Problem Set 4

Due: Beginning of class (at 2 p.m.) on Monday, Nov. 18.

1) In this problem, we will work through a mild but important generalization of our derivation of a power law for the Yule process. (This more general derivation could be interpreted as an approximation of the process giving rise to histograms of city sizes, paper citations, hyperlink in-degrees for the world wide web, in-degrees for routers on the Internet, book sales, movie co-appearances, power station in-degrees on the energy grid, and religion size.)

Consider a collection of objects, for example cities. Each object in this collection has a property k associated with it, e.g. the number of people in the city. This property appears to follow a power law (i.e. the probability that the object has the value k for this property is in some sense approximately proportional to $(k+c)^{-\alpha}$ for some fixed c and α). At each time step, one new city appears in the collection, starting with a population of k_0 people. Also over each time step, a total of m new people are added to the cities in the collection. Each is independently assigned a city at random with a probability proportional to $k+c$ for some constant c. Let $p_{k,n}$ denote the fraction of cities with populations of k people when there are n cities in the collection.

a) For just this problem part, consider the interpretation that the collection of objects is a collection of papers, where k is the number of citations of a given paper. Under this interpretation, $k_0 = 0$. Why then is c useful in assuming that citations are assigned to each paper with a probability proportional to $(k+c)^{-\alpha}$ rather than $k^{-\alpha}$?

b) By an argument analogous to that in the notes for the Yule process, obtain the master equation

$$(n+1)p_{k_0,n+1} = np_{k,n} + 1 - m\frac{k_0+c}{k_0+c+m}p_{k_0,n} \qquad \text{for } k = k_0, \quad \text{(1a)}$$

$$(n+1)p_{k,n+1} = np_{k,n} + m\frac{k-1+c}{k_0+c+m}p_{k-1,n} - m\frac{k+c}{k_0+c+m}p_{k,n} \qquad \text{for } k > k_0. \quad \text{(1b)}$$

In deriving Eqs. (1), you will have to neglect the possibility that a city receives more than one person in a given time step.

c) Show that the long-time solutions ("stationary solutions") of these equations—i.e. the p_k's defined by $p_k = \lim_{n\to\infty} p_{k,n}$—are

$$p_{k_0} = \frac{k_0+c+m}{(m+1)(k_0+c)+m}, \quad \text{(2a)}$$

$$p_k = \frac{\mathrm{B}(k+c,\alpha)}{\mathrm{B}(k_0+c,\alpha)}p_{k_0} \quad \text{for } k > k_0, \quad \text{(2b)}$$

where $\alpha = 2 + (k_0+c)/m$ and B is the beta function. (As we did before, we can conclude that since $\mathrm{B}(a,b) \sim a^{-b}$, this mildly more general Yule process also exhibits a power-law distribution $p_k \sim k^{-\alpha}$.)

2) You might think that for large tables of positive-valued data ranging over several orders of magnitude (accounting tables, census data, geographical statistics, physical constants like specific heats, and so on) the frequency of numbers starting with 1 would be 1/9, and likewise 1/9 for each of the other eight nonzero digits. In fact this is not the case. The percentage of numbers starting with 1 instead turns out to be about 30.1%, the percentage starting with 2 about 17.6%, the percentage beginning with 3 about 12.5% and so on, down to a 4.6% chance of a number starting with 9. This empirical observation is known as *Benford's law*, and a plot (here is one example) shows that the decay looks approximately like a power law. The precise theoretical percentage of numbers (in base 10) starting with the nonzero digit k is $\log_{10}(1 + 1/k)$. For example, the percentage of numbers starting with 2 is expected to be $\log_{10}(1 + 1/2) \approx 0.176091$.

a) *An intuitive explanation for Benford's law.* Draw a hypothetical normalized histogram of the data values, where the x axis is in log scale. This distribution should vary slowly and span many scales. For example, you might draw a log-normal distribution that spans at least six orders of magnitude. (With the x axis in log scale, this will appear as a normal distribution.) Now consider an arbitrary data value x. Write this out in (base 10) scientific notation. Write the condition that the leading number, which is on $[1, 10)$, starts with the digit k in terms of inequalities. Take the base-10 logarithm. Use this and your sketch to estimate the percentage of data values starting with the digit k. (You should get $\log_{10}(1 + 1/k)$.)

b) *One connection between Benford's law and power laws.* Suppose that you reexpress the data values in a base larger than 10, and suppose that the data values still range over several orders of magnitude in this new base. What power law does Benford's law approach for large k? (Hint: Convert the distribution of Benford's law to base e and perform a Taylor expansion about $1/k = 0$.)

3) We are now ready to give a complete set of connections between the other laws and distributions mentioned on the final slide of Lecture 7. (When we say "loosely equivalent" below, we mean in terms of mathematical form.)

a) *Bradford's law is loosely equivalent to Benford's law.* Suppose we have a collection of journals which have been ranked $1, 2, 3, \ldots$ in order of decreasing number of "relevant papers" (e.g. articles of interest). Let $R(n)$ be the total of relevant papers in the first k journals. Bradford's law can be stated as $R(k) = R(k^2) - R(k) = R(k^3) - R(k^2) = \cdots$, where k is any number greater than 1. Now let $r(k+1) = R(k+1) - R(k)$ express the number of relevant papers in the $(k + 1)$th journal. (Note the difference only makes sense for $k > 0$.) Show that $r(k+1) \propto \log_{10}(1+1/k)$, the form of Benford's law. (Hint: Use the fact that the only function that satisfies $\varphi(x^m) = m\varphi(x)$ is $\varphi(x) = a \log_b x$, where a is any constant and b is any base, to show that $R(k) \propto \log_{10} k$.)

b) *Bradford's law is loosely equivalent to Zipf's law.* Suppose the total number of journals is N. Normalize your function $R(k)$ to achieve the cumulative frequency function $F(k) = R(k)/R(N)$. Now let k range continuously from 1 to N and compute $f(k) = \frac{d}{dk}F(k)$ to show that $f(k) = \ln(N)k^{-1}$ (a strong form of Zipf's law).

c) *Zipf's law is loosely equivalent to a power law.* From the notes, a power law has a CCDF of $\Pr[X \geq x] = (x/x_{\min})^{-\alpha+1}$. Let x_k denote the kth largest value of N independent draws from this distribution. Then for very large N, we can expect the probability $\Pr[X \geq x_k]$ for the $(N+1)$th draw would be approximately k/N. Show that this observation implies that x_k is approximately $ck^{-\beta}$ for some constants c and β to be determined.

4) Consider a graph constructed as follows: To a single root node, attach z neighbor nodes. Then to each of these z neighbors, attach $z-1$ additional neighbors, and to each of these new neighbors $z-1$ more neighbors. (Thus the number of nodes grows like 1, $1+z$, $1+z+z(z-1)$, $1+z+z(z-1)+z(z-1)^2$, etc.) Keeping up this construction for infinite steps, we obtain the *Bethe lattice*. Now suppose that we change each edge of the Bethe lattice from closed to open with probability p. The paper REICH78 gives an elaborate and difficult but exact approach to finding the critical probability (i.e. "percolation threshold") at which an infinite cluster of open edges appears on the Bethe lattice. The result is $p_c = 1/(z-1)$. Can you give a much simpler, two-sentence argument for the plausibility of this result? (Hint: What is the expected number of open edges among the $z-1$ neighbors added to each node? If this expected number is less than one, what is the expected cluster size for the case in which $z = 2$? Now let this case serve as an approximation for the case of general z.)

5) *The SIS model with vaccination.* Suppose there are three populations: susceptible individuals with density S, infectious individuals with density I, and vaccinated individuals with density V. The total density of individuals is N, so $S + I + V = N$. Infectious individuals appear at a rate proportional to the probability that a susceptible individual and an infectious individual meet by chance, or SI. Infectious individuals recover and become susceptible again at a rate proportional to their density (a constant rate per capita).

a) From the above description, the dynamics of I are

$$\frac{d}{dt}I = \sigma I(N - V - I) - \rho I, \tag{3}$$

where σ is the rate of infection and ρ is the rate of recovery. Explain why this equation holds and give the units of σ and ρ.

b) Analyze Eq. (3) completely as we have done for other one-dimensional nonlinear systems. (By now, you should know the steps.) Show that in order to ensure that the disease is eradicated, a fraction $v > 1 - 1/R_0$ of the population must be vaccinated, where $R_0 = \sigma N/\rho$. (R_0 is called the *basic reproductive rate* of the pathogen.) The common flu this winter will have an R_0 of 2-3. If the SIS model was the appropriate model for this pathogen, how much of the University of Michigan community would need to be vaccinated for the number of flu cases to remain negligible?

c) In part (b), you should have found (and named) a bifurcation. If you haven't already, make a plot of I^* (the fixed points of I) versus σ for $\rho = 1$ to illustrate this bifurcation. Does this represent a "phase transition"? Why or why not?