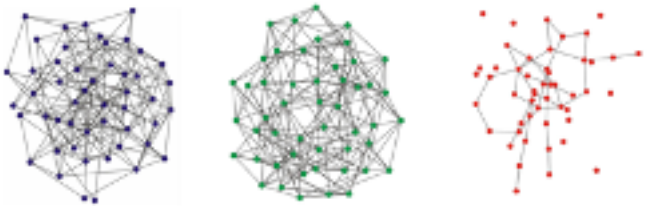

HOMEWORK ASSIGNMENT 3: GENERATIVE MODELS

ENGINEERING 5350
NETWORKS THEORY II FALL 2021
PROFESSOR GABRIEL CWILICH

PROBLEM 1: IT'S GOOD TO BE EARLY

Consider the Barabasi-Albert model of growth of a network of citations of papers in a new field of physics in which papers are published at a constant rate. The field is five years old and the network has now 8,000 papers.

(1) Suppose that you are the author of the very first paper published in the field. How many times more citations will you have (on average) compared with the person who published the paper number 20 in the field.

(2) How big should be the network (number of papers published in the field) when that guy has as many citations as you have right now now. At the current rate of publications, how long will he have to wait to achieve that milestone.

(3) Obtain an expression for the average number of citations per paper, for papers that were published between times t_1 and t_2, where the times are represented as a fraction of the total time the network has been growing. In other words $t_1 = \alpha\, t$ and $t_2 = \beta\, t$. Obviously your expression will depend on m, the number of links that are created when a node is introduced; in this case this represents the average length of the bibliography of the papers.

(4) Assume that the length of the average bibliography is 20 papers. Calculate, then:

- the average number of citations of the first 5% of the papers published in the field.
- the average number of citations of the papers in the 5% of papers published between the 50% and 55% papers published.

- the average number of citations of the last 5% of the papers published in the field.

(5) Interesting question for you to think, that the previous questions suggest. If I ask you for the average number of citations of the first paper published in the field, or the tenth, or the number 357, you need to know not only the size of the bibligraphies, but also the total number of papers in the field. But if I ask you to tell me just the average number of citations of a paper published six months ago, or a week ago, or a year ago, the total number of papers does not matter. Only for how long the network has been growing at its constant rate. Do you see a simple explanation for that?

PROBLEM 2: GROWTH WITHOUT PREFERENTIAL ATTACHMENT

This problem will teach you something that I already showed in the lecture on October 28, (in the section called The exact equations of the B-A model) under the title Growth without preferential attachment . That if you have a network that grows one node at a time, but its m links to nodes chosen at random among all the existing nodes in the network (instead of being chosen with preferential attachment probabilities) the network acquires in the limit of large n a fast decaying exponential distribution of degrees instead of a power law.

But it will teach me, and by extension to you, something more important. That before trying to simplify and approximate a problem, always check if it is not easier to actually solve it exactly. I solved the model without preferential attachment using the continuum approximation for the rate of growth of the instead of just going through the same steps of the exact B-A model that I solved just before in the notes. I got that idea from Barabasi, who does it like that in his book. It turns out that solving it exactly, is five times shorter and much easier to prove the result than doing the approximation, and that is what you will do here.

(1) Consider the rate equations for the Barabasi model, as I did in the lecture, but without preferential attachment. In other words consider that the probability of a certain node in the network acquiring a link when I introduce a new node, instead of being

$$m \times \frac{k}{2t} \quad or \quad m \times \frac{k}{2n} \quad is \quad instead \quad m \times \frac{1}{n}$$

Then derive the rate equations, exactly in the same way I did it for the B-A model, for p_k in terms of p_{k-1}, remembering that you need a special equation for

the case $k = m$. The equations will look very similar but a little simpler than the ones I obtained for the B-A model.

(2) Solve the equations (as usual in the limit $n \to \infty$. The solution is not just a little simpler than the B-A equation, they are just much much simpler. And you will see in no time that the solution becomes

$$p_k = C^k$$

where C is a constant smaller than one, indicating a clear exponential decay of the degrees.

(3) You can express the expression you obtained as

$$p_k = \exp^{-\lambda k} \quad with \quad \lambda = \ln \frac{1}{C}$$

Find C, and show that if you expand the logarithm in the expression above, you get the same result I obtained in the lecture (in a much more complicated way) for the exponential decay of the degrees. In other words, in this case solving the problem exactly is ten times simpler than doing the continuum approximation.

PROBLEM 3: SIMULATING THE BARABASI-ALBERT MODEL

The idea is to study and verify some of the properties of the model numerically.For some of the properties you can get away with using the canned versions of the algorithm. For example Mathematica can generate a network if you simply choose the number of nodes n of the network and the value of m, the number of links that each new node has (which in the Mathematica **BarabasiAlbertGraphDistribution** command is called k), and for certain things you can take advantage for that. If you need to measure things on a network as it grows, and not just the finished network, you will need to do a little programming.

I suggest that you work with a partner if that makes it easier. Remember that in my class of October 21, under the title *simulating the Price model* I made some comments on the very efficient algorithms to select the nodes to link to, using preferential attachment. Also before starting to do this assignment read all the things that are asked because, for example, for items (3), (4) and (5), you do not need to each time to run different simulations. You can run them once, and collect all the data you will need to complete those points.

(1) Dependence with m: Compare Barabasi-Albert networks of the same size n but with different values of the number of links per node m, and look at the distribution of the degrees p_k. Since

$$p_k = \frac{2m(m+1)}{k^3}$$

you expect that if you consider, for example, a network with $m = 2$ as compared with one with $m = 3$, the second one will have twice as many nodes of a certain degree than the first one. Verify if this is true, by picking a network large enough.

(2) Clustering coefficient : I did not have time to discuss in the lecture what is the clustering coefficient of the Barabasi-Albert model. Remember that in a random network (Erdos-Renyi) it goes to zero as $\frac{1}{n}$). In the case of the Barabasi-Albert it also goes to zero but much more slowly as

$$C = \frac{(\ln n)^2}{n}$$

If you want to see why this is so, there is an appendix in Chapter 5 of the Barabasi book where this result is derived in a few lines of algebra. But that means that the Barabasi model cannot explain fully the higher clustering of real social networks, although there are at least more triangles than in the completely random network.

Your task will be here to compare the clustering coefficient of the network for different values of n and see that it changes as the expression given above. Create several networks for each value of n and average the clustering coefficient for each n, and plot the clustering coefficient as a function of n and see how the curve compares with the theoretical expression given above

(3) How does the distributions of degrees converge to the final one?: Create a fairly large network (say $n \approx 10^4$) to study the distribution of the degrees. Remember that to check if something looks like a power law you need to plot p_k vs. k in a log-log plot. If $p_k \approx k^{-\alpha}$, then the log-log plot should be a straight line with slope $-\alpha$. In the Barabasi-Albert model α should be equal to 3.

Look how the distribution changes as the network grows. In other words get the distribution after you added the first 100 nodes, the first 500 nodes, the first 3000 nodes, etc to see how the distribution of the degrees looks at the different stages of growth. See if it converges to the final distribution.

(4) The cumulative distribution: The data in the previous step might be quite noisy (particularly for the higher degrees since there are very few nodes of those degrees), so it will be better to repeat what you did above

$$P(k) = \sum_{k'=k}^{\infty} p_{k'}$$

Remember that if $p_k \approx k^{-\alpha}$, then $P(k) \approx k^{-(\alpha-1)}$ NOTE: you do not need to repeat the simulations, just use the data you collected in the previous step but instead of plotting p_k vs k, plot $P(k)$ vs k

(5) Degree Dynamics: Follow the degree dynamics of a few nodes to see how they follow the theoretical prediction

$$k_i(t) = m\sqrt{\frac{t}{t_i}}$$

Pick a few nodes (say nodes introduced at $t_i = 10$, $t_i = 100$, $t_i = 500$, $t_i = 1000$, $t_i = 3000$) and follow how their degree changes as the network grows until it reaches its final size $n = 10,000$, and see if you can verify the square root behavior as a function of time mentioned above, and compare the curves for the nodes introduced at the different times, to see clearly the early mover's advantage.