Network Analysis and Modeling
CSCI 5352, Fall 2018
Prof. Dan Larremore
Problem Set 6, due 11/20

1. (100 pts total) Consider Price's model of a citation network, applied to publications in a single field.

   (a) (35 pts) Implement the simulation algorithm described in Chapter 13.1 of *Networks* [or 14.1 in the first edition].

   - For choices of $c = 3$ and $n = 10^6$, make a single figure showing the four complementary cumulative distribution functions $\Pr(K \geq k_{\text{in}})$ (the ccdf) for network in-degree k_{in}, one for each choice of $r = \{1, 2, 3, 4\}$.
   - Briefly discuss the impact of the uniform attachment mechanism on the distribution's shape and comment about the fraction of vertices with $k_{\text{in}} = 0$.

   (b) (30 pts) Reasonable values of the model parameters for real citation networks are $c = 12$ and $r = 5$.

   - For these choices, use your numerical simulation to calculate (i) the average number of citations to a paper (in-degree) in the first 10% of published papers (vertices) and (ii) the average number for a paper in the last 10%.
   - Briefly discuss the implications of your results with respect to the "first-mover advantage," and the corresponding bias in citation counts for the first papers published in a field.

   Hint: For a *good* estimate, average your answer over many repetitions of the simulation.

   (c) (15 pts) Visit the *Index of Complex Networks* at `icon.colorado.edu`. Under the ICON entry for "arXiv citation networks (1993-2003)," obtain both the network and dates files for the hep-ph citation network.

   - For (i) the first 10% and (ii) the last 10% of papers with submission dates, compute their average in-degree. Discuss any steps you took to convert the input data into a form on which you could perform these calculations.
   - Briefly discuss how well, and why, these empirical values agree or disagree with your model estimates from question (1b).

   Hint: You will need to "clean" these data a little in order to get good results. The two files contain non-identical lists of ids; there are 30,558 that occur in both files. To do the analysis, it will be useful to construct a list of pairs (i, t_i) of node ids and the dates they were created, in increasing order of t_i. Then think about how to transform the input citation network into a form by which to calculate the desired values.

   (d) (20 pts) Recall that Price's model is a dramatically simplified view of how nodes in a citation network accumulate new connections. Describe at least three ways that the "preferential attachment" mechanism is unrealistic in this context, and for each, suggest

a way that you could analyze a real citation network to demonstrate the difference between what the model predicts and what the real world shows.

(e) (20 pts extra credit) Now consider a variation of Price's model in which we remove the preferential attachment part. That is, each time a new vertex joins the network, each of its c edges attaches to an existing vertex with equal probability. Using the same parameter choices as in question (1a), produce a figure showing the ccdfs for both this model and Price's model, for $r = \{1, 4\}$. Briefly discuss the differences in terms of how citations (edges) are distributed across papers (vertices).