pr}(W = i) = W_i$.

    (b) Using the your result in part (a), argue that if

$$W = \sum_{j=1}^{U} V^{(j)}$$

    where $V^{(j)} \overset{d}{=} V$ then

$$F_W(x) = F_U(F_V(x)).$$

    Hint: write down the generating function of probability distribution of $\sum_{j=1}^{k} V^{(j)}$ in terms of $F_V(x)$.

17. (a) Again, given

$$W = \sum_{i=1}^{U} V^{(i)} \text{ with each } V^{(i)} \overset{d}{=} V$$

    where we know that

$$F_W(x) = F_U\left(F_V(x)\right),$$

    determine the mean of W in terms of the means of U and V.

    (b) For $W = U + V$, similarly find the mean of W in terms of U and V via generating functions. Your answer should make rather good sense.

18. Consider the family of generalized random networks with

$$P_k = a\delta_{k1} + (1-a)\delta_{k3}$$

where $0 \le a \le 1$.

General note: We worked through the $a = 1/2$ case in class so those notes should be rather helpful.

Determine the following (3 points each for a–d):

(a)
   i. The distribution of other friends for a node arrived along a randomly chosen direction of a randomly chosen edge, R_k.
   ii. The generating function $F_P(x)$.
   iii. The generating function $F_R(x)$, both directly from R_k and via $F_R(x) = F_P'(x)/F_P'(1)$.

(b) For which values of a a giant component exists, noting the critical value a_c if any phase transition is present.

(c)
   i. The generating function $F_\rho(x)$. Note: Do not expand the form you find.
   ii. The probability that a random edge leads to a subcomponent of finite size, $F_\rho(1)$.

(d)
   i. The generating function $F_\pi(x)$.
   ii. The fractional size of the largest component $S_1 = 1 - F_\pi(1)$ as a function of a.

19. By expanding $F_\rho(x)$ as a formal power series, find the probabilities that a random edge leads to components of finite size 1, 2, 3, 4, and 5, all as a function of a.

20. Using Python's NetworkX (or similar package in any language), simulate random networks with $N = 10^4$ nodes and determine the fractional size of the giant component as a function of a.

    Plot the simulation's output against your theoretical curve determined in the first question.

21. (3 + 3 + 3 + 3)

    Generalize the theory for the previous questions and solve for the same quantities and features in Q1a–Q1d for random networks with:

$$P_k = a\delta_{k1} + (1-a)\delta_{kk'}$$

    for fixed $k' \geq 2$ with $0 \leq a \leq 1$.

    Modifications:

    You will be able to do Q1a and Q1b exactly.

    Important: Please minimally set up and then solve Q1c and Q1d numerically (only) for $k' = 3, \ldots, 10$.

    Put everything on the same plot.

(a)
   i. The distribution of other friends for a node arrived along a randomly chosen direction of a randomly chosen edge, R_k.
   ii. The generating function $F_P(x)$.

iii. The generating function $F_R(x)$, both directly from R_k and via $F_R(x) = F'_P(x)/F'_P(1)$.

(b) For which values of a a giant component exists, noting the critical value a_c if any phase transition is present.

(c) i. The generating function $F_\rho(x)$. Note: Do not expand the form you find.

ii. The probability that a random edge leads to a subcomponent of finite size, $F_\rho(1)$.

(d) i. The generating function $F_\pi(x)$.

ii. The fractional size of the largest component $S_1 = 1 - F_\pi(1)$ as a function of a.

22. Plan: Work through some random bipartite calculations reproducing a few results from the classic Newman *et al.* paper [10]. Our stories are their stars, and our tropes are their movies.

Please note that we use a different convention for defining certain distributions, not just notation. It's a bit confusing. Okay, it's very confusing.

Here's a key to help:

Feature:	Our notation:	Newman *et al.* [10]:
First node type, symbol	stories, 🎞	movies, 0
Second node type, symbol	tropes, 💡	actors, 1
Number of type 1 nodes	$N_{\text{🎞}}$	M
Number of type 2 nodes	$N_{\text{💡}}$	N
Average affiliations of type 1 nodes	$\langle k \rangle_{\text{🎞}}$	ν
Average affiliations of type 2 nodes	$\langle k \rangle_{\text{💡}}$	μ
Affiliation distribution for type 1 nodes	$P_k^{(\text{🎞})}$	q_k
Affiliation distribution for type 1 nodes	$P_k^{(\text{💡})}$	p_k
P Generating function for type 1 nodes	$F_{P(\text{🎞})}$	g_0
P generating function for type 2 nodes	$F_{P(\text{💡})}$	f_0
R generating function for type 1 nodes	$F_{P(\text{🎞})}$	g_1
R generating function for type 2 nodes	$F_{P(\text{💡})}$	f_1
Induced P generating function for type 1 nodes	$F_{P^{(\text{🎞})}_{\text{ind}}}$	F_0
Induced P generating function for type 2 nodes	$F_{P^{(\text{💡})}_{\text{ind}}}$	G_0
Induced R generating function for type 1 nodes	$F_{R^{(\text{🎞})}_{\text{ind}}}$	F_1
Induced R generating function for type 2 nodes	$F_{R^{(\text{💡})}_{\text{ind}}}$	G_1

Note: You can of course use something simple like a and b instead of the film and lightbulb glyphs. Nevertheless, for notation happiness, feel free to use font awesome and the following structures:

```
\usepackage{fontawesome}

%% random biparite networks

\newcommand{\rbone}{\textnormal{\faFilm}}
\newcommand{\rbtwo}{\textnormal{\faLightbulbO}}

\newcommand{\rboneng}{N_{\rbone}}
\newcommand{\rbtwong}{N_{\rbtwo}}

\newcommand{\Prboneind}{P^{(\rbone)}_{\textnormal{ind},k}}
\newcommand{\Prbtwoind}{P^{(\rbtwo)}_{\textnormal{ind},k}}

\newcommand{\Rrboneind}{R^{(\rbone)}_{\textnormal{ind},k}}
\newcommand{\Rrbtwoind}{R^{(\rbtwo)}_{\textnormal{ind},k}}

\newcommand{\Prboneindplain}{P^{(\rbone)}_{\textnormal{ind}}}
\newcommand{\Prbtwoindplain}{P^{(\rbtwo)}_{\textnormal{ind}}}

\newcommand{\Rrboneindplain}{R^{(\rbone)}_{\textnormal{ind}}}
\newcommand{\Rrbtwoindplain}{R^{(\rbtwo)}_{\textnormal{ind}}}
```

Show that the triple-triangle clustering coefficient for the induced networks produced by an arbitrary random bipartite affiliation graph are

$$C_2^{(\text{🎞})} = \frac{N_{\text{💡}}}{N_{\text{🎞}}} \frac{F'''_{P^{(\text{💡})}}(1)}{F''_{P^{(\text{🎞})}_{\text{ind}}}(1)}$$

and

$$C_2^{(\text{💡})} = \frac{N_{\text{🎞}}}{N_{\text{💡}}} \frac{F'''_{P^{(\text{🎞})}}(1)}{F''_{P^{(\text{💡})}_{\text{ind}}}(1)}$$

23. (6 + 6 + 6) Consider the following bipartite affiliation graph degree distributions.

    (a) Fixed degree and fixed degree: $k_{\text{🎞}}$ and $k_{\text{💡}}$, both at least 1.

    (b) Poisson (mean $\langle k \rangle_{\text{🎞}}$) and fixed degree ($k_{\text{💡}}$):

    (c) Poisson and Poisson with mean degrees $\langle k \rangle_{\text{🎞}}$ and $\langle k \rangle_{\text{💡}}$.

For each case, determine these generating functions: $F_{P(\text{🎞})}(x)$, $F_{P(\text{💡})}(x)$, $F_{R(\text{🎞})}(x)$, $F_{R(\text{💡})}(x)$, $F_{P^{(\text{🎞})}_{\text{ind}}}(x)$, $F_{R^{(\text{🎞})}_{\text{ind}}}(x)$, $F_{P^{(\text{💡})}_{\text{ind}}}(x)$, and $F_{R^{(\text{💡})}_{\text{ind}}}(x)$.

24. For the three bipartite graphs given above, determine the condition for a giant component in both induced networks, i.e.,

$$\langle k \rangle_{R,\text{🎞},\text{ind}} \equiv \langle k \rangle_{R,\text{💡},\text{ind}} > 1$$

where

$$\begin{aligned}\langle k \rangle_{R,\text{🎞},\text{ind}} = \langle k \rangle_{R,\text{💡},\text{ind}} &= \frac{F''_{P(\text{💡})}(1)}{F'_{P(\text{💡})}(1)} \frac{F''_{P(\text{🎞})}(1)}{F'_{P(\text{🎞})}(1)} \\ &= \frac{\langle k(k-1) \rangle_{\text{🎞}}}{\langle k \rangle_{\text{🎞}}} \frac{\langle k(k-1) \rangle_{\text{💡}}}{\langle k \rangle_{\text{💡}}}.\end{aligned}$$

25. Using whatever network package you like, construct random bipartite affiliation networks to reproduce Fig. 7 from [10]:

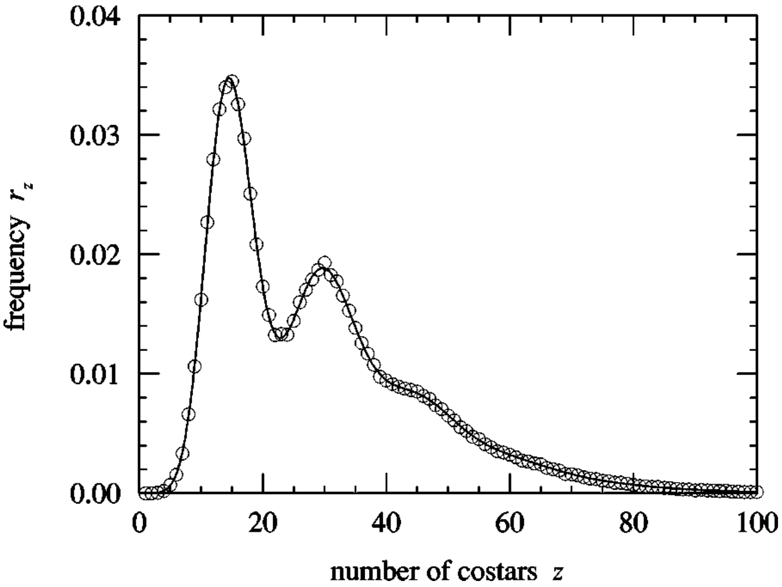


FIG. 7. The frequency distribution of numbers of co-stars of an actor in a bipartite graph with $\mu = 1.5$ and $\nu = 15$. The points are simulation results for $M = 10\,000$ and $N = 100\,000$. The line is the exact solution, Eqs. (89) and (90). The error bars on the numerical results are smaller than the points.

Consider this to be $P^{(\text{💡})}_{\text{ind},k}$, the probability a trope shares appears alongside k other tropes in stories.

Parameters: $N_{\text{🎞}} = 10^4$, $N_{\text{💡}} = 10^5$, $\langle k \rangle_{\text{🎞}} = 15$, and $\langle k \rangle_{\text{💡}} = 1.5$.

26. Plot the induced distribution $P^{(\text{🎞})}_{\textrm{ind},k}$, the probability a story is connected to k other stories through shared tropes.

27. (optional)

    Derive equation 89 in [10] for the degree distribution:

    $$P^{(\text{🎞})}_{\textrm{ind},k} = \frac{(\langle k \rangle_{\text{💡}})^k}{k!} e^{\langle k \rangle_{\text{🎞}}(e^{-\langle k \rangle_{\text{💡}}}-1)} \sum_{i=1}^{k} \left\{ {k \atop i} \right\} \left[\langle k \rangle_{\text{🎞}} e^{-\langle k \rangle_{\text{💡}}} \right]^i ,$$

    where

    $$\left\{ {k \atop i} \right\} = \sum_{j=1}^{i} \frac{(-1)^{i-j}}{j!(i-j)!} j^k$$

    is the Stirling number of the second kind.

28. (optional) Add the theoretical curve obtained above to the plot you generated before that.

29. Data snaring and wrangling:

    Find two (2) interesting, large network data sets online. The networks may be weighted or not, directed or undirected.

    Transform each network's representation into row, column, and weight vectors as per the first assignment. The row vector contains the node at the start of an edge, the column vector the ends, and the weights, well, the weight of the edge.

    Include a one line description for each network along with a link to the data source.

    This time round, if you haven't already, please give NetworkX a shot too.

    Please submit your data via email with the subject heading
    "CocoNuTS: Network submission for "The Darkest Timeline" ↗".

    In the next assignment, we'll examine all submitted networks. Possibly.

    For questions 30–35:

    Consider the simple spreading mechanism on generalized random networks for which each link has a probability $\beta \leq 1$ of successfully transmitting a disease.

    We assume that this transmission probability is tested only once: either a link will or will not be able to send an infection one way or the other (this is a bond percolation problem). We'll call these edges 'active.'

    Denote the degree distribution of the network as P_k and the corresponding generating function as F_P. In class, we wrote down the probability that a node has k active edges as

    $$\tilde{P}_k = \beta^k \sum_{i=k}^{\infty} \binom{i}{k} (1-\beta)^{i-k} P_i .$$

30. Given a random network with degree distribution P_k, find $F_{\tilde{P}}$, the generating function for $\tilde{P}_k$, in terms of F_P.

31. Find the generating function for $\tilde{R}_k$, the analogous version of R_k, the probability that a random friend has k other friends.

32. For standard random (ER) networks, use your results from the preceding questions to find the critical value of $\langle k \rangle$ above which global spreading occurs.

33. Find an expression connecting the three quantities β, the average degree $\langle k \rangle$, and the size of the giant component $\tilde{S}_1$.

34. What is the slope of the $\tilde{S}_1$ curve near the critical point for ER networks?

35. Using whichever method you find most exciting, plot how $\tilde{S}_1$ depends on $\langle k \rangle$ for $\beta = 1$, $\beta = 0.8$, and $\beta = 0.5$.

36. Using either generating function methods (original) or the physical approach (better) from slides on contagion, reproduce the following pieces from Watts's 2002 paper [11] on global cascades on random networks:

    (a) The cascade windows diagram in Fig. 1.

    (b) The vulnerable and triggering component curves in Fig. 2b.

    Note 1: Only the vulnerable component was determined theoretically in [11]. The slides go further and determine the triggering component's size.

    Note 2: This question is all theory but you will need to solve the second and third problems numerically.

37. Using Gleeson and Calahane's iterative equations below, derive the contagion condition for a vanishing seed by taking the limit $\phi_0 \to 0$ and $t \to \infty$.

$$\phi_{t+1} = \phi_0 + (1 - \phi_0) \sum_{k=0}^{\infty} P_k \sum_{j=0}^{k} \binom{k}{j} \theta_t^j (1 - \theta_t)^{k-j} B_{kj},$$

$$\theta_{t+1} = \phi_0 + (1 - \phi_0) \sum_{k=1}^{\infty} \frac{k P_k}{\langle k \rangle} \sum_{j=0}^{k-1} \binom{k-1}{j} {\theta_t}^j (1 - \theta_t)^{k-1-j} B_{kj},$$

where $\theta_0 = \phi_0$, and B_{kj} is the probability that a degree k node becomes active when j of its neighbors are active.

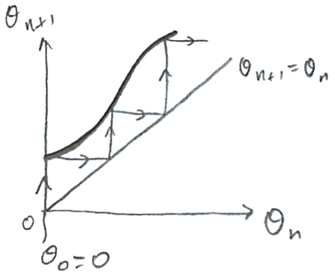
Recall that by contagion condition, we mean the requirements of a random network for macroscopic spreading to occur.

To connect the paper's model and notation to those of our lectures, given a specific response function F and a threshold model, the B_{kj} are given by $B_{kj} = F(j/k)$.

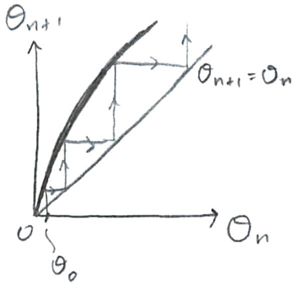
Allow B_{k0} to be arbitrary (i.e., not necessarily 0 as for simple threshold functions).

Here's a graphical hint for the three cases you need to consider as $\theta_0 \to 0$:

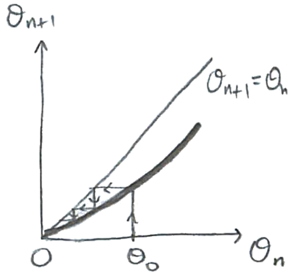
Success:



Sucesss:



Fail:



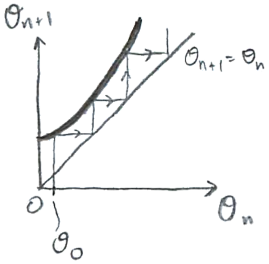
38. Derive equation 4 in Gleeson and Cahalane (2007) [12]:

$$C_\ell = \sum_{k=\ell+1}^{\infty} \sum_{j=0}^{\ell} \binom{k-1}{\ell} \binom{\ell}{j} (-1)^{\ell+j} \frac{kP_k}{\langle k \rangle} F\left(\frac{j}{k}\right).$$

39. (9 pts)

(a) Derive equation 6 in Gleeson and Cahalane (2007), which is a second order approximation to the cascade condition for vanishing seeds.

Here's an example of how this must work:

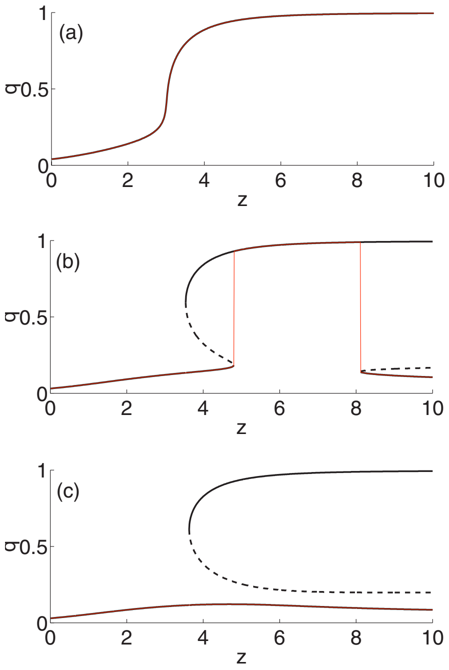


(b) Hence reproduce the dashed analytic curve shown in Figure 1 of their paper.

(c) Explain why there are jumps in the cascade window outline that do not occur at reciprocals of the integers.

40. (6 pts)

(a) By numerically finding the fixed points of $\theta_{t+1} = G(\theta_t; 0)$, reproduce Figure 3 in Gleeson and Cahalane (2007):



(b) Also plot $G(\theta_t; 0)$ for an average threshold $\phi_*(= R)$ of 0.371 for $\langle k \rangle (= z) = 1, 2, 3, \ldots, 10.$

Add the cobweb diagram for a $\phi_0 = 0$ seed.

Do this by creating a recursive plotting script in matlab, for example.

You can use the following Matlab scripts/data as a basis, and most of the work is done. You'll need to improve the plots with some labels, and interpret them properly. The first function calls the other two.

https://pdodds.w3.uvm.edu//share/matlab/Gfunction.m
https://pdodds.w3.uvm.edu//share/matlab/gleeson_{f}ig3_{0}2.mat
https://pdodds.w3.uvm.edu//share/matlab/cobweb3.m

(c) Discuss how the stable points move with $\langle k \rangle$.

Note: $\phi_* = 0.371$ matches plot (b) in Figure 3 of [12].

41. We've figured out in class that for large enough N (and $\langle k \rangle$ fixed), a random network always has a Poisson degree distribution:

$$P(k; \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$$

where $\lambda = \langle k \rangle$. And as we've discussed, we don't find these networks in the real world (they don't arise due to simple mechanisms). Let's investigate this oddness a little further.

Compute the expected size of the largest degree in an infinite random network given $\langle k \rangle$ and as a function of increasing sample size N. In other words, in selecting (with replacement) N degrees from a pure Poisson distribution with mean $\langle k \rangle$, what's the expected minimum value of the largest degree $\min k_{\max}$?

A good way to compute $k_{\max}$ is to equate it to the value for which we expect $1/N$ of our random selections to exceed. (We had a question in 300 along these lines for power-law size distributions.)

Hint—Of course we'll be using Stirling's Approximation.:
http://www.youtube.com/watch?v=uK5yakuX59M

42. In 1-d, consider a population density $\rho(x) = cx^{-\gamma}$ for $x \geq 1$ and $\gamma > 2$ (note that $c = \gamma - 1$).

    Find the ideal distribution for N sources where N is large.

    Hint: draw yourself a clear picture of what's going on.

    Hint: guess the form of the locations of the centers and work from there.

    Also: Feel free to do some numerics to see how things work.

43. Repeat the above treatment for $\rho(x) = \lambda e^{-\lambda x}$ for $x \geq 0$.

44. Yes, even more on power law size distributions. It's good for you.

    For the probability distribution $P(x) = cx^{-\gamma}$, $0 < a \leq x \leq b$, compute the mean absolute displacement (MAD), which is given by $\langle |X - \langle X \rangle| \rangle$ where $\langle \cdot \rangle$ represents expected value. As always, simplify your expression as much as possible.

    MAD is a more reasonable estimate for the width of a distribution, but we like variance σ^2 because the calculations are much prettier. Really.

45. In the limit of $b \to \infty$, show that MAD asymptotically behave as:

    $$\langle |X - \langle X \rangle| \rangle = \frac{2(\gamma - 2)^{(\gamma-3)}}{(\gamma - 1)^{(\gamma-2)}} a.$$

    How does this compare with the behavior of the variance? (See the last question of Assignment todo???.)

46. Using the CCDF and standard linear regression, measure the exponent $\gamma - 1$ as a function of the upper limit of the scaling window, with a fixed lower limit of $k_{\min} = 200$.

    Please plot γ as a function of $k_{\max}$, including 95% confidence intervals.

    Note that the break in scaling should mess things up but we're interested here in how stable the estimate of γ is up until the break point.

    Comment on the stability of γ over variable window sizes.

    Pro Tip: your upper limit values should be distributed evenly in log space.

47. (3 + 3 + 3)

    Estimating the rare:

    Google's raw data is for word frequency $k \geq 200$ so let's deal with that issue now.

    From Assignment 2, we had for word frequency in the range $200 \leq k \leq 10^7$, a fit for the CCDF of

    $$N_{\geq k} \sim 3.46 \times 10^8 k^{-0.661},$$

    ignoring errors.

    (a) Using the above fit, create a complete hypothetical N_k by expanding N_k back for $k = 1$ to $k = 199$, and plot the result in double-log space (meaning log-log space).

    (b) Compute the mean and variance of this reconstructed distribution.

    (c) Estimate:

        i. the hypothetical fraction of words that appear once out of all words (think of words as organisms here),

        ii. the hypothetical total number and fraction of unique words in Google's data set (think at the species level now),

        iii. and what fraction of total words are left out of the Google data set by providing only those with counts $k \geq 200$ (back to words as organisms).

48. Starting from here: http://mskcc.convio.net/pdf/cycle_{f}or_{s}urvival/cfs_{c}ancer_{f}act_{s}heet1.pdf, explore the "rare are legion" aspect of heavy-tailed distributions for cancer.

49. Explain the scaling of RPM for engines.

50. **Zombies!**

    (Optional. But taking practical precautions for your survival in the event of a global zombie attack is not optional.)

    Network version of the SZR model:

    Based on the work of Munz *et al.* [13], we will model Zombie attacks on generalized random networks (the paper is here).

    There are three states: S, susceptible, Z, zombie, and, R, removed.

For the random mixing model studied by Munz et al., the differential equations are

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \theta - \beta SZ - \delta S,$$
$$\frac{\mathrm{d}Z}{\mathrm{d}t} = \beta SZ + \zeta R - \alpha SZ,$$
$$\text{and } \frac{\mathrm{d}R}{\mathrm{d}t} = \delta S + \alpha SZ - \zeta R,$$

where

θ is the birth rate of new susceptibles;

β is the rate at which susceptibles who bump into zombies become zombies

δ is the background, non-zombie related death rate for susceptibles;

ζ is the rate at which the dead (removed) are resurrected as zombies;

and α is the rate at which susceptibles defeat zombies (through traditional methods shown in movies).

For our purposes, consider a random network with degree distribution P_k containing completely susceptible individuals and discrete time updates. We'll now think of the parameters above as probabilities, and ignore birth and death processes ($\theta = \delta = 0$).

We'll further assume that if a susceptible takes out a zombie, the latter cannot resurrect. So this means there's a fourth category, let's call it D for definitely dead.

Assume that in each time step, all edges convey interactions, meaning each individual interacts with each of their neighbors.

Under what conditions (P_k and spreading parameters) will local zombification be guaranteed to take off (i.e., grow exponentially, at least in the short term), given one randomly chosen individual becomes the first zombie?

(The long term dynamics will likely be complicated so we will focus on the initial dynamics.)

See http://www.wired.com/wiredscience/2009/08/zombies/ for more information/enjoyment.

51. (12 pts) Consider a family of undirected random networks with degree distribution

$$P_k = c\delta_{k1} + (1-c)\delta_{k3},$$

where δ_{ij} is the Kronecker delta function, and where c is a constant to be determined below. Also allow nodes to be correlated according to the following node-node mixing probabilities.

Conditional probability version, $P(k\,|k'\,)$:

$$\begin{aligned} P(1\,|\,1) &= \frac{1}{2}(1+r), \\ P(3\,|\,1) &= \frac{1}{2}(1-r), \\ P(1\,|\,3) &= \frac{1}{2}(1-r), \\ \text{and } P(3\,|\,3) &= \frac{1}{2}(1+r). \end{aligned}$$

where $-1 \le r \le 1$ is the family's tunable parameter.

Newman's correlation probability version:

$$E = [e_{ij}] = \begin{bmatrix} e_{00} & e_{02} \\ e_{20} & e_{22} \end{bmatrix} = \frac{1}{4}\begin{bmatrix} (1+r) & (1-r) \\ (1-r) & (1+r) \end{bmatrix}$$

where e_{ij} is the probability that a randomly chosen edge connects a node of degree $i+1$ an a node of degree $j+1$, and only the non-zero values of E are shown.

For the following questions, you can use either formulation.

(a) Determine c so that purely disassortative networks are achievable if r is tuned to -1.

(b) Determine which networks in this family have a giant component. In other words, find the values of r for which a giant component exists.

Note which value (or values) of r mark a phase transition.

(c) Analytically determine the size of the giant component as a function of r.

(d) Determine the size of the largest component containing only degree 3 nodes as a function of r.

Hint: allow degree 3 nodes to be always vulnerable ($\beta_{3i} = 1$ for $i = 0$, 1, 2, and 3) and degree 1 nodes to be never vulnerable ($\beta_{1i} = 0$ for $i = 0$ and 1).

52. Spreading on assortative networks: Starting from

$$\theta_{j,t+1} = G_j(\vec{\theta}_t) = \phi_0 + (1-\phi_0)\times$$

$$\sum_{k=1}^{\infty} \frac{e_{j-1,k-1}}{R_{j-1}} \sum_{i=0}^{k-1} \binom{k-1}{i} \theta_{k,t}^{i}(1-\theta_{k,t})^{k-1-i} B_{ki}.$$

show the matrix for which we must have the largest eigenvalue greater than 1 for spreading to occur is

$$\frac{\partial G_j(\vec{0})}{\partial \theta_{k,t}} = \frac{e_{j-1,k-1}}{R_{j-1}}(k-1)(\beta_{k1} - \beta_{k0}).$$

53. Show that for uncorrelated networks, i.e, when $e_{jk} = R_j R_k$, the above condition collapses to the standard condition

$$\sum_{k=1}^{\infty}(k-1)\frac{kP_k}{\langle k\rangle}(\beta_{k1} - \beta_{k0}) > 1.$$

54. (3 + 3 + 3) Optional:

Solve Krapivsky-Redner's model for the pure linear attachment kernel $A_k = k$.

Starting point:

$$n_k = \frac{1}{2}(k-1)n_{k-1} - \frac{1}{2}kn_k + \delta_{k1}$$

with $n_0 = 0$.

(a) Determine n_1.

(b) Find a recursion relation for n_k in terms of n_{k-1}.

(c) Now find

$$n_k = \frac{4}{k(k+1)(k+2)}$$

for all k and hence determine γ.

55. (3 + 3) Optional:

From lectures:

(a) Starting from the recursion relation

$$n_k = \frac{A_{k-1}}{\mu + A_k} n_{k-1},$$

and $n_1 = \mu/(\mu + A_1)$, show that the expression for n_k for the Krapivsky-Redner model with an asymptotically linear attachment kernel A_k is:

$$\frac{\mu}{A_k}\prod_{j=1}^{k}\frac{1}{1+\frac{\mu}{A_j}}.$$

(b) Now show that if $A_k \to k$ for $k \to \infty$ (or for large k), we obtain $n_k \to k^{-\mu-1}$.

56. (3 + 3 + 3) Optional:

From lectures, complete the analysis for the Krapivsky-Redner model with attachment kernel:

$$A_1 = \alpha \text{ and } A_k = k \text{ for } k \geq 2.$$

Find the scaling exponent $\gamma = \mu + 1$ by finding μ. From lectures, we assumed a linear growth in the sum of the attachment kernel weights $\mu t = \sum_{k=1}^{\infty} N_k(t) A_k$, with $\mu = 2$ for the standard kernel $A_k = k$.

We arrived at this expression for μ which you can use as your starting point:

$$1 = \sum_{k=1}^{\infty} \prod_{j=1}^{k} \frac{1}{1 + \frac{\mu}{A_j}}$$

(a) Show that the above expression leads to

$$\frac{\mu}{\alpha} = \sum_{k=2}^{\infty} \frac{\Gamma(k+1)\Gamma(2+\mu)}{\Gamma(k+\mu+1)}$$

Hint: you'll want to separate out the $j = 1$ case for which $A_j = \alpha$.

(b) Now use result that [9]

$$\sum_{k=2}^{\infty} \frac{\Gamma(a+k)}{\Gamma(b+k)} = \frac{\Gamma(a+2)}{(b-a-1)\Gamma(b+1)}$$

to find the connection

$$\mu(\mu - 1) = 2\alpha,$$

and show this leads to

$$\mu = \frac{1 + \sqrt{1 + 8\alpha}}{2}.$$

(c) Interpret how varying α affects the exponent γ, explaining why $\alpha < 1$ and $\alpha > 1$ lead to the particular values of γ that they do.

57. (10 pts) What is the clustering coefficient C for a standard random network with degree distribution P_k? Compute C for the following two cases:

(a) N is finite and links between nodes exist with probability p.

(b) The random network is infinite with mean degree $\langle k \rangle = z$.

Use the definition $C = \frac{3\#\text{triangles}}{\#\text{triples}}$, or equivalently, that C is the probability that if a is connected to b and c, then b and c are connected.

(c) What's the interpretation for the local structure of infinite random networks given your answer to (b)?

58. (25 pts) Generating functions and giant components. In this question, you will use generating functions to obtain a number of results we found in class for standard random networks.

(a) For an infinite standard random network with average degree $\langle k \rangle = z$, compute the generating function for the degree distribution P_k.

(Recall the degree distribution is Poisson: $P_k = e^{-z} z^k / k!$, $k \geq 0$.)

(b) Using your answer to (a) and the joyous properties of generating functions, show that $\langle k \rangle = z$ and that the degree variance is $\langle k^2 \rangle = z^2 + z$.

(c) Find the generating function for the $\{\tilde{q}_k\}$, where q_k is the probability that a node arrived at by following a random direction on a randomly chosen edge has k outgoing edges.

(d) Using your result for (c), determine the average number of outgoing edges from a randomly-arrived-at-along-a-random-edge node.

(e) Based on (d), what is the condition on z for a standard random network to have a giant component?

(Hint: you need to find for what values of z, a randomly chosen neighbor will, on average, have at least one other neighbor.)

59. (15 pts) In Krapivsky and Redner's treatment of growing random network for linear attachment kernels, they assumed $\sum_{k=1}^{\infty} n_k A_k = \mu t$ and found that μ must be such that

$$1 = \sum_{k=1}^{\infty} \prod_{j=1}^{k} \left(1 + \frac{\mu}{A_j} \right)^{-1}.$$

(a) Show that when the attachment kernel is purely linear, $A_j = j$, and when $\mu = 2$, the above equation above is satisfied.

(b) Bonus question territory: Krapivsky and Redner also looked at the specific attachment kernel $A_1 = \alpha$ and $A_k = k$ for $k > 1$, where $\alpha > 0$. They determined that the resulting degree distribution has a power-law tale obeying $k^{-\gamma}$ where $\gamma = (3 + \sqrt{1 + 8\alpha})/2$.

Using this modified linear attachment kernel, show that $\mu(\mu - 1) = 2\alpha$.

60. (20 pts)

Aspects of Kleinberg's search problem in one dimension:

Consider N nodes connected in a 1-d line graph (i.e., a sequence of N nodes lying on a line, with adjacent nodes connected), labelled $i = 1$ to N.

Take our starting node to be at one end of the line, say $i = 1$, and the target node to be at the other end, $i = N$.

Let node $i = 1$ have exactly one long distance link (i.e., a shortcut link).

In attempting to construct a searchable network, we add a link from our start node $i = 1$ to another node $j = 2, \ldots, N$ with probability $cr^{-\alpha}$, where c is a constant of proportionality and $r = j - i$ is the distance between i and j. (Normally, we add links to all nodes but for this question, we're only interested in what happens with the first step from $i = 1$.)

(a) Compute the constant of proportionality c (Hint: the sum over the probabilities of attaching to all other nodes must be unity; use an integral approximation again.)

(b) Show that for $\alpha = 1$, the chance of the link from node $i = 1$ reaching a node at position $j \geq N/2$ is on the order of $1/\ln N$. (This effectively means that by moving along the line starting at i, we should find a shortcut to the other half of the line within a factor of $\ln N$ steps. This is pretty good.)

(c) For $\alpha > 1$, show that the probability of i having a shortcut to the other half of the line decays as an inverse power of N. (This means that our shortcut is likely too close to i and won't help us jump to the other half of the line.)

(d) If $\alpha < 1$, our shortcut will link to the other half of the line with a finite, constant probability, independent of N for large N. So what's the drawback here?

61. More of a note:

- Newman[14]:

$$C_3 = \frac{6 \times \#\text{triangles}}{\#\text{ordered pairs}}$$

- Now count each triple twice
- Same as C_2 but interpretation is different
- Probability that a friend of i's friend is also i's friend.

- C_1 = probability that two friends of a randomly chosen node are connected
- C_2 = probability that two nodes are connected given they have a friend in common.
- $C_3 (= C_2)$ = probability that a node's friend of a friend is also a friend of that node.

- For sparse networks, C_1 tends to discount highly connected nodes.
- While C_1 is a measure of clustering, it doesn't quite as simple interpretation as C_2.
- Some variability in which measure is used in the literature.

- Not always clear which one is being used...

62. Generating functions and giant components: In this question, you will use generating functions to obtain a number of results we found in class for standard random networks.

    (a) For an infinite standard random network (Erdös-Rényi/ER network) with average degree $\langle k \rangle$, compute the generating function F_P for the degree distribution P_k.
    (Recall the degree distribution is Poisson: $P_k = e^{-\langle k \rangle} \langle k \rangle^k / k!$, $k \geq 0$.)

    (b) Show that $F_P'(1) = \langle k \rangle$ (as it should).

    (c) Using the joyous properties of generating functions, show that the second moment of the degree distribution is $\langle k^2 \rangle = \langle k \rangle^2 + \langle k \rangle$.

    (d) Find the generating function for the degree distribution P_k of a finite random network with N nodes and an edge probability of p.

    (e) Show that the generating function for the finite ER network tends to the generating function for the infinite one. Do this by taking the limit $N \to \infty$ and $p \to 0$ such that $p(N-1) = \langle k \rangle$ remains constant.

63. (a) Continuing on from Q1 for infinite standard random networks, find the generating function $F_R(x)$ for the $\{R_k\}$, where R_k is the probability that a node arrived at by following a random direction on a randomly chosen edge has k outgoing edges.

    (b) Now determine the average number of outgoing edges from a randomly-arrived-at-along-a-random-edge node.

    (c) Given your findings above, what is the condition on $\langle k \rangle$ for a standard random network to have a giant component?
    (Hint: you need to find for what values of $\langle k \rangle$, a randomly chosen neighbor will, on average, have at least one other neighbor.)

64. Consider the simple spreading mechanism on generalized random networks for which each link has a probability $\beta \leq 1$ of successfully transmitting a disease.

    We assume that this transmission probability is tested only once: either a link will or will not be able to send an infection one way or the other (this is a bond percolation problem). We'll call these edges 'active.'

    Denote the degree distribution of the network as P_k and the corresponding generating function as F_P. In class, we wrote down the probability that a node has k active edges as

$$P_k' = \beta^k \sum_{i=k}^{\infty} \binom{i}{k} (1-\beta)^{i-k} P_i.$$

(a) Given a random network with degree distribution P_k, find $F_{P'}$, the generating function for P'_k, in terms of F_P.

(b) Find the generating function for R'_k, the analogous version of R_k, the probability that a random friend has k other friends.

65. (a) For standard random networks, use your results for Q3 to find an expression connecting β, the average degree $\langle k \rangle$, and the size of the giant component S'_1.

    (b) What is slope of the S'_1 curve near the critical point for ER networks?

    (c) Using whichever method you find most exciting, plot how S'_1 depends on $\langle k \rangle$ for $\beta = 1$, $\beta = 0.8$, and $\beta = 0.5$.

66. Consider a network with a degree distribution that obeys a power law and is otherwise random.

    Assume that the network is drawn from an ensemble of networks which have N nodes whose degrees are drawn from the probability distribution $P_k = ck^{-\gamma}$ where $k \geq 1$ and $2 < \gamma < 3$.

    (a) Estimate $\min k_{\max}$, the approximate minimum of the largest degree in the network, finding how it depends on N. (Hint: we expect on the order of 1 of the N nodes to have a degree of $\min k_{\max}$ or greater.)

    (b) Determine the average degree of nodes with degree $k \geq \min k_{\max}$ to find how the expected value of $k_{\max}$ scales with N.

Repeats:

I. Supply networks and allometry:

Consider a set of rectangular areas with side lengths L_1 and L_2 such that $L_1 \propto A^{\gamma_1}$ $L_2 \propto A^{\gamma_2}$ where A is area and $\gamma_1 + \gamma_2 = 1$. Assume $\gamma_1 > \gamma_2$.

Now imagine that material has to be distributed from a central source in each of these areas to sinks distributed with density $\rho(A)$, and that these sinks draw the same amount of material per unit time independent of L_1 and L_2.

1. Find an exact form for how the volume of the most efficient distribution network scales with overall area $A = L_1 L_2$. (Hint: you will have to set up a double integration over the rectangle.)

2. If network volume must remain a constant fraction of overall area, determine the maximal scaling of sink density ρ with A.

II. Size-density law:

In two dimensions, the size-density law for distributed source density $D(\vec{x})$ given a sink density $\rho(\vec{x})$ states that $D \propto \rho^{2/3}$. We showed in class that an approximate argument that minimizes the average distance between sinks and nearest sources gives the 2/3 exponent.

1. Repeat this argument for the d-dimensional case and find the general form of the exponent β in $D \propto \rho^{\beta}$.

- We will explore real networks throughout the course performing some key measurements introduced in Principles of Complex Systems.
- you are encouraged to use Python (along with, for example, NetworkX or graph-tools).
- Data is available in two compressed formats:
  - Matlab + text (tgz): https://pdodds.w3.uvm.edu/teaching/courses/2023-2024pocsverse/data/303complexnetworks-data-package.tgz
  - Matlab + text (zip): https://pdodds.w3.uvm.edu/teaching/courses/2023-2024pocsverse/data/303complexnetworks-data-package.zip

  and can also be found on the course website (helpfully) under data.
- The main Matlab file containing everything is networkdata_combined.mat.
- For directed networks, the ijth entry of the adjacency matrix represents the weight of the link from node i to node j. Adjacency matrices for undirected networks are symmetric.
- For all questions below, treat each network as undirected unless otherwise instructed.
- For this assignment, convert all weights on links to 1, if the network is weighted.
- You do not have to use Matlab for your basic analyses. Python would be a preferred route for many.
- The supplied text versions may be of use for visualization using gml.
- The Matlab command spy will give you a quick plot of a sparse adjacency matrix.
- Real data sets used here are taken from Mark Newman's compilation (and linked-to sites) at http://www-personal.umich.edu/~mejn/netdata/.

1. Record in a table the following basic characteristics:
   - N, the number of nodes;
   - m, the total number of links;
   - Whether the network is undirected or directed based on the symmetry of the adjacency matrix;
   - $\langle k \rangle$, the average degree ($\langle k_{\mathrm{in}} \rangle$ and $\langle k_{\mathrm{out}} \rangle$ if the network is directed);
   - The maximum degree k^{max} (for both out-degree and in-degree if the network is directed);
   - The minimum degree k^{min} (for both out-degree and in-degree if the network is directed).

2. (3+3)

   (a) Plot the degree distribution P_k as a function of k. In the case that P_k versus k is uninformative, also produce plots that are clarifying. For example, $\log_{10} P_k$ versus $\log_{10} k$.
   (Note: Always use base 10.)

   (b) See if you can characterize the distributions you find (e.g., exponential, power law, etc.).

3. Measure the clustering coefficient C_2 where

$$C_2 = \frac{3 \times \#\mathrm{triangles}}{\#\mathrm{triples}}.$$

   For directed networks, transform them into undirected ones first.

   One approach is to compute C_2 as

$$C_2 = \frac{3 \times \frac{1}{6}\mathrm{Tr}A^3}{\frac{1}{2}\left(\sum_{ij}[A^2]_{ij} - \mathrm{Tr}A^2\right)}.$$

   Note: avoiding computing A^3 is important and can be done.

- We will explore real networks throughout the course performing some key measurements introduced in Principles of Complex Systems.
- you are encouraged to use Python (along with, for example, NetworkX or graph-tools).
- Data is available in two compressed formats:

  - Matlab + text (tgz): https://pdodds.w3.uvm.edu/teaching/courses/2023-2024pocsverse/data/303complexnetworks-data-package.tgz
  - Matlab + text (zip): https://pdodds.w3.uvm.edu/teaching/courses/2023-2024pocsverse/data/303complexnetworks-data-package.zip

  and can also be found on the course website (helpfully) under data.

- The main Matlab file containing everything is networkdata_combined.mat.
- For directed networks, the ijth entry of the adjacency matrix represents the weight of the link from node i to node j. Adjacency matrices for undirected networks are symmetric.
- For all questions below, treat each network as undirected unless otherwise instructed.
- For this assignment, convert all weights on links to 1, if the network is weighted.
- You do not have to use Matlab for your basic analyses. Python would be a preferred route for many.
- The supplied text versions may be of use for visualization using gml.
- The Matlab command spy will give you a quick plot of a sparse adjacency matrix.
- Real data sets used here are taken from Mark Newman's compilation (and linked-to sites) at http://www-personal.umich.edu/~mejn/netdata/.

1. Okay, let's get back to the 6 networks we explored in the first assignment. Questions 2 through 4 will focus on them.

   Measure the degree-degree assortativity. This is the standard Pearson correlation coefficient and the focus is on links, and then the nodes at the end of each link.

   For undirected networks, we need to think about how we choose the ordering of an edge's two degrees when we perform the correlation. Which degree goes first? Or should we include both orderings? How about randomly choosing the ordering? Does it matter?

   For directed networks, various correlations are possible (in-in, in-out, etc.). For this question, measure the correlation of the in-degree of the source node and the out-degree of the destination node for each link.

2. Produce plots of the adjacency matrices.

3. Using a network visualization tool of your choice, produce plots of the networks (if possible, depending on size).

   For the smaller ones, please label the nodes numerically.

4. (3 + 3)

   Consider a modified version of the Barabàsi-Albert (BA) model [8] where two possible mechanisms are now in play. As in the original model, start with m_0 nodes at time $t = 0$. Let's make these initial guys connected such that each has degree 1. The two mechanisms are:

   M1: With probability p, a new node of degree 1 is added to the network. At time $t + 1$, a node connects to an existing node j with probability

   $$P(\text{connect to node } j) = \frac{k_j}{\sum_{i=1}^{N(t)} k_i} \tag{6}$$

   where k_j is the degree of node j and $N(t)$ is the number of nodes in the system at time t.

   M2: With probability $q = 1 - p$, a randomly chosen node adds a new edge, connecting to node j with the same preferential attachment probability as above.

   Note that in the limit $q = 0$, we retrieve the original BA model (with the difference that we are adding one link at a time rather than m here).

   In the long time limit $t \to \infty$, what is the expected form of the degree distribution P_k?

   Do we move out of the original model's universality class?

   Different analytic approaches are possible including a modification of the BA paper, or a Simon-like one (see also Krapivsky and Redner [9]).

   Hint: You can attempt to solve the problem exactly and you'll find an integrating factor story.

   Another hint, moment of mercy: Approximate the differential equation by considering large t (this will simplify the denominators).

   (3 points for set up, 3 for solving.)

5. Optional:

   Watch "Remedial Chaos Theory."

   Community, S3E04.

   https://en.wikipedia.org/wiki/Remedial_Chaos_Theory

References

[1] J. M. Carlson and J. Doyle. Highly optimized tolerance: A mechanism for power laws in designed systems. *Phys. Rev. E*, 60(2):1412–1427, 1999. pdf

[2] J. M. Carlson and J. Doyle. Highly Optimized Tolerance: Robustness and design in complex systems. *Phys. Rev. Lett.*, 84(11):2529–2532, 2000. pdf

[3] D. J. Watts and S. J. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998. pdf

[4] C. E. Shannon. A mathematical theory of communication. *The Bell System Tech. J.*, 27:379–423,623–656, 1948. pdf

[5] L. Jost. Entropy and diversity. *Oikos*, 113:363–375, 2006. pdf

[6] P. W. Anderson. More is different. *Science*, 177(4047):393–396, 1972. pdf

[7] R. M. May. Will a large complex system be stable? *Nature*, 238:413–414, 1972. pdf

[8] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–511, 1999. pdf

[9] P. L. Krapivsky and S. Redner. Organization of growing random networks. *Phys. Rev. E*, 63:066123, 2001. pdf

[10] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64:026118, 2001. pdf

[11] D. J. Watts. A simple model of global cascades on random networks. *Proc. Natl. Acad. Sci.*, 99(9):5766–5771, 2002. pdf

[12] J. P. Gleeson and D. J. Cahalane. Seed size strongly affects cascades on random networks. *Phys. Rev. E*, 75:056103, 2007. pdf

[13] P. Munz, I. Hudea, J. Imad, and R. J. Smith? When zombies attack!: Mathematical modelling of an outbreak of zombie infection. In J. M. Tchuenche and C. Chiyaka, editors, *Infectious Disease Modelling Research Progress*, pages 133–150. Nova Science Publishers, Inc., 2009. pdf

[14] M. E. J. Newman. The structure and function of complex networks. *SIAM Rev.*, 45(2):167–256, 2003. pdf